SURVEY OF COMPUTING POWER NETWORK

Qianying Zhao[1], Bo Lei [2], Min Wei [3]
[1]China Telecom Research Institute, Beijing, China, zhaoqy50@chinatelecom.cn, [2] China Telecom Research Institute, Beijing, China, leibo@chinatelecom.cn, [3] China Telecom Research Institute, Beijing, China, weim6@chinatelecom.cn

NOTE: Corresponding author: Qianying Zhao, zhaoqy50@chinatelecom.cn

***Abstract** – In September 2021, ITU-T released the first international recommendation for computing power networks (CPNs), ITU-T Y.2501 jointly led by China Telecom, China Unicom and Huawei, and started the Y.2500-Y.2599 series of recommendations for computing power networks. The computing power network is a new type of network that integrates multilevel computing resources, realizes the efficient coordination of the cloud, edge and network, improves the utilization efficiency of computing resources and provides users with the optimal computing and network resource services. Since the first proposal of the computing power network in 2019, it has made breakthroughs in its research and standardization work. As a new field of international standards, also a new field of the communication industry, this paper gives a comprehensive analysis of its definition, framework, service model, key technologies and current progress and challenges, summarizing the research direction of future computing power networks which will have strong implications for the research and implementation of the computing power network.*

***Keywords** – Cloud computing, computing power, computing power network, content delivery network, edge computing*

1. INTRODUCTION

The development of 5G and edge computing has spawned many emerging services. They differ from previous services, as such services have more diverse demands on computing power and networks.

1.1 Low latency requirements

Different services and scenarios have different requirements for the network. Latency is one of the performance indicators and a decisive factor for user experience. Therefore, low latency has become one of the essential requirements of emerging services, which has led to the birth of edge computing.

Typical low-latency scenarios, such as Virtual Reality (VR), Augmented Reality (AR), etc. In such scenarios, user participation is high and the latency requirements mainly come from the smooth interaction between people or between people and devices. In VR and AR scenarios, devices are required to respond to signals given by humans in a timely manner. Therefore, the optimization of human-computer interaction experience places a high requirement on latency of 10 ms [1]. In industrial Internet scenarios, the realization of high-efficiency automated production in factories requires real-time operational control. If some steps in the production do not receive instructions on time, it will affect the product quality and even cause the system to crash. Therefore, industrial Internet also puts forward higher requirements for latency, reaching 1~10ms [1].

The low-latency requirement of the service is usually achieved by moving the computing nodes to the edge. Ubiquitous computing scenarios can also be achieved by selecting a resource node at a reasonable distance from the service. However, when a single node cannot meet the requirements, it is also necessary to use multiple edge nodes for services through collaborative technology.

1.2 High mobility requirements

Similar to the mobile phone service, users are usually moving from the initiation of the service request to the end of the service. Therefore, computing power services need to meet high mobility requirements. Typical scenarios are live broadcast and Internet of vehicles. Live broadcast, for example, anchor X is conducting a travel live broadcast. During a live broadcast, the anchor needs to go to city B from city A, and share the scenery along the way between city A and city B. Live broadcast services need to be able to provide a beauty feature, with real-time interaction with the audience (such as barrage communication, voice and video connections, live streaming commerce, etc.). In the future, VR live broadcasts will require stronger image rendering, greater computing power and lower latency. As the location of the anchor continues to change, using the same computing resources during the change will increase the physical distance between the service and computing, leading to the increase in latency. Therefore, it is possible to use computing power networks to select matching resources for users in real time based on the location and resource conditions, so as to realize the accompanying resources and improve the user experience. Therefore, computing power service location needs to be adjusted according to the user's location in real time. In Internet of vehicles scenarios, mobility is its most prominent feature. During driving, vehicles must perceive and respond to changes in the environment on time in complex traffic environments. It requires low-latency communication between vehicles and vehicles, pedestrians and road facilities. However, with the rapid evolution of vehicle location, a node serving computing power will quickly change from a low-latency node to a high-latency node. To meet its low latency requirements, it is necessary to constantly switch the location of computing service nodes, achieving the millisecond-level latency required for the Internet of vehicles.

High mobility requirements require network- wide resource view and cooperation with artificial intelligence technology to realize intelligent orchestration and scheduling of services, realizing seamless service switching in high mobility scenarios.

1.3 Large computing power requirements

In high-precision scientific research scenarios such as weapons research and development, climate simulation and gene sequencing, their computing tasks are intensive and require extremely high accuracy. It can take years to process if done on a single computer. Taking scientific research as an example, the large computing power scenario has new demands on the network and computing. On the one hand, the bandwidth demand for large-scale scientific data transmission has reached 100 Gbps, and will reach more than terabits per second in the future. In addition, some scientific devices have proposed exascale floating point calculations per second (EFlops). Relying on the original method of building a dedicated computing power pool, the cost is high and the resource utilization efficiency is low.

