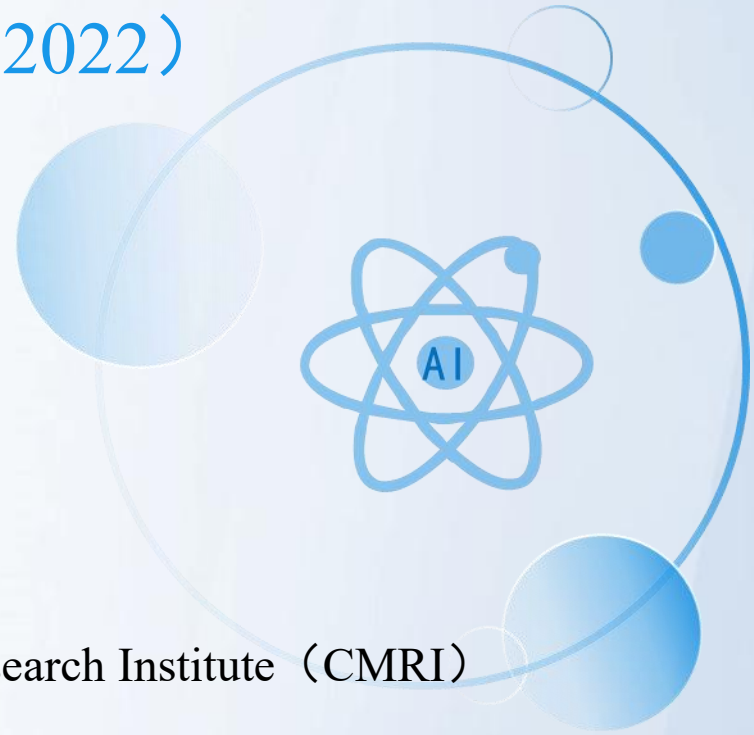
中国移动 China Mobile | 研究院 CMRI

6G Native AI Architecture and Technologies White Paper

（2022）



China Mobile Research Institute（CMRI）

Contents

1 Preface

Artificial intelligence (AI) has developed rapidly in the past decade, which has surpassed the Human Intelligence in modeling the nonlinear laws of big data samples and online accurate decision-making in interaction with the environment, and has achieved great success in the fields of computer vision, natural language processing, and robot control. The reason for the rapid development of AI is, on one hand, the breakthrough of artificial intelligence algorithms represented by deep learning and reinforcement learning; on the other hand, the rapid decline on the cost and popularization of artificial intelligence computing power represented by GPUs.

Since 5G, AI has gradually been widely used in mobile communication networks, such as network configuration optimization at the network management level, resource scheduling optimization at the network element level, and even the physical layer of the air interface. In addition, there are also more and more AI applications on the terminal side. Towards the future, 6G network needs to facilitate the digitalization and intelligentization of thousands of industries, and it needs to provide intelligent services with less latency and better performance than cloud intelligence. For operators, network operation costs need to be greatly reduced, and network operation and maintenance needs to evolve from local intelligent scenarios to high-level network autonomy.

At present, AI applications are mainly based on centralized cloud resources. Cloud servers aggregate large amounts of data, utilize centralized computing power to preprocess them, and train and validate AI models. However, transmitting a large amount of raw data in the network will not only put enormous pressure on network transmission bandwidth and performance indicators (such as latency), but also bring great challenges to data privacy protection. Besides, due to the lack of computing power, algorithms and data, there is still much room for improvement in the intelligent applications on the terminal side.

In the face of the above challenges, it is necessary to introduce native AI capabilities into the network, abandon the patched mode of AI applications, and realize the deep integration of communication connection, computing, data and AI models at the network architecture level, where the distributed computing power and data in the network are fully utilized for the coordination mechanisms between multiple nodes and between terminals and the network, and realization of the integration of distributed and centralized processing. In this way, not only data privacy can be protected, the efficiency of data processing, the timeliness of decision-making and reasoning, and utilization efficiency of network nodes can also be improved. This white paper first introduces the driving forces and application scenarios of native intelligence. The demand for native AI support by 6G network is derived from the current status of intelligent network applications, the requirements on high-level network autonomy, ubiquitous intelligence, high value network services, extreme service experience, and network safety and trustworthiness. Then, the paper elaborates on the definition and scope of native AI, and proposes the deep integration of AI computing power, data, algorithms and network connections. Besides, the new concepts of 6G native AI are introduced including AI service quality (QoAIS), orchestration of AI workflows of its full life cycle, computing and communication integration, and integration of native AI and

digital twins. The new architecture driven by native AI is proposed and described in detail, including data plane, smart plane and extended control plane and user plane, and new technologies are introduced including AI model orchestration, distributed model training, distributed model inference, pre-validation and optimization of digital twins. Finally, the future research directions are prospected.

2 Driving Forces

The application of artificial intelligence technology in 5G networks has promoted the intelligent development of mobile communication networks and vertical industries, but the application mode of patching and plug-in hinders the effectiveness of AI applications. At the same time, the application and exploration of artificial intelligence in all walks of life has put forward requirements for new basic capabilities of future networks. To realize the vision of ubiquitous intelligence, 6G networks need to build native AI capabilities.

2.1 Challenges of 5G intelligent networks

In the 5G era, intelligent network practices requires the integration of AI technologies with the hardware, software, systems, and processes of 5G communication networks, and AI is used to help communication networks realize intelligent planning, construction, maintenance and optimization, so as to improve quality and efficiency and reduce cost. The utilization of AI promotes the technological and systematic transformation of the network itself, enables agile business innovation, and promotes the construction of intelligent networks, including cloud network equipment intelligence, network operation intelligence, and network service intelligence. In 5G network, AI is mainly used for optimization of communication connections and service processes. Although the service cloud has been introduced, network and cloud is loosely coupled since the 5G architecture, protocol functions and processes have been finalized, and only incremental iterations can be made to the existing architecture solutions..

The challenges of 5G intelligent network practices based on patched-on and plug-in AI are as follows:

- The lack of a unified framework leads to the lack of effective verification and guarantee methods for AI performance. The verification of AI application effects is carried out after the fact, so the overall end-to-end process is long and complex, and the intermediate process generally requires a lot of human intervention. The relative large impact on the network makes it difficult to quickly promote AI in the process of applying it to the existing network.
- The plug-in mode is difficult to achieve a fully automatic closed-loop of pre-validation, online evaluation and optimization. AI model training usually requires the preparation of a large amount of training data. In the plug-in mode, it is difficult to collect and label data on the existing network in a centralized manner, and the transmission and storage

overhead is also large, resulting in a long iteration cycle of the AI model, high training overhead, slow convergence, and poor model generalization.

- In the plug-in mode, computing power, data, models and network connections belong to different technical systems, and no standardized interfaces and interaction rules are defined between them. Cross-system collaboration is carried out on the management plane, leading to extra latency of seconds and even minutes and the unguaranteed quality of service.

2.2 New scenarios of 6G ubiquitous intelligence

Native AI refers to the network supporting AI through native design patterns at the architectural level, rather than patched-on or plug-in patterns. The driving forces of the native design pattern are as follows:

- **The network provides ubiquitous AI services:** To realize the vision of ubiquitous intelligence, the 6G network needs to help the digitalization and intelligentation of thousands of industries and realize the on-demand supply of intelligent capabilities " anytime, anywhere ". Compared with cloud service providers, 6G networks need to provide intelligent capability services with higher real-time performance and better performance and at the same time provide federal intelligence between industries to realize cross-domain intelligent integration and sharing. On the other hand, due to the large amount of data in the terminal, the computing power of the terminal is also getting stronger and stronger. Considering the data privacy requirements, the native intelligent collaborative network and the computing power, communication connection and algorithm model of the terminal and other resources are required, such as computing power offloading. , model orchestration, etc., to provide 2C customers with the ultimate business experience and high-value new services.
- **AI provides services for the network:** 6G networks need to achieve a high level of autonomy, security and trustworthiness. At present, the level of network autonomy is not high (the level of autonomous driving network is about 2.2), and it is necessary to introduce native AI capabilities in the network to support the perception and realization of the intentions of operators and users, and realize the self-design, self-implementation, self-optimization, and self-development of the network. Evolution, and ultimately achieve a high level of network autonomy. In addition, in the future, the network will carry more diversified services, serve more application scenarios, and carry more types of data. Therefore, the network will face a large number of new and complex attack methods. The security capabilities based on native AI are embedded in all aspects of the 6G network to realize autonomous detection of threats, autonomous defense or assist in defense.

It can be seen from the above driving force analysis that in addition to meeting basic communication needs, 6G networks also need to consider the integration of computing, data, models/algorithms, etc., that is, 6G needs to be designed through native AI at the architectural

level to meet the Diversified new business scenarios and network autonomous optimization requirements for network AI include AI applied to network optimization and user experience (such as air interfaces rewritten with AI), as well as various AI services required by third parties .

3 Definition and Scope

Deep integration with AI will be considered in 6G design stage, i.e., 6G native AI. Unlike 5G, which treats AI functions as added-on features, 6G native AI will exercise an end-to-end orchestration and control of computing power, data, and models. The key ingredients, such as connections, computing, data and AI algorithms/models are inherently integrated and meanwhile the on-demand orchestrating those key ingredients into wireless, transmission, core network, etc. are supported, which provides the inherent intelligence capabilities required for high-level network autonomy and diversified business needs. That is the native AI capability of 6G, which will make network intelligence more efficient and perform better. At the same time, the network intelligence will also be expanded accordingly, not only accelerating the continuous optimization of network performance, but also providing intelligent service capabilities, which enables the digital and intelligent transformation of various industries. Network intelligence will continue to evolve in the 6G era, promoting a truly intelligent native network.

3.1 Definition of 6G Native AI

6G network native AI is operating within the 6G network architecture, which provides data collection, data preprocessing, model training, model inference, model evaluation and other entire life cycle AI workflows. The key ingredients of AI services, e.g., computing power, data, algorithms, connections, and network functions, protocols, procedures are deeply integrated into the initial design of 6G network. 6G network native AI aims at providing real-time and efficient intelligent services and capabilities for high-level network autonomy, ubiquitous intelligence for industrial users, ultimate service experience for subscribers, and native network security.