To meet the large computing power requirements, high-performance computing technology can be used to supply resources for large computing power services by building super-computing centres and cloud computing centres. Besides, idle computing power (cloud computing, edge computing, etc.) around the data source can be scheduled through the network in the future to make up for the computing power gap.

1.4 Tidal requirements

The tidal requirements of services are often overlooked, but it is common. Take the video surveillance of office buildings as an example. During the daytime, there is a large flow of people in the office area, and the data to be processed and analysed will be more complicated; while at night, when there is no one in the building, the video picture is almost static and it does not need to be processed too much. Therefore, in this scenario, the computing demand for the same service fluctuates significantly over time. Thus, the computing power service needs to adjust the provided resources as the services requirements change to maximize resource utilization.

To meet tidal requirements, the computing power service needs to combine artificial intelligence technology to realize the prediction of resource usage through service perception methods such as traffic prediction. Under normal circumstances, the requirements for computing power services for each service are several, not just one. Taking the Internet of vehicles as an example, it has apparent high mobility and low latency characteristics, and the computing demand for complex environments is also high. Therefore, computing resources need to be combined with the network and emerging technologies to realize the multidimensional computing service requirements.

In order to meet the multidimensional computing power service requirements of the business in the scenario of ubiquitous computing power, the computing power network came into being. In September 2021, ITU-T released the first international recommendation for computing power networks, ITU-T Y.2501 "Computing power network —framework and architecture". In ITU-T Y.2501, a computing power network is defined as a new type of network that realizes optimized resource allocation, by distributing computing, storage, network and other resource information of service nodes through a network control plane (such as a centralized controller, distributed routing protocol, etc.). CPN combines network context and user requirements to provide optimal distribution, association, transaction and scheduling of computing, storage and network resources. [2]

This paper will introduce the computing power network technology in detail with regard to ITU-T Y.2501. It is organized as follows. Section 2 specifies the ecosystem of CPN, introducing the participants in a computing power network and how the computing power network provides services to users. Section 3 introduces the functional architecture of a CPN, giving an overall understanding of the computing power network from the perspective of its architecture. Section 4 discusses the key technologies in the computing power network, including the concept and current situation of the key technologies. Section 5 describes the key technologies that are easily confused with the computing power network at an early research stage and expounding the similarities and differences of the computing power with these technologies to help readers better understand the concept and function of the computing power network. Section 6 introduces the progression of standards of computing power networks in ITU in regard to ITU-T Y.2501, which will help to further carry out standardization work. Section 7 describes the key milestones and current application practices of the computing power network, helping readers better understand the development process and current progress of the computing power network. Finally, Section 8 makes a conclusion of this paper, analyses the challenges in computing power networks and its future direction, for the reference of researchers in computing power networks.

2. ECOSYETEM OF CPN

The framework of the computing power network is given in ITU-T Y.2501. The framework specifies the roles in the CPN ecosystem including computing power network consumer, computing power network provider, computing power network transaction platform, computing power network control plane, network operator, etc., as shown in **Fig 1**[2].

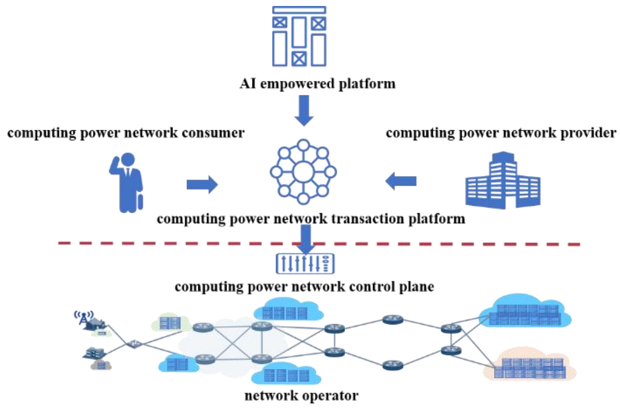

Fig 1 - Framework of computing power network

- Computing power network consumer: Units or individuals who consume computing power resources and network resources. They make various requirements in terms of cost, performance and security according to their services.
- Computing power network provider: This refers to the units or individuals that can provide computing power resources. The computing resource pool can be a small-scale edge computing node, a large or medium-sized cloud computing node, a metropolitan computing node or a super-computing centre, etc. Therefore, the provider can be a telecommunication operator, a large cloud service provider, a small or medium-sized enterprise, a super-computing centre, or even an individual.
- Computing power network transaction platform: A platform that allows computing power network providers and computing power network consumers to make transactions. It can be a public transaction, that is, the computing power network consumers know exactly who provide the computing power resources. It can also be an anonymous transaction, that is, computing power network consumers do not need to know who are the computing power network providers, and the computing power network transaction platform is responsible for transaction reliability and computing security. In addition, on this transaction platform, not only computing resource transactions, but network resource transactions need to be completed at the same time according to location and service requirements.

- Computing power network control plane: This collects computing power information, network information, etc. and sends them to the computing power network transaction platform for computing power network consumers to select appropriate computing power resources and network resources, and then provides computing power network consumers with the optimal computing power resource allocation and network connection solution.
- Network operator: An operator that provides connection services which connects users and computing power resources, and which can provide different levels of connection services according to user requirements.
- AI empowered platform: As an additional module in the computing power network framework, it can provide computing power applications for computing power network consumers as well as AI-based auxiliary operations for computing power network providers.

The different combinations of the roles in the computing power network realize the business model of the computing power network.

The goal of a CPN is to provide integrated computing power services to different types of users. CPN consumers could find proper computing power resources (basic resources) or service resources through a CPN. According to the uses of a CPN, we divide the users of CPN into two categories: CPN consumers and CPN providers. According to the category which the CPN consumers and CPN providers belong to, there are four application models of CPN at present, which we describe in the following subsections [3].

Direct model

The direct model is shown in **Fig 2** and the serial number indicates the workflow. In this model, the CPN consumer is individual users and the CPN provider is basic resource providers. Individual users will use basic resources finally. Individual users find a suitable computing resource node through the CPN, and the CPN builds network connection intelligently between individual users and resource nodes.

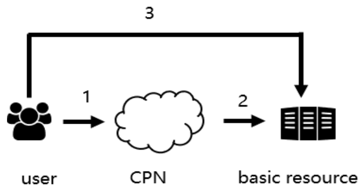

Fig 2- Direct model

User-driven model

The user-driven model is shown in **Fig 3** and the serial number indicates the workflow. In this model, the CPN consumer is individual users and the CPN provider is basic resource providers. A service platform will use the resources finally. Individual users find a suitable computing resource node through the CPN, notify the service platform to deploy the service on that node and then this service will be used by the users.

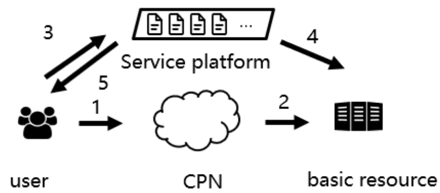

Fig 3- User-driven model

Indirect model

The indirect model is shown in **Fig 4** and the serial number indicates the workflow. In this model, the CPN consumer is service platforms and the CPN provider is basic resource providers. According to its service plans and/or business prediction, the service platform can find a suitable node that can cover a certain area and deploy service capacity by using a CPN. And individual users use the service in the service platform. Since individual users do not use the CPN directly, we call this model indirect.

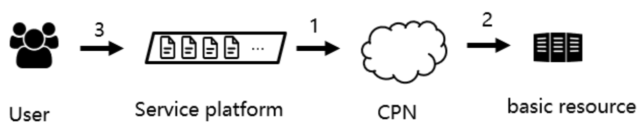

Fig 4 - Indirect model

Service model

The service model is shown in **Fig 5** and the serial number indicates the workflow. In this model, the CPN consumer is individual users and the computing power provider is the service platforms, and an individual user can find services through the CPN. The service platform has already deployed their services in basic resources (service platform can find basic resources through the CPN or other ways). And it announces service information through basic resource nodes in the computing power network. Then, the individual user can find the services through the CPN.

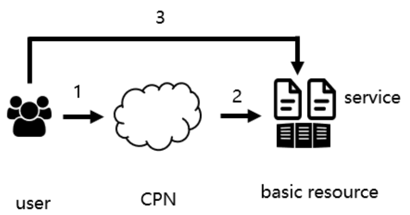

Fig 5 - Service model

3. FUNCTIONAL ARCHITECTURE OF CPN

The functional architecture of a computing power network is depicted in **Fig 6**. It consists of four layers which provide the aforementioned functionalities by interacting with themselves[2].

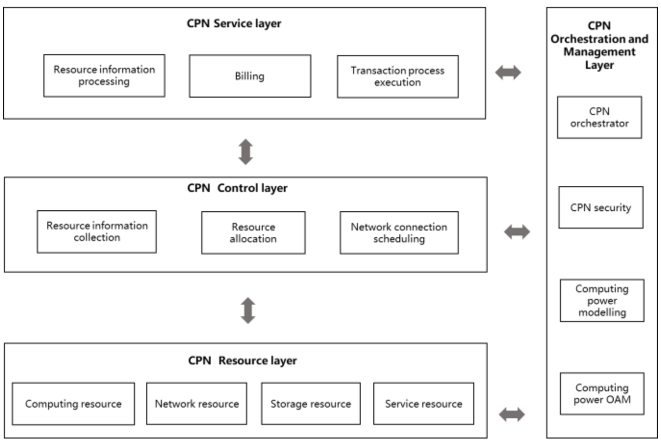

Fig 6 - Computing power network functional architecture

CPN resource layer

The CPN resource layer is where the resources reside provided by computing power network providers and network operators. This includes resources typically used in resource nodes (cloud computing node, edge computing node, etc.) such as computing resources (servers, etc.), network resources (switches, routers, etc.), storage resources (storage devices) and also the deployed services that run on the servers. In this layer, the resources are various, heterogeneous and belong to many providers. Using unified identification can realize the unified authentication and resource scheduling of different vendors and heterogeneous computing power resources.