3.2 Scope of 6G Native AI

The existing mobile communication network is mainly connection-oriented data transmission, which requires the transmission link guarantee based on QoS (Quality of Service) (such as data rate, delay, etc.). However, the native intelligence is required to implement end-to-end control and orchestration of computing power, models and data. Thus, there are enormous gap towards the network design and implementation and operation. Therefore, it is critical to consider the above distinct requirements at the beginning of 6G network design. On one hand, new concepts such as quality assurance of AI-based service, end-to-end orchestration and computing and communication integration should be introduced, on the other hand, new architecture designs such

as AI data plane, AI intelligent plane, extended control plane and user plane should also be considered.

The quality of AI service assessment and assurance should be built for native AI, and then based on quality of AI service the end-to-end AI life-cycle orchestration will be implemented, including computing power, AI models, data and connections.

Native AI requires deep integration of computing and communication. Considering that the capabilities of native AI are distributed into a large number of network nodes, which are usually restricted in data collection, computing power, bandwidth, and delay, and thus it is vitally important to adopt the co-design of computing and communication resources coordination. In addition, it is necessary to rethink the network architecture, protocols and functions, which are adaptable to air interface transmission and the performance optimization of native AI.

4 New Idea

It is quite complicated to integrate AI with the traditional connection oriented network at the beginning of 6G network design, which requires cross-domain expertise. It is vital to think out of box of the traditional design paradigm and incorporate fresh AI ingredients and concepts. We believe that the assessment and assurance of AI service quality, the orchestration and management of AI life cycle, and the deep integration of AI computing and communication will become the basic concepts of native AI systems.

Facing various industries and scenarios, there are diversified demands for 6G native AI network. The first question, we need to answer, is how to translate user demands into network AI service capabilities? We propose the concept of AI service quality, namely QoAIS (Quality of AI Service), and believe that the network should provide an assessment and assurance for QoAIS. Next, how to evaluate and continuously satisfy QoAIS and implement QoAIS assurance requires the involvement of the management plane, control plane, and user plane. From a management plane perspective, we propose the orchestration and management of the AI life cycle workflow, i.e., the semi-static allocation of network resource, such as computing power, data, algorithms, and connections, to satisfy QoAIS requirements; from control and user plane perspective, it is important to allocate real-time network resource to continuously satisfy QoAIS, in which the deep integration of AI computing and communication is the key.

4.1 Quality of AI service (QoAIS)

QoAIS is a set of metrics profiles of AI service quality assessment and the corresponding assurance mechanisms. [1]. 6G networks will provide inherent AI capabilities, which can serve a variety of intelligent applications, namely AIaaS. Considering that the requirements of different intelligent scenarios towards the quality of AI services are expected to be highly diverse, and thus, a set of indicators is required to express user-level needs and network orchestration and control (including AI models/algorithms, computing power, data, connections, etc.) in a quantitative or hierarchical way.

The native 6G AI services can be categorized into the following types, i.e., AI data, AI training, AI inference, AI validation and etc. Each type of AI service requires a different set of QoAIS. The QoS of the communication service in traditional communication networks mainly considers the performance related to the connectivity, such as delay and data rate (MBR, GBR, etc.). Besides that, the 6G network will introduce a variety of resource dimensions for AI service orchestration and control, such as distributed heterogeneous computing resources, storage resources, data resources, and AI models/algorithms. Therefore, 6G native AI service quality should be evaluated from multiple dimensions of network resources, such as connection, computing, algorithm, and data. At the same time, with the implementation of the "carbon neutrality" and "carbon peak" policies, the global industry of intelligent applications will pay more attention to data security and privacy, and network automation. In the future, performance-related KPIs will no longer be the only indicators to be highlighted, and the requirements for security, privacy, autonomy and resource overhead will gradually play more important roles and become crucial dimensions for evaluating AI service quality. Therefore, from the initial design, the QoAIS indicators needs to consider performance, overhead, security, privacy and autonomy and other aspects as well.

Table 4.1-1： QoAIS indicators of AI training service

Types of AI Services	**Evaluation dimensions**	**QoAIS indicators**
AI training	performance	Performance bounds, training time, generalization, reusability, robustness, interpretability, consistency between loss function and optimization objective, fairness
	overhead*	Storage overhead, computing overhead, transmission overhead, energy consumption
	Safety*	Storage security, computing security, transmission security
	privacy*	Data privacy level, algorithm privacy level
	autonomy	Fully autonomous, partially manually controllable, and fully manually controllable

Note*: Common evaluation indicators between different types of AI services

Among them, the performance bounds are the upper and lower bounds for evaluating the model performance, such as model accuracy, recall rate. Generalization refers to the ability of a pre-trained model adaptive to make predictions on new data. Reusability is the ability of a model to continue to function in case of application scenarios change. Robustness refers to the capability that the model can still work even if the input data is distorted, attacked or uncertain. Interpretability refers to the degree to which a model's internal mechanisms can be understood. Consistency between the loss function and the optimization goal refers to the degree of consistency between the design of the loss function and the optimization goal during the model

training process, for example, whether the number of variables considered in the loss function completely covers the optimization goals. Autonomy refers to the requirements for autonomous operation and manual intervention in the workflow of AI data/training/validation/inference services, reflecting the degree of automation of AI services. Autonomy is divided into three levels: complete autonomy (full automated AI service, without any manual intervention in the whole process), partial manual intervention (some workflows of AI service are automated, while others require manual assistance), all manual controllable (All aspects of the AI service workflows are handled manually).

In addition to the evaluation dimensions shown in the table above, QoAIS can also include application specific performance indicators. Taking channel state information compression as an example [2], normalized mean square error (NMSE) or cosine similarity can be selected as the KPIs for channel recovery accuracy, or link-level/system-level indicators (such as block error bit rate or throughput, etc.) as KPIs reflecting the impact of channel feedback accuracy on system performance. In addition, QoAIS can also include the availability of AI services, the response time of AI services (from the user initiating the request to the first response message of the AI service) and other general evaluation indicators not related to the specific type of AI service.

QoAIS is a key input for the network native AI orchestration and control. The top-level QoAIS will be broken down by network AI management and orchestration system, and then be mapped to the specific QoS requirements for data, algorithms, computing, connections, and etc.

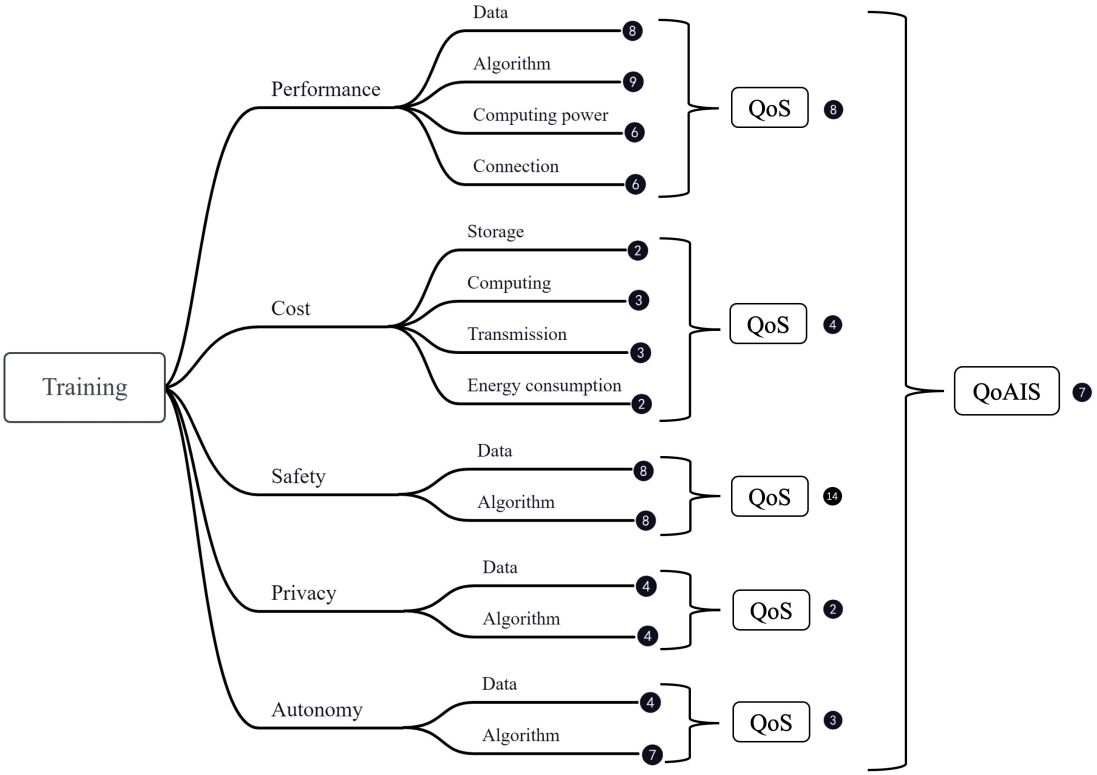


Figure 4.1-1 : QoS indicators decomposed into QoS indicators in each resource dimension

The above figure shows the mapping between each indicator of QoAIS and the corresponding QoS metrics. The overall QoAIS indicators of AI services are decomposed into

QoAIS indicators in each indicator dimension, and further mapped to QoS indicators in each resource dimension, which are guaranteed by the mechanisms from the management plane, the control plane and the user plane's perspective. The QoS indicators in each resource dimension in the figure can be divided into quantitative indicators and leveled indicators (such as security level, privacy level, and autonomy level). For the former category of indicators, the quantification schemes for some indicators are mature or relatively easy to formulate (such as training time, algorithm performance, calculation accuracy, various resource costs, etc.), while there are currently no quantitative evaluation methods for other indicators (such as model robustness, reusability, generalization and interpretability, etc.), as shown in Table 4.1-2. Therefore, it is a key issue to design sufficiently open and inclusive network architecture in the initial stage so that the mature quantitative technologies of the above indicators can be gradually introduced.

Table 4.1-2 : Mapping of AI training service performance QoAIS to each resource dimension

Metric dimension	QoAIS indicator	resource dimension	Quantitative indicators	No quantitative metrics yet
performance	Performance bounds, training time, generalization, reusability, robustness, interpretability, optimization target matching, fairness	data	Feature redundancy, completeness, data accuracy, and data preparation time	Sample space balance, integrity, sample distribution dynamics
		algorithm	Performance index bounds, training time, convergence, optimization target matching degree	Robustness, reusability, generalization, interpretability, fairness
		computing power	Computational accuracy, duration, efficiency	
		connect	Bandwidth and jitter, delay and jitter, bit error rate and jitter, reliability, etc.	

In terms of quality assessment and assurance mechanism, there are still some problems for the QoS mechanism of 5G network, such as coarse service differentiation granularities, long optimization and adjustment period, and inability of radio resource management adaptive to dynamic fluctuation of network and services. Therefore, it is also necessary to consider how to design end-to-end efficient QoAIS mechanisms and procedures when proposing QoAIS indicators for assessing and assuring AI services in 6G networks.

Extended question:

1. After the introduction of AI services in the network, users may have different requirements for security and privacy. The QoS and security design are independent in traditional communication networks, and how to co-design QoS and security for 6G native AI?
2. Currently there are no mature quantitative evaluation methods for some QoAIS indicators (such as model generalization, interpretability, and reusability [3]). How to design an inclusive architecture to accommodate gradually introduced quantitative technologies for such indicators?

4.2 AI lifecycle orchestration management

The AI life cycle refers to the life cycle of an AI workflow in the network, that is, the generation, execution, monitoring, evaluation, optimization, completion, and deletion of an AI workflow. Network Native AI Workflow refers to one or more tasks that the network needs to complete step by step in order to complete an AI service. Currently, AI has a similar end-to-end workflow in various industrial applications [4], which can be divided into four flows: data management, model learning, model validation, and model deployment. Figure 4.2-1 shows a common AI end-to-end workflow.

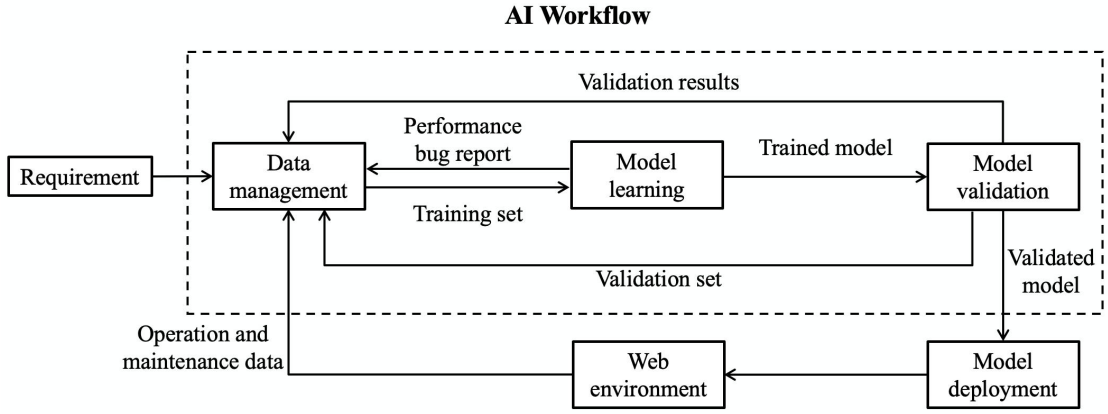


Figure 4.2-1 : Common AI end-to-end workflow

At present, in the practice of 5G network intelligence, most of the AI workflows is implemented offline, independent from the network operating environment, and the silo-development model is adopted between different intelligent applications (that is, for each intelligent application the research and development is carried out independently without resource coordination and sharing), which leads to low efficiency and high cost. The 6G network will provide a complete operating environment for end-to-end AI workflows of various intelligent applications.

Depending on the type of AI service, the AI workflow may include different tasks. Not all of the corresponding workflows are end-to-end. For example, the workflow of AI data service only includes tasks related to data management; AI validation workflows can include both data management and model validation related tasks, or only model validation tasks; AI training workflow can include only model learning, or both data management and model learning, depending on whether the data provided by users is sufficient for the quality requirements. The AI

inference workflow can only include tasks related to model deployment or can include tasks related to data management and model deployment simultaneously. For an intelligent application requesting multiple AI services at the same time, the corresponding workflow may last end-to-end. Figure 4.2-2 shows the relationship between the native AI workflow and AI services in the 6G network.

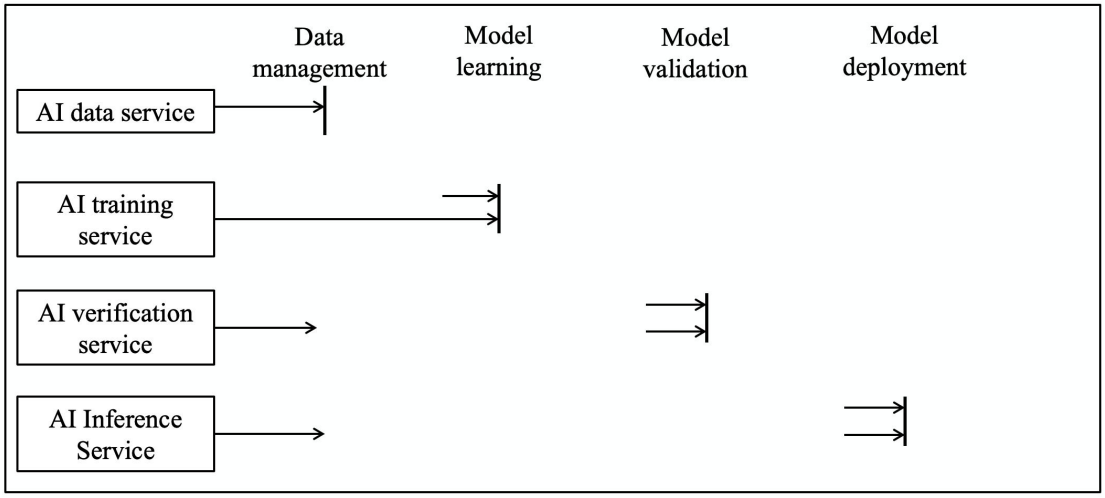


Figure 4.2-2: diagram of native AI services and workflows in 6G networks

The 6G network generates the required workflow and relative tasks for each AI service, and then orchestrates the respective resources (computing, algorithms, data, connections, etc.) for each task in the workflow to ensure continuous assurance of the QoAIS. In this process, the management plane is responsible for collecting performance monitoring data, evaluating QoAIS, and analyzing the impact of the task design and resource management, so as to continuously optimize the schemes and strategies and realize intelligent orchestration management.

Extended question:

1. In order to ensure the continuous satisfaction of QoAIS, is it adequate to rely solely on management plane to orchestrate resources required by the workflow? Does the control plane need to be involved? How do management and control work together?

4.3 Deep integration of AI computing and communication

In order to ensure the continuous achievement of QoAIS, in addition to realizing intelligent orchestration and management of the AI workflows from the management plane, it is also necessary to realize the deep integration of AI computing and communication on the control plane and user plane.

The computing resources in the traditional communication network mainly serve the communication services. The computing resources are integrated in the network equipment processing board, and the computing resources are allocated based on the predefined procedures of the communication services. In contrast with communication services, AI services are computing intensive. In recent years, various processor architectures (GPU, NPU, DPU, TPU, etc.) have been emerging to improve computing efficiency and reduce energy consumption. The key

requirements for computing of native AI services in 6G networks are high computing efficiency, low energy consumption, and low latency. Although the computing efficiency of centralized computing resources in the cloud is high, it is hard to meet low-latency requirements for edge AI applications. The computing power resources of edge nodes are limited, while the scale is large and the real-time performance is better. Therefore, it is preferable to coordinate computing resources between edge and cloud, which is expected to meet computing performance requirements for various AI services.

The edge computing capabilities have been introduced in 5G MEC solutions to provide low-latency services, however the network connection and computing are loosely coupled, and there is room for further improvement in terms of efficiency, deployment cost, security, and privacy protection. For example, in the 5G MEC solution [5], the user plane function UPF in core network can be co-located with the MEC, but they are still two relatively independent systems from logical architecture and control management‘s perspectives. When the adjustments are required for network connection and computing power simultaneously, it is coordinated through the management plane, which leads to relatively large delay. On the other hand, the computing resources deployed in the cloud, edge and device are distributed and heterogeneous. If the coordination is adaptive to dynamic and complex radio environment in a real-time way, it is crucial to provide real-time support from control plane and user plane’s perspective.

Taking mobile networks as an example, there are potential three co-design modes for the deep integration of AI computing and communication illustrated in Figure 4.2-3.

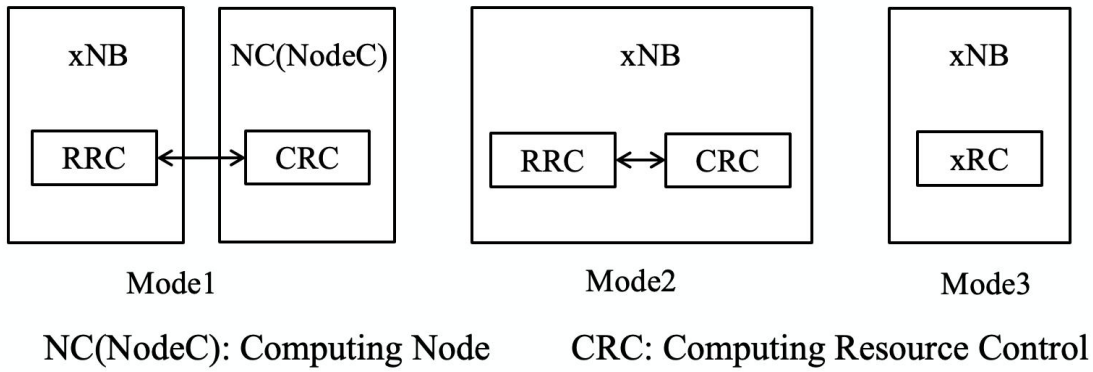


Figure 4.2-3： Three modes of AI computing and communication co-design

Mode 1: A new logical computing unit (NC, Computing Node) is introduced into the mobile network architecture, which is independent from the base station. CRC (Computing Resource Control) in NC interacts with RRC (Radio Resource Control) in xNB control plane through specified interfaces. The advantage of this mode is that it allows inter-vendor deployment between base stations and computing unit vendors, however, relatively long interaction delay may be introduced.