CPN control layer

The CPN control layer (realized by the CPN control plane) collects the information from the CPN resource layer and sends it to the service layer for further processing. After receiving the processing results from the CPN service layer, the CPN control layer will preoccupy the resources and establish a network connection.

The computing power network control layer has three basic functions: resource information collection function, resource allocation function and network connection scheduling function:

- Resource information collection function: The computing power network control layer collects various resource information, including, but not limited to, computing power resource information, network resource information, storage resource information, algorithm resource information, etc. and generates a resource information table.
- Resource allocation function: According to the CPN consumer's requirement or processing results from the CPN service layer, the computing power network control layer checks the resource information table, then makes a resource allocation strategy and sends it to the CPN providers, such as notifying the computing power network providers when and how many computing resources will be occupied and refreshing their resource information.
- Network connection scheduling function: Network connection requirements are obtained according to the resource allocation strategy, such as among which points the network connection should be established, the bandwidth of each network connection and the quality of service needed to be provided. According to these network connection requirements, the corresponding network resources are scheduled and the network connections are established. (Note: the network connection here indicates not a just traditional communication pipeline, it may also require the deployment of corresponding network elements, such as 5G UPF, vBRAS, vCPE and other access control network elements, according to the network connection requirements.)

CPN service layer

The CPN service layer can realize the functions of the CPN transaction platform mentioned in Section 2. The service layer supports the following functionalities:

- Resource information processing: The CPN service layer obtains various computing power resource information and network resource information from the computing power network control layer. According to user requirements and resource information, it provides optional resources and reasonable prices based on the construction cost, maintenance cost, scarcity and competition relationship of them.
- Billing: This includes two kinds of bills. One is paying bills for computing power consumers according to the occupation statistics of computing power resources and network resources. The other is income bills for computing power providers and network operators according to the supplement of computing power resources and network resources.

Transaction process execution: This executes transactions including analysing user requirements, generating resources view, monitoring the resource usage, etc.

CPN orchestration and management layer

The CPN orchestration and management layer can realize orchestration, security, modelling and OAM function for the CPN.

- The CPN orchestrator is in charge of the orchestration and management of CPN resources and services.
- The CPN security module is responsible for applying security-related controls to mitigate the security threats in CPN environments.
- A computing power modelling module is used for computing power modelling according to various services.
- Computing power OAM realizes the operation, administration and maintenance.

4. KEY TECHNOLOGIES IN CPN

To realize the integrated supply of cloud-network resources, the computing power network needs to gradually solve various technical challenges such as computing power measurement, sensing, routing, transaction, and scheduling. It is worth pointing out that not all these challenges can be solved to provide integrated cloud-network services, but it is a gradual evolution and enhancement process [4]. This section will introduce the key technologies in the computing power network.

4.1 Computing power measurement

Unlike electricity, computing resources cannot be quantified simply in units like "kWh/kWh", especially considering the different types of CPUs, GPUs, FPGAs, ASICs and other chips, it is difficult to make a unified measurement. Therefore, a consensus is needed to quantify heterogeneous computing resources and diverse business needs based on standard specifications, establish a unified description language, and give computing resources a standard unit that can be measured and charged.

At present, industry research institutions, industry alliances and standards organizations have recognized this problem and have launched research work from different perspectives, but no unified conclusion has been formed yet. In the existing application cases, coarse-grained units such as virtual machines and containers are still the main measurement methods.

4.2 Computing power perception

Computing power perception can also be thought of as the perception of all types of resource information, even including the perception of user requirements. However, there are two ways of thinking here. One is for the resource owner to take the initiative to provide resource information and inform the user through the network or centralized systems such as cloud management and resource management; the other is for the network or centralized systems to take the initiative to probe the resource information. Both technical routes are currently in the process of continuous development. For the perception of user demand, the initial stage can use the user intent-driven approach to provide resource demand information actively. And later, with the maturity of artificial intelligence algorithms, traffic prediction models combined with AI deep neural network algorithms can be used to predict resource demand, resource consumption and other aspects to achieve resource pre-allocation, speed up resource deployment and improve overall resource utilization.

4.3 Computing power routing

Computing power routing organically integrates network resource information with computing power resource information. It provides a user-centred view of computing power resources so that users can clearly understand the distribution and offer of various computing power resources to determine the optimal resource combination. It has become the key technical development direction of computing network convergence services.

In line with the traditional network routing method, computing power routing can also adopt distributed, centralized and hybrid schemes. The distributed scheme extends IP protocols to add resource information and service demand information to achieve information flooding and resource routing. The centralized scheme uses centralized control units such as software-defined network (SDN) controllers to collect network information and computing power information and then present them to users in a unified manner. The hybrid solution combines the characteristics of both approaches, using distributed protocols to distribute resource information, and then using centralized control units for unified processing.

4.4 Computing power transaction

With the development of intelligent business, the use of computing power resources has brought about new characteristics of high frequency and a short time. For example, the training part of the artificial intelligence algorithm needs to complete the model training under large data samples in a short time. Still, after the training is completed, the inference part only needs to use a relatively small amount of computing power resources. Therefore, the traditional way of leasing resources by month or day can no longer meet the requirements, and a new resource trading system needs to be designed by combining the characteristics of high frequency and a short time to compress the trading cycle to the hour or even minute level.

In addition, limited by the construction characteristics of edge computing, there will be many supply sides in the edge computing market. How to realize high-frequency transactions with the participation of multiple parties has become a new research point. From the current research, the use of a distributed ledger system based on blockchain technology can meet such requirements, but the specific implementation is subject to further research and demonstration.

4.5 Computing power orchestration

Under the support of high-frequency, multi-party and heterogeneous resource transactions, it is necessary to provide resources quickly according to the transaction content, such as allocating computing power resources and establishing network connections quickly, and being able to release resources quickly and updating resource information after the use is completed.

5. RELATED TECHNOLOGIES

At present, research has been carried out in multiple directions around providing the optimal service for services. Therefore, there are some technical terms that sound similar, such as grid computing, cloud computing, edge computing and content delivery network, etc. The concepts of these technologies and computing power network are often confused by beginners. This section will analyse the similarities and differences between these technologies and computing power network to sort the relationship between these technologies.

5.1 Grid computing

Grid computing studies how to divide a problem that requires a tremendous amount of computing power to solve into many small parts, then distribute these parts to many computers for processing, and finally combine the results of these calculations to get the final result. Its purpose is to use the idle computing power of the CPUs of computers on the Internet to solve large-scale computing problems[5].

The computing power network has similarities with grid computing, but they are essentially different. Firstly, the origin is different. The computing power network focuses more on the commercial value of resources. That is, resource providers can lease the computing resources, storage resources, or other resources to resources' consumers to obtain the corresponding profit. However, grid computing focuses on contributing free resources and volunteering to contribute to some scientific research projects [6].

Secondly, their premise assumptions are different. A computing power network assumes that computing power resources are abundant which can simultaneously serve multiple users and various services. While, the premise assumption of grid computing is that resources are scarce in that era, therefore, computing resources needed to be assembled to serve a single service. Consequently, the computing power network cannot be regarded as a replica of grid computing. Certainly, the idea of sharing computing power resources advocated by grid computing will continue to develop in the computing power network.

5.2 Cloud computing

In order to solve the coordination problem of computing resources and network resources, the cloud computing field has also proposed a variety of solutions[7], such as cloud-network collaboration, that is, it incorporates network resources into the cloud management system. Common practices are: one is, the network opens its capabilities to the cloud's cloud management system, and the cloud management system uniformly schedules computing resources, storage resources, and network resources; the other is, the cloud management system sends network requests to the network control unit, for instance, a network collaborative orchestrator, the network control unit schedules the network according to cloud service requirements. Obviously, the first step is to select the cloud service and then determine the network connection. Therefore, a cloud service provider can connect multiple networks, and even use technologies such as a Soft Defined Wide Area Network (SD-WAN) to achieve cross-domain connections across different network operators.

However, the computing power network solves the problem from another perspective. The computing power pool sends its computing power resource information to the network control plane, and then distributes this information through the network control plane (centralized controller or distributed routing protocols). When users have services' requirements, they can choose the most suitable computing power pool and network path by analysing the network information and computing power information recorded in the routing table. Obviously, the computing power network needs to select the network first, and then select the computing power pool (cloud computing service nodes or edge computing service nodes).

If there is only one optional network service provider and only one cloud service provider / computing power provider, then there is not much difference between cloud-network collaboration and a computing power network. However, in reality, there are many network service providers, and more cloud service providers / computing power providers. In this situation, the difference between cloud-network collaboration and a computing power network is quite large. In the cloud-network collaboration solution, the users first select a cloud service provider, or even a specific cloud resource pool or edge computing node, and then select the network connection product and optimal network path among multiple network service providers. In the computing power network solution, it is necessary to first determine the network service provider and then select the computing node from multiple computing resources according to the requirements of the services on the latency and other indicators, also combining with the network situation.

5.3 Edge computing

Edge computing is a kind of distributed computing in which processing and storage takes place at or near the edge [8].

The generation of computing power networks is closely related to edge computing. On the one hand, the birth of edge computing satisfies the low-latency requirements of new services. It solves the congestion caused by a large amount of data in the backbone network. However, on the other hand, it also changes the traffic model of the network, resulting in multilevel computing nodes all over the network [9]. How to realize the coordinated scheduling of multilevel resource nodes and the flexible deployment of applications, so as to provide users with a consistent service experience, has also become critical. Edge computing reduces the latency by shortening the physical distance, while the computing power network reduces the latency by selecting the nearest computing node for the services. The computing power network focuses on the optimal selection, allocation and scheduling of computing power resources, and can support edge computing to better serve low-latency services.