Mode 2: The computing unit as an inner function is built in the base station. The RRC and CRC interact with each other via an internal interface. The advantage of this mode is that better performance may be achieved, and meanwhile, radio communication resources and computing resources are independently controlled and coordinated on demand, which ensures the scalability of RRC and CRC design to some extent.

Mode 3: The logical computing unit is built in the base station and RRC and CRC are merged into a unified resource control entity (xRC), which controls the connection and computing resources at the same time. The advantage is that the control decisions on connection and computing resources can be made at the same time, which leads to the best performance of the coordinated connection and computing resource control. However, it is more complex to design such mechanism and meanwhile scalability issue may come into being.

By implementing deep integration of computing and communication on the control plane, more efficient measures can be provided to continuously achieve the QoAIS targets. The advantage is that when some QoAIS indicator is deteriorating, some policy or scheme can be quickly optimized. For example, if connection bandwidth is restricted and local computing power is sufficient, it is preferable to compress the AI data with more sophisticated algorithms with high recovery accuracy and lower transmission bandwidth; when the connection bandwidth is sufficient and stable, but the local computing power is limited, it is preferable to increase the local computing capabilities by collaborating with surrounding nodes. For user plane, the deep integration of AI computing and communication is mainly reflected in the joint design and optimization of AI computing protocols and communication protocols. In terms of computing protocols, for the same AI computing task, different protocols and configuration parameters may be required for heterogeneous computing resources, which eventually have an impact on computing accuracy and computing time. In terms of communication protocols, various types of data from AI task (such as model intermediate results, model weights, model gradients, etc.) shall be processed optimally considering the instability of bandwidth and channel state, such as the coding and encoding of source and channel. In addition, since the computation and communication of AI tasks are often sequentially processed in time domain, it provides a possibility to allow joint design and optimization.

.

Extended question:

1. How to effectively consider co-design of computing and communication by coordinating management plane and control plane, to achieve a balanced network resource allocation, and better resource and energy consumption efficiency?
2. How to jointly design and optimize the computing protocol and communication protocol of AI tasks from user plane's perspective to meet the performance and overhead requirements at the same time?

5 New Architecture

The integration of AI resource ingredients into 6G network architecture design is the most fundamental feature for 6G native AI. As the three basic ingredients of AI (data, algorithms and computing power) have become as fundamental resources as network connections, the design of the corresponding architecture, interfaces, protocols should be reflected through the entire AI life cycle. In addition to its own internal management, control, processing and transmission, each resource ingredient will also cooperate with others to meet QoAIS requirements. Therefore, unlike

5G network, new data plane, smart plane, and computing plane will be defined in 6G network, and traditional control plane and user plane are expected to be extended as well. The following figure 5-1 shows the logical architecture of the 6G native AI network.

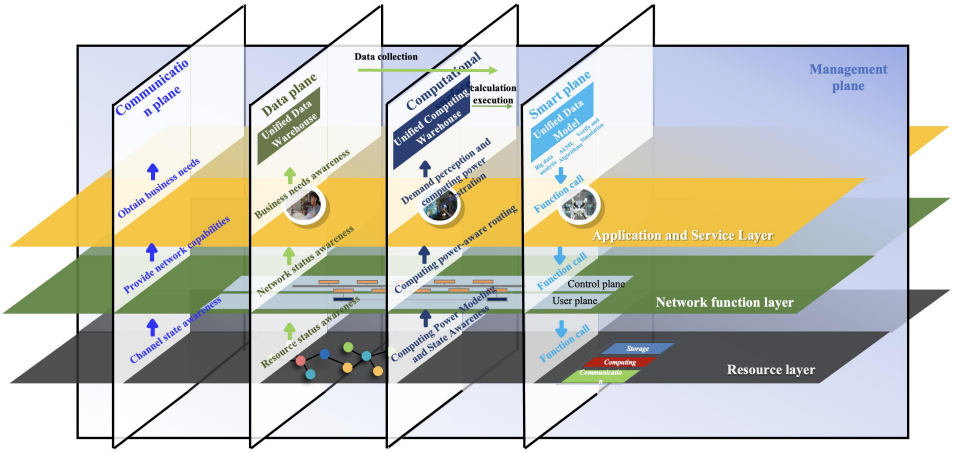


Figure 5-1: logical architecture of the 6G native AI network

Horizontally 6G networks can be divided into resource layer, network function layer and application and service layer ("three layers"). The resource layer provides radio access, computing, storage and other underlying resources, and provides corresponding support and services for the upper layers. In network function layer a specific network function is formed by combining one or more network functions together to provide network service capabilities to application and service layers. The application and service layer provides corresponding support for customers' business needs. In terms of vertical logical architecture, except the "communication plane" that carries traditional communication services, New data plane, computing plane and smart plane are introduced in 6G network. The data plane is responsible for data collection, cleaning, processing and storage in end-to-end network, and provides data services to other layers. The computing plane provides a unified computing power warehouse, perceives computing power requirements, manages computing tasks, provides computing routing, computing modeling, and meanwhile provides computing services for other layers. The intelligent plane provides the operating environment for full life-cycle of native AI. It invokes the services provided by data plane and computing plane, and provides intelligent services for other layers. The management plane provides operation and maintenance of all other layers and planes.

For 6G native AI, the implementation of the following new concepts is mainly reflected in the "three layers" and the intelligent plane, data plane, computing plane and management plane. It is worth mentioning that the control plane and user plane belong to the network function layer. The control plane and user plane traditionally are oriented to support communication services. After the introduction of the data plane, computing plane and intelligent plane, some new service data will be generated from these planes, such as data collected and transmitted on the data plane, input, output, and intermediate data of computing tasks on the computing plane, parameters of AI models on the intelligent plane and etc. In order to provide support such new "service", the control plane and user plane are also extended as well.

This chapter focuses on the data plane, smart plane, and extended control and user planes.

5.1 Data plane

5G network intelligence practices [8] show that the data collection is difficult, and the data quality is hard to guarantee. There are limited specified interfaces for data collection in previous network architecture and protocol design, and the data collection partly relies on implementation, such as deep packet inspection or data probing, and meanwhile it is hard to guarantee data collection in a timely manner. There are some problems for management-based data collection, such as few types of data, long collection period (more than 15 minutes), inconsistent data formats, naming, and calculation from different vendors, and it is also difficult to open southbound network management interface. At the same time, due to the instability of data collection, the transmission loss, the limited storage in the network management equipment, and the difficulty in obtaining labels, the collected data is often not of good quality, such as missing, non-labeled or label errors. Before AI model training, it takes a lot of time and labor cost to pre-process the input data.

Facing the above challenges, “data plane” [7] is introduced to 6G network architecture. The data elements in the data plane will cover internal and external data of the network, including service data, user data, network data, sensing data and so on. The data services include data collection, data preprocessing, data storage, data access, data sharing and collaboration, etc. Basic data services have the following technical characteristics: support for trusted authentication, authorization, access, efficient data storage and management, on-demand data collection, data preprocessing and aggregation, open interface for external access, etc. Figure 5.1-1 shows the logical functional architecture of the 6G network data plane.

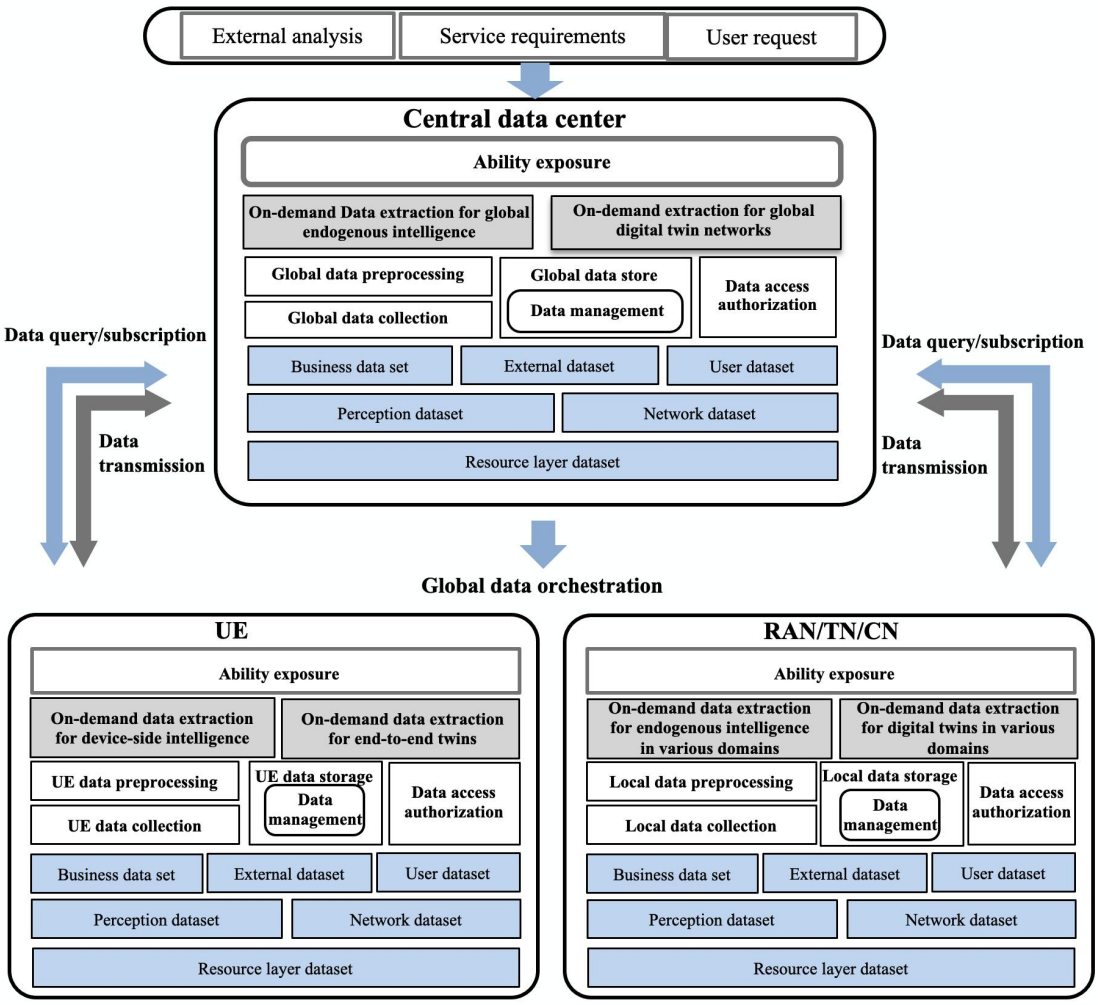


Figure 5.1-1 : 6G network data plane functional architecture

The data plane architecture of the 6G network consists of a central data center and local data centers in each domain, which adopts a hybrid centralized and distributed architecture. The central data center stores the end-to-end global data, and orchestrates the data globally on demand; the local data centers within each domain store and manage the data collected from the local network, and provide data services for various upper-layer applications.