5.4 Content delivery network

The idea of a Content Delivery Network (CDN) is to avoid bottlenecks and links on the Internet that may affect the speed and stability of data transmission to make content transmission faster and more stable. Through a unified CDN scheduling server, the CDN finds a server node with a closer distance for the user according to the location applied by the user and shortens the response time between the user and the server by shortening physical distance and improving user experience. The computing power network has the same working method as a CDN.

However, different from CDN, the CDN is mainly oriented to the content storage service and realizes the arrangement and management of the storage resources in the network. In contrast, the computing power network is oriented to computing, so its implementation is more complicated than for a CDN, and the scheduling is also more refined.

6. STANDARDIZATION PROGRESS IN ITU

At present, ITU has carried out a series of standardization work around the overall architecture of Y.2501.

In March 2020, China Unicom and China Telecom started a new work item, Q.CPN “Computing power network signaling architecture”. This Recommendation provides the overview for requirements of the CPN, interface CPN reference model, signalling requirements for interfaces CPN. The draft Recommendation has been updated in July 2021 and has contained the architecture and functional entities, signalling requirement, signalling procedure of managing resources, and the information flow of deploying service. [10]

In March 2020, China Unicom, China Telecom and China Academy of Information and Communications Technology （CAICT）started a new work item Q.BNG-INC “Requirements and signalling of intelligence control for the border network gateway in computing power network”. The draft Recommendation was updated in December 2021. This Recommendation is focused on the general requirements, service requirements, network intelligence control requirements, along with the architecture and procedure of intelligence control for the BNG devices in computing power networks. [11] This Recommendation gives the architectures of intelligence control for the BNG devices in a CPN from two scenarios. One scenario is edge computing interconnect. In this scenario, the servers in the edge cloud are connected to the BNG. The edge clouds are interconnected by the distributed routing protocol running on the BNG through the underlay network, which decides that the inter-domain resource information collection and computing power scheduling should be conducted in a distributed way. The other scenario is data centre interconnect. In this scenario the BNG device could be a WAN gateway, which is able to manage the intra-domain computing power scheduling in a centralized way. Different from the scenario of edge computing interconnect, the BNG devices for the data centre interconnect are all connected to an inter-domain WAN controller.

In March 2021, China Telecom, China Unicom and CNIC CAS started a new work item Q.CPN-TP-SA “Signalling architecture of transaction platform in CPN”. [12]This draft Recommendation provides the signalling architecture of transaction platforms in a CPN. This Recommendation provides a signalling architecture of a CPN transaction platform. It also introduces the signalling procedure, signalling requirements of a CPN transaction platform and defines the protocols used for reference points.

In July 2021, China Telecom started a new work item Y.NGNe-O-CPN-reqts "Requirements and framework of NGNe orchestration enhancements for supporting computing power network". This draft Recommendation provides an overview of NGNe orchestration enhancements for supporting computing power networks and defines the requirements and framework of NGNe orchestration enhancements for the support of computing power network in an NGNe environment. The draft Recommendation was updated in December 2021. This Recommendation believes that to fulfil the objective of a computing power network, it requires cooperation from network and cloud service providers which might be the most difficult part for implementation of this idea. Therefore, the involvement of NGNe orchestration would be appropriate which has the capabilities to coordinate the network and cloud resources as a whole. [13]

In July 2021, China Unicom, China Telecom, CAICT and Inspur started a new work item Y.ASA-CPN:” Computing power Network authentication scheduling architecture”. This draft Recommendation provides requirements for computing power network scheduling technology architecture and a corresponding trusted authentication mechanism, computing power resource management, unified scheduling, and a computing power identification system. [14]

In December 2021, China Unicom, China Telecom, Beijing University of Posts and Telecommunications and CAICT started a new work item-Y.SFO “Requirements and framework of Service Function Orchestration based on Service Function Chain (SFC)”. This draft Recommendation provides the scenarios of service function orchestration based on the SFC, such as based on user requirement, to realize function services deployed and dynamically adjusted on demand and based on the resource situation, to realize the resource optimization and load balance of service functions, specifies its general requirements and provide a basic framework. This Recommendation provides two scenarios, one is an SFC solution based on function orchestration, and the other is an SFC solution based on performance orchestration. [15]

7. MILESTONE PROGRESS

Research on the computing power network has only been going and it has caused waves in the industry due to its huge potential demand. At present, China's major operators, manufacturers and academic institutions have joined the research on computing power networks. The computing power network has made a number of significant advances.