Since the data required and generated by AI services also is processed by data plane, such as training samples, AI model parameters, model intermediate results, model gradients, inference samples, and inference results, etc. Various data services provided by data plane can be invoked throughout the entire life cycle of AI workflows. For example, trusted services provided by data plane can be invoked to guarantee the trustworthiness requirements of AI service defined in QoAIS [9]; computing and transmission overheads can be reduced by invoking on-demand data collection and preprocessing services.

In the 6G network, trustworthiness is expected to be a key requirement for data services [10]. The credibility of data services is mainly reflected in the stages of data collection, data storage, data access, data sharing and collaboration. Regarding data collection, data privacy, fairness, reproducibility and robustness are major considerations. Data privacy is mainly guaranteed by some data processing technologies, such as debias sampling and annotation, data sources tracing,

data anonymousness and differential privacy. Data fairness is mainly assessed via quantitative indicators, such as the correlation coefficient of variables, loss function, complete Cartesian product, etc. The reproducibility and robustness of data collection can be guaranteed by data source tracing.

Extended question:

1. How to support the openness and use of internal data in network entities from the network architecture level?
2. How to support the on-demand data extraction from the network architecture level? Including the type of collected data, the amount of collected data, the collection method, the data preprocessing method, etc.

5.2 Smart plane

In the previous chapter, the new concepts involving the design of the management plane, control plane and user plane have been introduced. These new mechanisms provide a complete operating environment for the entire life cycle of various AI workflows, in order to satisfy the requirements of QoAIS. This complete operating environment is called as the "smart plane" of the 6G network. Figure 5.2-1 shows the functional architecture design of the smart plane of the 6G network.

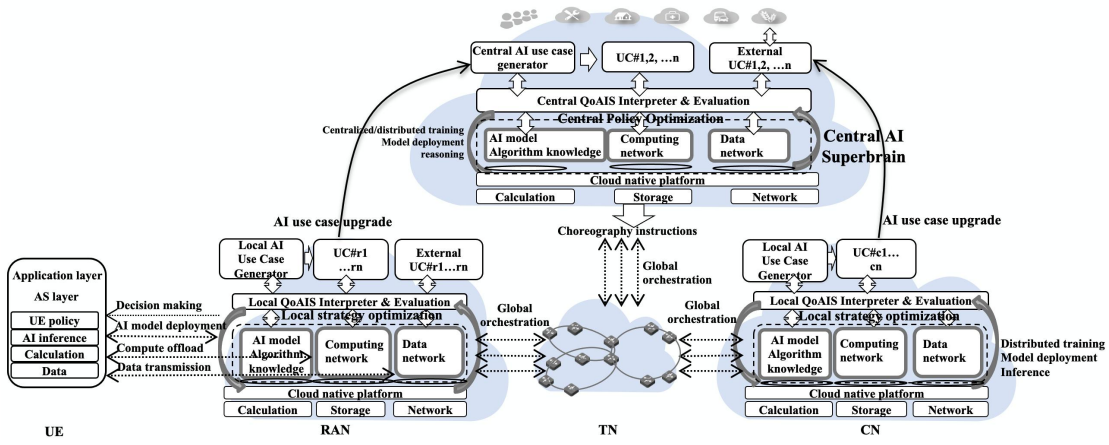


Figure 5.2-1: 6G network smart plane functional architecture

The smart plane of 6G native AI network has the following technical characteristics:

The first technical feature is the self-generation of AI use cases. An AI use case is a one-time AI service request made by a user to the network. An AI use case may involve one or more types of network native AI services (such as AI training, validation, and inference services), which is regulated by the AI use case description. From this description, the network can learn information on intelligent application scenarios, input and output data, model selection, model training, model validation, and decisions on model outputs. The network can generate AI use case descriptions by itself based on its own data analysis or external instructions. The management plane is responsible for managing all AI use cases, scheduling and implementing AI use cases, generating the

corresponding AI services, AI workflows, and QoAIS requirements, and provisioning network resources (including data, algorithms, computing power, connections, etc.). Figure 5.2-2 shows the logical relationship between AI use cases, AI services, AI workflows, and AI tasks.

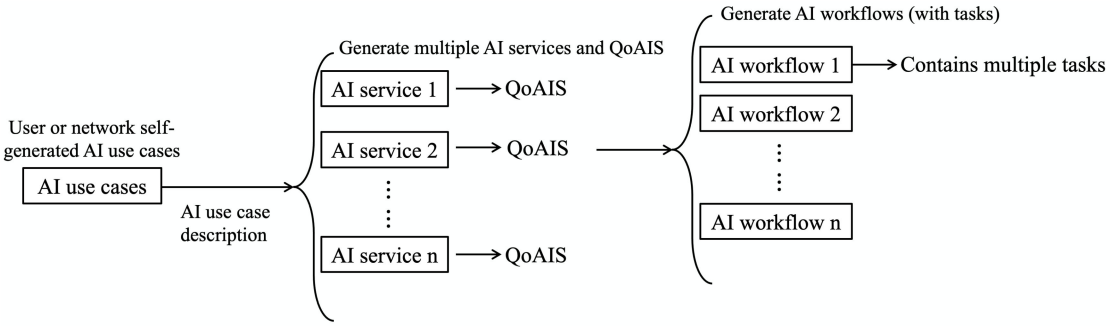


Figure 5.2-2: The logical relationship between AI use cases, AI services, AI workflows, and AI tasks

Second, QoAIS generation. QoAIS is used for quality evaluation for native AI services in the network. An AI service corresponds to a set of QoAIS, and the QoAIS corresponding to an AI use case is composed of the respective QoAIS of all AI services it contains. When the network receives an AI use case, it needs to know the QoAIS requirements corresponding to the use case, so that it can be decomposed into the specific requirements for the orchestration, scheduling, and control of various network resources. The AI use case description can be imported externally and can be generated internally, e.g., AI use cases may be generated based on upper-layer intent information.

Third, the entire life cycle of AI workflow is carried within the network. Various AI workflows can be generated by the network management plane, including data collection, preprocessing, data expansion, and data analysis; model selection, training, parameter adjustment; model verification, integration, monitoring, and updating, etc. And then the required resources are orchestrated, monitored, and optimized to meet QoAIS requirements. In such scenarios high level automation is required without human intervention.

Fourth, the management plane, control plane and user plane collaborate with each other to ensure the continuous achievement of QoAIS, which is mainly achieved through the orchestration and control of the three key ingredients of AI (algorithm, computing power, data) and network function (connection). The management plane is responsible for resource scheduling in the initial stage and non real-time resource allocation adjustment. The control plane and user plane perform real-time QoS assurance according to the dynamic changes of the network environment.

Fifth, the combination of AI centralized and distributed architecture. The central AI supercomputer has sufficient computing power, massive storage capacity. It is suitable for intelligent applications with large scale models, high performance requirements, and non real-time requirements. The central AI unit in each domain of the wireless, transmission, and core network acts as a centralized AI engine in the respective domain, and is responsible for the AI use cases that can be completed in the local domain. The edge nodes distributed in each domain have limited computing power and storage, and will support intelligent applications with high real-time requirements through cooperation between network entities. When the QoAIS of an AI use case in the local domain cannot be achieved within the single domain (such as lack of feature data of

other domains, lack of computing resources), the use case will be escalated to the central AI supercomputer and achieved through global resource orchestration. This hybrid architecture can relieve the performance pressure caused by a single centralized architectures .s

5.3 Extended control and user planes

The control plane and user plane of the existing mobile communication network are designed to meet the quality requirements of traditional communication services (including voice, data packet transmission, streaming media, etc.), and their main purpose is to provide connections for data transmission, support user mobility, and ensure service experience. In terms of resource types, dedicated computing resources are used, and the demand for computing and storage resources is not high. Unlike traditional communication services, AI services are data-intensive and computing-intensive services. New resource dimensions will be introduced for the native AI services including heterogeneous computing and storage resources, new computing tasks, as well as the AI data required and generated by AI services. It is necessary to design a management and control mechanism for new dimension resources, and at the same time, it is important to design an efficient user plane mechanism for the input, output and in-process data of AI services, that is, AI services will become a special user of 6G networks. These will greatly expand the control plane and user plane in traditional mobile communication networks.

We call the new control plane and user plane with extended protocols and procedures to support QoAIS as AI Control Plane (AI CP) and AI User Plane (AI UP) respectively. Table 5.3-1 shows the comparison of AI CP and AI UP with CP and UP in traditional mobile communication networks.