7.1 Milestones

In September 2019, the first paper on computing power networks was published, explaining the concept of a computing power network. [16]

In October 2019, the first ITU-T draft Recommendation of computing power networks, ITU-T Y.CPN-arch was initialized. [17]

In November 2019, the first white paper of computing power networks, "China Union Computing Network White Paper" was released. [18]

In January 2020, a computing power network prototype was demonstrated at ITU-T for the first time.

In November 2020, the first computing power network book "Edge Computing and Computing Power Network" was released. [19]

In December 2020, the prototype of a computing power network transaction platform was displayed.

In September 2021, ITU-T released the first international standard for computing power networks, ITU-T Y.2501, gave the series number Y.2500-Y.2599 for computing power networks, starting a new field of international standards. [2]

In September 2021, BBF completed its first computing power network standard. [20]

7.2 Recent implementation

Plenty of study work has been carried out on the computing power network technologies in the early stage, and it is currently developing from theoretical research to application. At present, the latest application practice of computing power networks is under preparation, and this section will introduce a typical case.

In this case, there are the following key components: CPN transaction platform, CPN controller and computing power gateways, as shown in **Fig 7** .

The CPN transaction platform is related to the CPN service layer in the computing power network architecture and realizes the process of computing power transaction.

The CPN controller realizes the function of the CPN control layer, and is responsible for collecting, analysing and processing computing power information and network information. In this solution, the computing power network controller consists of the current cloud management platform and network management platform (SDN controller) and some decision-making functions.

A computing power gateway is placed at the exit of the resource pool to realize the exchange and distribution of computing power resource information, report and cooperate with the CPN transaction platform and CPN controller to complete the entire computing power network process. The computing power gateway can offload the functions of the cloud management platform and network management platform in varying degrees depending on the solution.

The workflow of the solution is:

Step 1: The computing resource pools send their information to the computing power network gateways. The CPN controller collects global computing power and network information from the computing power gateway and other network nodes through the network management platform (SDN controller) and cloud management platform, then sends it to the CPN transaction platform after processing.

Step2: The users send resource requirements to the CPN transaction platform, and the user selects or the transaction platform recommends suitable resources for them according to their requirements.

Step 3: According to the transaction status between the user and the resource providers on the CPN transaction platform, the CPN controller sends the corresponding resource allocation strategy to each resource pool, such as notifying the computing power resource providers, how many computing resources will be occupied during the time period and the resource information data recorded by the platform will be refreshed at the same time.

Step 4: According to the allocation of network resources, the network management platform obtains the network connection requirements, such as what kind of connections needs to be established and what kind of service quality assurance should be provided. Based on these service requirements, the corresponding network connections are established.

Step 5: Business release; users could use these computing resources and network resources.

This solution is relatively easy to be implemented at the current stage. Since the functions of the cloud management platform and network management platform are mature in the existing network, this solution only needs to develop some functions for unifying these two platforms to realize the basic principle of computing power network. And in this solution, the computing power gateway only has the ability to collect computing power information. In fact, it only has a simple processing and forwarding ability, but in other solutions, it could do more work to replace the cloud management platform and network management platform.

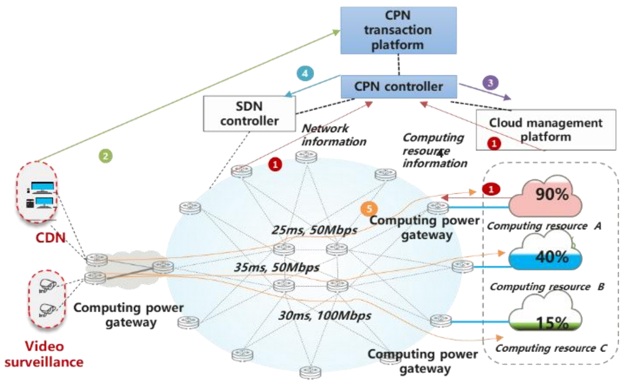

Fig 7 - Computing power network solution

8. CONCLUSION

This paper introduces a new technical field, computing power networks. Computing power network technology improves the current end-to-end model of computing at the edge of the network and supports a distributed resource mode. It can change the service mode that the network is only used as a transmission pipeline, and massive services are reused, which improves the utilization efficiency of computing resources; it can, especially, greatly improve the utilization efficiency of edge computing and solve the scalability problem of resource-constrained nodes such as edge computing. From CPN providers, the utilization of resources can be improved, costs can be reduced and resource benefits can be maximized through the reasonable arrangement of resources. Computing power networks can also give full play to the role of the network, reducing the resource occupancy rate of the management and control system, and reducing the construction cost. It can also promote the vigorous development of network-related technical fields in the future, which is conducive to breaking through the closure of the traditional computing industry and building a more open and win-win industrial ecological environment for CPN consumers and CPN providers. The computing power network technology has gradually evolved from technical research to practical implementation. However, there are still many challenges in its development. For example, the key technical standards have not yet been unified. Due to the heterogeneous nature of computing power, the promotion of relevant standards requires the participation of more partners; the functions of prototype products are not yet perfect, and according to the computing power network concepts and different implementation methods, the current computing power network prototype products are still relatively simple in function, and most of them are used in centralized computing power network solutions. The development of devices supporting a distributed computing power network solution is difficult. The existing devices can only support laboratory verification. In actual use, it will face problems such as a high refresh rate of computing power information which is hard to support for network equipment. Therefore, in the following work， it is necessary to break through the key technical difficulties of computing power measurement and computing power modelling in the future and focus on the standardization of key technologies. In terms of equipment, it is also necessary to design a more reasonable solution considering the capabilities of the current network equipment and further develop computing power gateways and computing power network controllers to meet higher business requirements. In addition, it is also necessary to start the research on the combination of computing power networks and 6G. On the one hand, the powerful access capability of 6G can be used to enable the computing power network, and on the other hand, the computing power network can be used to provide strong computing power support for 6G networks and 6G services.