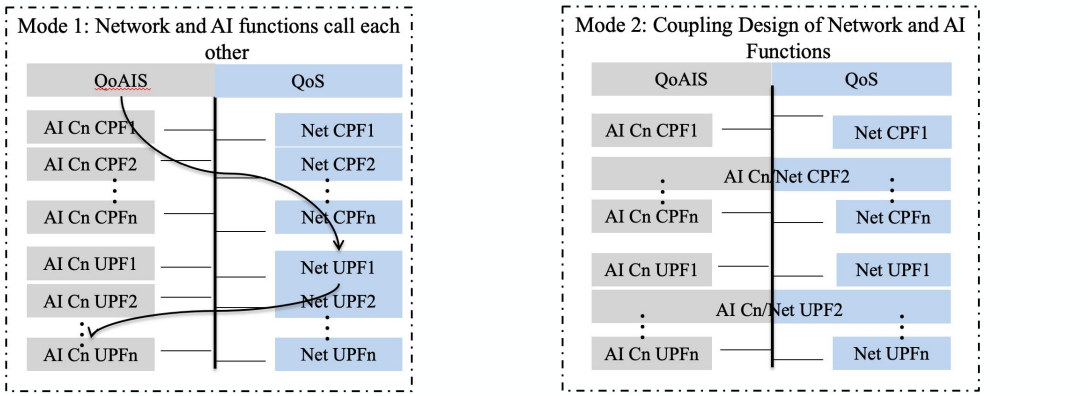
Table 5.3-1: Comparison of AI CP and AI UP with CP and UP in traditional mobile communication networks

traditional communication business	**Native AI Services**				
Connection	Connection	Computing power	Algorithm	Data	Multidimensional resources
NF CP	control mechanisms for AI connections	control mechanisms for AI computing power	control mechanisms for AI algorithm self-optimization	AI on-demand dynamic data collection and processing control mechanism	AI CP

NF UP	transport mechanisms for AI data	execution mechanisms for AI computing tasks	process flow for AI algorithm self-optimization	processing mechanisms for the data AI needs	AI UP

The reason why the traditional control plane and user plane are often end-to-end since traditional communication services usually involve terminals, wireless, transmission and core network domains. However, the workflows of AI services no longer last end-to-end. Instead, the control plane of AI is responsible for control multiple dimensional resources to complete a specific task rather than end-to-end communication, and similarly the user plane of AI is composed of data processing on multiple dimensional resources.

There is an enormous gap between QoAIS and QoS of communication services; for example, AI data may include training samples, inference results, model parameters, intermediate calculation results of training/inference, model gradients, etc. And the transmission mode, data type, data volume of AI data is quite different from those of data from communication services. In addition, the impact caused by radio channel changes, user mobility, and user distribution are different between AI services and communication services. It is not known yet whether the existing control and transmission mechanism of communication networks still applicable. It may be necessary to design a special connection control mechanism and data transmission protocol for AI services, or it may be feasible to take advantage of the same functional module to serve both traditional communication services and AI services. Figure 5.3-1 shows two possible modes between AI connection and traditional connection regarding control mechanism and data transmission protocol.



AI Cn CPF： AI-connected control-related functions
AI Cn UPF： AI-connected data transfer-related functions
Net CPF： Control-related functions of traditional communication connections
Net UPF： Data transfer related functions of traditional communication connections

Figure 5.3-1: The relationship between AI connection and traditional connection in control mechanism and data transmission protocol

Extended question：

1. What are the pros and cons of providing AI services via the application layer? Is it capable to provide AI services by utilizing the existing control and user plane of communication network? If not, what areas need to be improved or innovated?
2. What is the relationship between connections, computing power, algorithms, and data resources regarding the data processing and control mechanisms?

6 New technologies

For 6G native AI network, each stage of AI life cycle, including data collection, model selection, model training, model inference, performance evaluation and optimization, requires corresponding technologies support. Unlike cloud AI with centralized deployment, a large scale distributed nodes in a wireless network usually require collaboration to complete AI tasks. It is indispensable to design specific mechanisms considering computing and communication integration. Depending on the specific AI task, the collaboration approach and integration mechanism might be different. On the other hand, in order to ensure that AI training and inference does not have a negative impact on network performance, it is particularly important to setup a digital twin network for a physical network. New techniques related to model orchestration, training, and inference will be introduced in the first three sections, and then the interaction between native AI and digital twin will be described in the last section.

6.1 AI model selection and finetuning

The AI model orchestration is one of the key technologies for the orchestration and management of the entire AI life cycle. During the entire life cycle management of AI models, including model training, model validation, model inference, and model transfer , it is important to select the appropriate baseline model considering model complexity, inference overhead, retraining overhead, especially in the case of the dynamic changes in computing power and communication resources for wireless network.

Through theoretical analysis and experiments, it is found that the depth and complexity of the AI model have a significant impact on the performance of the model. The more complex the AI model, the greater the probability of a more optimal solution tends to be [11]. But the more complex the model, the more overhead required to retraining. In order to reduce training overhead (such as training data collection, computing power), the common practice is to introduce model retraining, that is, select a baseline model with good performance well trained on a large data set of the source domain, and then use the target domain data to retrain the baseline model to learn the distribution bias between the source and target domains.

A key problem that needs to be solved for model retraining is to choose a suitable baseline model structure and weights. To apply AI model selection and retraining to wireless communication systems, the following factors need to be considered:

1. Due to the difference caused by nonlinear power amplifier of transceivers in real system, complex air interface fluctuation, user distribution and movement and other factors, it is vitally important to introduce retraining to improve the adaptability of AI model. However, it is difficult to accumulate adequate channel state information, and thus it is necessary to utilize small samples to retrain the baseline model to achieve better transfer performance in the target domain.
2. It is also necessary to consider retraining the baseline model on distributed nodes, or even on terminals. Due to limited computing resources and power consumption, it is important to choose a suitable baseline model while ensuring low retraining overhead.

The performance and cost comparison are listed in Table 6.1-1 under three training setting . Among them, basic learning is based on random weights of a basic model; transfer learning uses a large amount of data in the source domain to train the weights of the basic model; similarly, meta-learning [12] also uses the training samples in the source domain, and it aims at learning meta-knowledge, i.e. transferable characteristics, rather than simply fitting the training samples .

Table 6.1-1: Comparison of performance and cost of under three training setting

	basic learning	**transfer learning**	**meta-learning**
source domain model	Basic model random weights	Learning model structure and weights based on large amounts of data	Learning from large amounts of data Meta knowledge
source domain model training overhead	without	Larger overhead (1x)	Max (10 times)
target domain performance	Small samples: low performance Large samples: high performance	Large distribution differences: medium performance Small distribution differences: high performance	High performance for small samples
Retraining overhead	high cost	low overhead	low overhead

From the above table, it is important to extend the model orchestration function of AI retraining to implement the selection of AI baseline models. Specifically, in order to enable optimal selection of the AI baseline model training method, the data distribution difference between the source domain and the target domain should be evaluated first. For example, if there is a large distribution difference and sufficient training resources, it is preferable to consider meta-learning to train the baseline model on the source domain dataset. Moreover, the data distribution of the source domain dataset can be analyzed to arrange the transfer feature dataset to speed up the baseline model training. If the distribution differences are small and training resources are relatively limited, it is preferable to consider transfer learning to train the baseline

model. In addition, the performance of the baseline model with the best performance in the source domain is not necessarily the best candidate in the target domain. We can also leverage the data characteristics and distribution of the target domain to enable optimal selection of the baseline model.

6.2 AI model training with terminal and network collaboration

Distributed AI model training refers to utilizing distributed computing resources deployed on the cloud, edge, and terminals to perform AI model training, which can improve computing resource utilization, enhance model performance, and protect data privacy. As described in Section 4.3, with the deep integration of computing and communication, the coordination of distributed heterogeneous computing resources is considered to adapt to the dynamic and complex communication environment in a real-time way, which leads to real-time support from the control plane and user plane, see chapters 5.3 Extended Control and User Plane.

For data-driven AI model training, traditional methods include model training based on centralized computing power and data, or model training based on distributed parallel computing. The latter is usually processed in a computer cluster, i.e., splitting data or model into different computing nodes for parallel computing, and producing the final result on centralized computing nodes by aggregation processing [13]. Because the computer cluster network condition is relatively stable and reliable, and the training dataset distribution is usually known, the theoretical modeling is relatively easy, and the performance of model training is easy to guarantee. However, in the mobile communication network, there are some complex issues, e.g., unstable channel quality, user mobility, and non-IID distribution of training data. Therefore, it is unavoidable to design more complex distributed model training schemes in mobile network than in a computer cluster.

At present, many technical frameworks of distributed AI model training have been proposed in industries and academia, such as (layered) federated learning [14], group learning [15], multi-agent learning [16], and model segmentation-based learning [17][18] et al. However, most of research is based on certain theoretical or ideal assumptions, rather than complex network environments. In this case, can the performance of model training be guaranteed? Is the communication resource overhead and efficiency acceptable? These are issues to be studied.

It is believed that when training a distributed AI model between terminals and base stations in a wireless network, a large volume of intermediate data will be generated during the training process, and as a consequence, the radio resources will be frequently occupied. In addition, air interface transmission delay and bit error rate will degrade the training results. In order to ensure the convergence of the model and meanwhile improve the utilization of radio resources of the air interface, it is a worthwhile to introduce a higher-order model learning algorithm with higher efficiency [19][20][21]. Since model learning algorithms of different orders (zero-order, first-order stochastic gradient descent, second-order Newton method, etc.) have their own

advantages and disadvantages in terms of training speed and resource overhead, it is beneficial to dynamically adjust the learning algorithm according to the wireless channel state. Figure 6.2-1 is a diagram of the dynamic selection of various learning algorithms, and Figure 6.2-2 shows the functional interaction designed to introduce this dynamic selection mechanism.

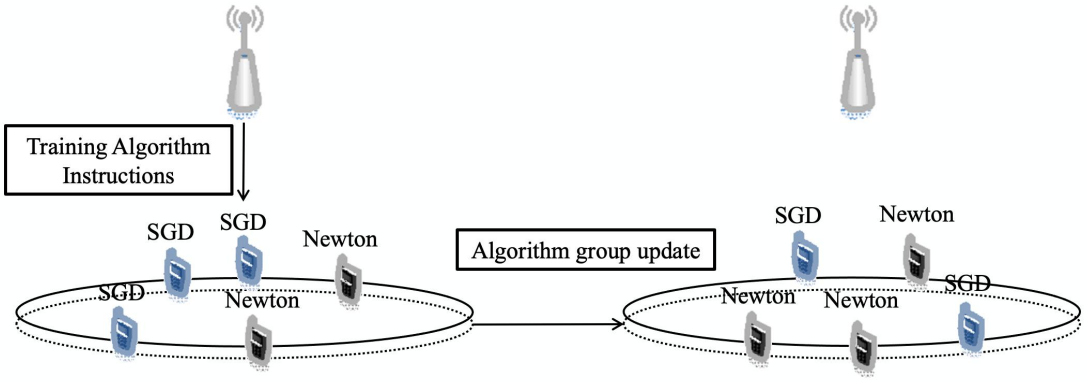


Figure 6.2-1: Schematic diagram of the dynamic selection of various learning algorithms

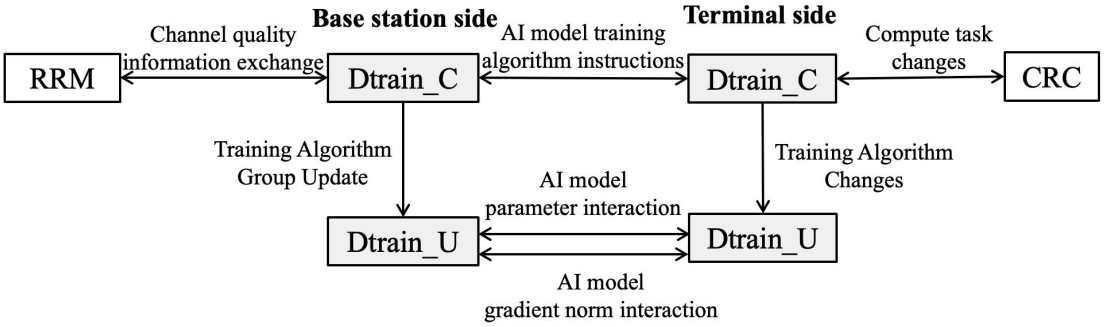


Figure 6.2-2 : Functional interaction of dynamic selection mechanisms of various learning algorithms

As mentioned in 5.3, a new control and data transmission protocol is required to be introduced for AI connection in air interface, which are represented by Dtrain_C and Dtrain_U respectively in the above figure. Among them, Dtrain_C is the control function entity responsible for collaborate the terminals and the base station to perform AI model training. In the aforementioned scenario, this entity dynamically adjusts the model learning algorithm according to the change of the channel condition. Dtrain_U is the service plane functional entity responsible for the AI model training collaboration between the terminal and the base station. It includes a dedicated protocol stack required to carry information such as model parameters, gradients or gradient norms between the base station and the terminals. Since the transmission scheme of the above-mentioned information data is different from that of traditional communication services, and the reliability requirements are also different, it is necessary to re-design the air interface transmission protocol accordingly.

Extended question:

1. The technical improvement by considering both network connection characteristics and AI model training characteristics can theoretically improve the feasibility and performance of 6G native AI network, but it does not fundamentally change the data-driven paradigm. Is it possible to explore a new model-driven training paradigm, so as to implement self-growth of algorithms/models?

6.3 AI model inference with terminal and network collaboration

Distributed AI model collaborative inference refers to the utilization of distributed computing resources on the cloud, edge, and terminal to perform AI model inference, which can improve the utilization of computing resources, compensate insufficient computing power on terminals, and protect data privacy.

Collaborative inference based on model partitioning is a distributed collaborative inference framework in wireless networks proposed by the industry in recent years. When a terminal needs to complete a model inference task and its own computing power is not sufficient, it can request the assistance of the computing resources from the network side to jointly complete inference task. The decision that needs to be made is how to split the model, for example, for the following deep neural model in Figure6.3-1, the beginning two layers are split from the remaining layers, which means the left part of the neural model is running on the terminal side, and the right part is running on the network side.

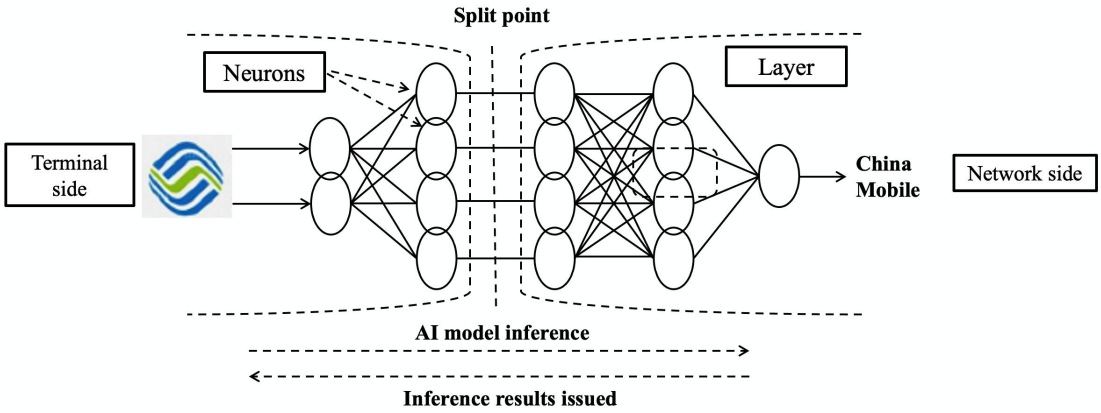


Figure 6.3-1: Diagram of end-to-end cooperative inference

Latency and accuracy are two important performance indicators for AI inference. The decision on the model split point will affect the computation overhead of the terminal side and the base station side, as well as the amount of data that needs to be transmitted over the air interface. Therefore, the factors that should be taken into account may include model splitting point, terminal computing resource allocation, air interface radio resource allocation, and base station computing resource allocation, etc., and thus it is important to coordinate scheduling of communication and computing resources timely. Especially when there are multiple

heterogeneous computing resources on both the terminal and the base station, the allocation of the computing resources will have a direct impact on the inference delay. Therefore, it is necessary to consider how to decide on the model split point, the computing resource allocation scheme on the terminal side and the base station side at the same time, and make timely adjustments as the network condition changes, so as to ensure the continuous achievement of inference target. It can also be considered to take advantage of a better decision-making scheme via reinforcement learning, and the relevant design is shown in the following table.

Table 6.3-1: Design of variables related to end-to-end collaborative inference scheme based on reinforcement learning

variable	**base station side**	**terminal side**
condition (State)	The remaining allocatable computing power of various types of computing resources in the base station, the transmission bandwidth between computing units (optional), the remaining allocatable uplink and downlink air interface channel transmission resources on the base station side, and the terminal uplink channel quality	The remaining allocatable computing power of various types of computing resources in the terminal, the transmission bandwidth between computing units (optional), and the quality of the terminal's downlink channel
Action	The base station side is responsible for calculating some model parameters, uplink and downlink bandwidth allocation, and base station side computing resource allocation	The terminal side is responsible for the calculation of some model parameters and the terminal side computing resource allocation
Reward	Base station side inference energy consumption, etc.	Inference performance indicators, terminal inference energy consumption, etc.

Figure 6.3-2 shows the functional interactions for this mechanism. As mentioned in 5.3, a new control mechanism and data transmission protocol needs to be introduced for AI connection, which are represented by Dinfer_C and Dinfer_U respectively in the following figure. The control entity Dinfer_C can dynamically decide on the model split point, the joint allocation of wireless resources and computing resources according to the changes of radio resources and computing resources, and the data transmission entity Dinfer_U is responsible for calculating and transmitting the intermediate results between the base station and the terminals.

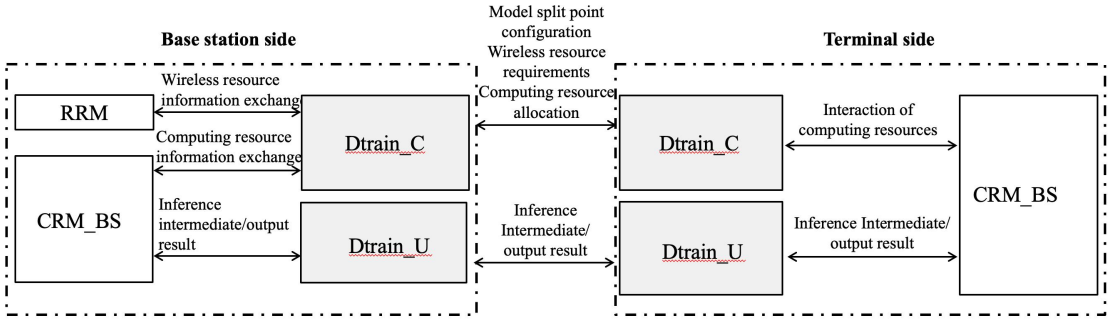


Figure 6.3-2: Functions and interactions of the end-to-end collaborative inference scheme

Extended question:

1. For AI services (such as inference) that require low latency and high reliability, how to precisely allocate computing resources? Whether it is possible to model both heterogeneous computing resources and connection QoS to achieve optimal resource allocation?

6.4 AI performance pre-validation based on digital twin

A digital twin network is a network composed of physical network entities and their digital twin entities, and real-time interaction can be performed between the physical entities and the digital twin entities. The twin digital entities corresponding to the physical entities can be constructed by data collection and simulation. In this system, various network management and applications can utilize the digital twin to efficiently analyze, diagnose, simulate and control the physical entities based on data and models [6][7].

The model validation in AI model life-cycle management utilizes the validation dataset to select the appropriate trained model, but the validation dataset and the training dataset are usually distributed in the same way. How to improve the generalization performance of the model in the digital twin is one of the key technical problems to be solved. Generating samples in more scenarios in digital twin can reduce the overhead of data collection, and meanwhile, improve generalization performance by data augmentation to diversify data samples.

Conditional adversarial generative network (CGAN) [22] can dynamically generate environmental models that conform to specific distributions due to dynamically changing environmental conditions. The environmental models may include user distribution models, radio channel models, user service models, network state models, network resource allocation models, etc. As shown in Figure 6.4-1, a CGAN is introduced into the physical network, and the random sequence and certain semantic environmental conditions are used as input, and the Nash equilibrium is achieved through the adversarial training of the generation model and the discriminative model. The generation model generated by the adversarial training is sent to the digital twin, where more environmental data can be generated by selectively changing the environmental conditions to conforms to specific distributions.