ACKNOWLEDGEMENT

This work is supported by the National Key R&D Program of China (No. 2020YFB1806700).

REFERENCES

[1] Xutong Zuo, et al, “Low-latency networking:architecture, key scenarios and research prospect” Journal on Communications, vol. 40 No.8, 2019.

[2] ITU, Recommendation ITU-T Y.2501, “Computing Power Network- framework and architecture”, 2021.

[3] Bo Lei, et al, “Computing Power Network: An Interworking Architecture of Computing and Network Based on IP Extension”, 2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR), Paris, 2021.

[4] Maozheng Li, et al, “cloud and network convergence: Digital Information Infrastructure in the Computing Power Era”, Beijing, 2022, pp.135-137.

[5] Lan Foster, et al, “The grid: blueprint for a new computing infrastructure”, San Fransisco: Morgan Kaufmann publisher, 1999.

[6] Zhengyu Dong, “Research on task scheduling algorithm in Grid Computing”, ChongQing University, 2009.

[7] S. Azodolmolky, P. Wieder and R. Yahyapour, "Cloud Computing Networking: Challenges and Opportunities for Innovations", IEEE Communication Magazine, vol. 51, no. 7, pp. 54-62, July 2013.

[8] ISO. ISO/IEC TR 23188:2020(en) : Information technology — Cloud computing — Edge computing landscape . (2020-02) .https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:23188:ed-1:v1:en.

[9] Bo Lei, et al, “Overview of edge computing and computing power network” ZTE Communications, vol. 27, 2021, pp.:3-6.

[10] ITU, Draft Recommendation ITU-T Q.CPN “Signalling requirements for Computing power Network”, 2021.

[11] ITU, Draft Recommendation ITU-T Q.BNG-INC “Requirements and signalling of intelligence control for the border network gateway in computing power network”, 2021.

[12] ITU, Draft Recommendation ITU-T Q.CPN-TP-SA “Signalling architecture of transaction platform in CPN”, 2021.

[13] ITU, Draft Recommendation ITU-T Y.NGNe-O-CPN-reqts: "Requirements and framework of NGNe orchestration enhancements for supporting computing power network", 2021.

[14] ITU, Draft Recommendation ITU-T Y.ASA-CPN: “Computing power Network authentication scheduling architecture”, 2021.

[15] ITU, Draft Recommendation ITU-T Y.SFO “Requirements and framework of Service Function Orchestration based on Service Function Chain”, 2021.

[16] Bo Lei, et al, “A new edge computing solution based on the interworking among cloud, network and edge: Computing power network,” Telecommunication Science, vol. 9, 2019, pp.:44-51.

[17] ITU, Draft Recommendation ITU-T Y. CPN-arch: “framework and architecture of computing power network,”.2019.

[18] China Unicom, “China Unicom computing power network white paper,” Beijing: China Unicom, 2019.

[19] Bo Lei, et al, “Edge computing and Computing power network,” Beijing, 2020.

[20] BBF,” NPIF - SDN and NFV - Metro Computing Network (MCN)”, 2021.

AUTHORS

Qianying Zhao is now an engineer of the future network research center of China telecom research institute. Qianying Zhao received her Master's degree in Sant'Anna School of Advanced Studies, Pisa, Italy, in 2018. Her current research interests include computing power networks and edge computing. She is the author of two technical books and has published seven papers in journals.

Bo Lei is now the deputy director of the Network Technology Research Institute of the China telecom research institute. Bo Lei received his Master's degree in telecommunication engineering from Beijing University of Posts and Telecommunications, Beijing, P. R. China, in 2006. His current research interests include future network architecture, new network technology, computing power networks and 5G application verification. Bo Lei now leads the future network research center focusing on future networks. He is the first author of two technical books and has published more than 30 papers in top journals and international conferences and filed more than 30 patents.

Min Wei is now an engineer of future network research center of China telecom research institute. She received her Master's degree from Beijing University of Posts and Telecommunications, Beijing, P. R. China, in 2015. Her current research interest is computing power networks.