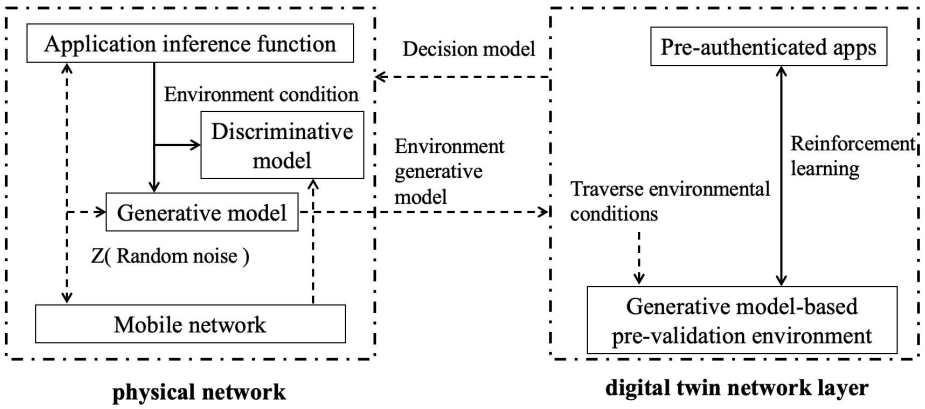


Figure 6.4-1: Model performance pre-validation process based on digital twin

In addition to introducing a CGAN to generate diversified data samples, in order to achieve pre-validation of more scenarios, the reinforcement learning can be introduced to search data sample space. Considering that the state and action space may be very large, and the cost of exploring all sample space is high. On one hand, a distributed search agent can be introduced to speed up the search speed, and on the other hand, an efficient sample space search algorithm needs to be introduced, which may consider the feedback of the physical network. In addition, static or semi-static environment models can be introduced in some dimensions to reduce the exploration cost of the sample space.

In addition, a two-way closed-loop optimization mechanism is introduced between the physical network and the digital twin. The pre-validation applications in the digital twin collect abundant samples via the interaction with the pre-validation environment, and then train the AI model to generate the required decision models for the physical network. In addition, the decision model may generate deviations, and passes them to the digital twin for further correction in the pre-validation environment .s

What needs to be further considered is that the GAN is currently mainly used in the field of image processing, and the indicators of model convergence are used to measure the diversity and restoration degree of image (such as IS, FID). Thus, it is necessary to design indicators suitable for the distribution characteristics of mobile network samples.

7 Summary and Outlook

From 5G network intelligence practice, it is found that there are many shortcomings for external AI. However, since the 5G architecture, protocol and procedures have been finalized, we can only make incremental improvements on the existing solutions. At the same time, in 6G era, intelligent scenarios will be more extensive. In addition to serving high-level autonomy and network performance optimization itself, and providing the ultimate service experience to customers, it will also help the digital-intelligence transformation of thousands of industries. The intelligent applications will be greatly enriched, and the demand for intelligent service

performance will be multi-dimensional. All of these require a paradigm shift from external AI to native AI.

The 6G native AI requires new concepts, new architectures, and new technologies. In terms of new concepts, this white paper proposes that the 6G network have an evaluation system and a closed-loop guarantee mechanism, i.e., QoAIS. Under the guidance of QoAIS, multiple dimensional resources, such as computing, data, algorithms, connections can be coordinated to serve full life cycle of AI workflows including data collection, data preprocessing, model training, model inference, and model evaluation. At the same time, this white paper also proposes to integrate native AI and digital twin, the AI models and workflows can be pre-verified in digital twin.

In terms of new architecture, this white paper proposes that the 6G network will add new data planes and smart planes, and meanwhile greatly expand the traditional control plane and user plane. Among them, the data plane will provide basic data transmission services for native AI and digital twin, the intelligent plane will provide a complete operating environment for the full life cycle of AI workflows, and the extended control plane and user plane is proposed to deeply integrate communication and computing. In terms of new technologies, this white paper lists key technologies in model orchestration, training, inference, and the integration of native AI and digital twin.

At present, the industry has gradually reached a consensus on the requirements, concepts and scopes of native AI in 6G networks. The network architecture and key technologies of native AI are still under active research and discussion. Designing an intelligent native network architecture for the rich intelligent applications in 6G era requires not only a deep understanding of traditional mobile communication networks, but also an accurate grasp of the requirements of various potential intelligent services in the future, as well as in-depth understanding of full life cycle AI.

To this end, we jointly launched the 6GANA (6G Alliance of Network AI) forum in December 2020 with 18 members from operators, equipment vendors, Internet service providers, and universities. 6GANA is positioned as a global forum, focusing on the continuous exploration and promotion of 6G AI-related technologies, standardization and industrialization. It aims at conducting joint research through the entire ecosystem, including ICT (such as chip manufacturers, network infrastructure providers, mobile network operators), vertical industries, AI service providers, AI solution providers, AI academic and other stakeholders. A consensus has been formed to promote AI to become a new capability and service for 6G networks [23]. 6GANA TG2 is a working group under 6GANA responsible for the study of network architecture. It will identify the basic technical requirements of 6G network native AI, study the key enabling technologies, its impact on 6G network architecture, and its impact on standardization, and build an overall framework for 6G native AI network. Facing this goal, TG2 members have comprehensively collected and summarized ten key technical issues that are widely discussed by the industries in December 2021, and formed Ten Questions about 6G Native AI Network Architecture. The content of this white paper provides reference answers to some of the key technical questions in "Ten Questions about 6G Native AI Network Architecture".

Finally, we propose that all partners in the industry chain work together to innovate, focusing on the following key technical issues to conduct in-depth research and extensive discussions:

- What aspects will be reflected in the quality requirements (QoAIS) of AI services for diversified intelligent applications in the future? What new evaluation dimensions will emerge compared to traditional QoS? How to support the generation and evaluation of the above indicators from the network architecture?
- In order to ensure the continuous achievement of QoAIS, how to integrate or collaborate different resource dimensions (data, computing power, algorithms, and connections) from the management plane, control plane, and user plane's perspectives?
- How to support the openness of internal network data from the network architecture level? How to support the on-demand dynamic extraction and processing of data by native AI from the network architecture level?
- Can we reuse the control and user plane protocols of traditional communication for AI services? What needs to be improved?
- What is the relationship between native AI and digital twin? How to support the deep integration of the two from the network architecture?
- If a variety of native AI technologies is used without manual intervention and results in network problems, how to trace and recover the problem?
- How to ensure the credibility of AI model? If a model works well in evaluation stage, some problems to cause poor performance is encountered during inference stage , how to identify the problems in time, and how to handle exceptions?

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
QoS	Quality of Service
QoAIS	Quality of AI Service
AIaaS	AI as a Service
NMSE	Normalized mean square error
KPI	Key Performance Indicator
GPU	Graphics Processing Unit
NPU	Neural-network Processing Unit
DPU	Data Processing Unit
TPU	Tensor Processing Unit
MEC	Mobile Edge Computing
UPF	User Plane Function
CPF	Control Plane Function
RRC	Radio Resource Control
CRC	Computing Resource Control
xRC	x Resource Control
CP	Control Plane
UP	User Plane
Dtrain_C	Distributed Training ControlPlane Unit
Dtrain_U	Distributed Training UserPlane Unit
CGAN	Conditional Generative Adversarial Networks

Authors

This whitepaper is co-authored by:

Future Lab, CMRI: Juan Deng, Gang Li, Qingbi Zheng, Zirui Wen, Chenkang Pan, Qixing Wang, Guangyi Liu

AI Center, CMRI: Guangyu Li, Haitao Cai, Yanping Liang, Peng Zhao, Li Yu

Reference

[1] 刘光毅,邓娟,郑青碧,李刚,孙欣,黄宇红.6G 智慧内生：技术挑战、架构和关键特征[J].移动通信,2021,45(04):68-78.

[2] Wen C K, Shih W T, Jin S. Deep learning for massive MIMO CSI feedback[J]. IEEE Wireless Communications Letters, 2018, 7(5): 748-751.

[3] Liu H, Wang Y, Fan W, et al. Trustworthy ai: A computational perspective[J]. arXiv preprint arXiv:2107.06641, 2021.3GPP TS 38.323, “NR; Packet Data Convergence Protocol (PDCP) specification.”

[4] Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges[J]. ACM Computing Surveys (CSUR), 2021, 54(5): 1-39.

[5] 马洪源，肖子玉，卜忠贵，赵远.5G 边缘计算技术及应用展望[J].电信科学,2019,35(06):114-123.

[6] Deng J, Zheng Q, Liu G, et al. A Digital Twin Approach for Self-optimization of Mobile Networks[C]//2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2021: 1-6.

[7] Liu G, Li N, Deng J, et al. 6G Mobile Network Architecture-SOLIDS: Driving Forces, Features, and Functional Topology[J]. Engineering, 2021.

[8] 中国移动：中国移动自动驾驶网络白皮书[R/OL].(2021)[2021-11-28]. https://www.sohu.com/a/492610271_121015326

[9] Li B, Qi P, Liu B, et al. Trustworthy AI: From Principles to Practices[J]. arXiv preprint arXiv:2110.01167, 2021.

[10] Toreini E, Aitken M, Coopamootoo K, et al. The relationship between trust in AI and trustworthy machine learning technologies[C]//Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020: 272-283

[11] Choromanska A, Henaff M, Mathieu M, et al. The loss surfaces of multilayer networks[C]//Artificial intelligence and statistics. PMLR, 2015: 192-204.

[12] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning. PMLR, 2017: 1126-1135.

[13] Shukur H, Zeebaree S R M, Ahmed A J, et al. A state of art survey for concurrent computation and clustering of parallel computing for distributed systems[J]. Journal of Applied Science and Technology Trends, 2020, 1(4): 148-154.

[14] Liu L, Zhang J, Song S H, et al. Client-edge-cloud hierarchical federated learning[C]//ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020: 1-6.

[15] Warnat-Herresthal S, Schultze H, Shastry K L, et al. Swarm Learning for decentralized and confidential clinical machine learning[J]. Nature, 2021, 594(7862): 265-270.

[16] Xu X, Li R, Zhao Z, et al. Stigmergic Independent Reinforcement Learning for Multiagent Collaboration[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[17] Shao J, Zhang J. Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems[C]//2020 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2020: 1-6.

[18] Teerapittayanon S, McDanel B, Kung H T. Distributed deep neural networks over the cloud, the edge and end devices[C]//2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017: 328-339.

[19] Li T, Sahu A K, Zaheer M, et al. Feddane: A federated newton-type method[C]//2019 53rd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2019: 1227-1231.

[20] Hua S, Yang K, Shi Y. On-device federated learning via second-order optimization with over-the-air computation[C]//2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall). IEEE, 2019: 1-5.

[21] Ghosh A, Maity R K, Mazumdar A, et al. Communication efficient distributed approximate Newton method[C]//2020 IEEE International Symposium on Information Theory (ISIT). IEEE, 2020: 2539-2544.

[22] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[23] 6GANA Whitepaper

Digital Twin ; Ubiquitous Intelligence