International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.2501

(09/2021)

SERIES Y: GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

Next Generation Networks – Computing power networks

Computing power network – Framework and architecture

Recommendation ITU-T Y.2501

ITU-T Y-SERIES RECOMMENDATIONS

GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

GLOBAL INFORMATION INFRASTRUCTURE	
General	Y.100–Y.199
Services, applications and middleware	Y.200–Y.299
Network aspects	Y.300–Y.399
Interfaces and protocols	Y.400–Y.499
Numbering, addressing and naming	Y.500–Y.599
Operation, administration and maintenance	Y.600–Y.699
Security	Y.700–Y.799
Performances	Y.800–Y.899
INTERNET PROTOCOL ASPECTS	
General	Y.1000–Y.1099
Services and applications	Y.1100–Y.1199
Architecture, access, network capabilities and resource management	Y.1200–Y.1299
Transport	Y.1300–Y.1399
Interworking	Y.1400–Y.1499
Quality of service and network performance	Y.1500–Y.1599
Signalling	Y.1600–Y.1699
Operation, administration and maintenance	Y.1700–Y.1799
Charging	Y.1800–Y.1899
IPTV over NGN	Y.1900–Y.1999
NEXT GENERATION NETWORKS	
Frameworks and functional architecture models	Y.2000–Y.2099
Quality of Service and performance	Y.2100–Y.2199
Service aspects: Service capabilities and service architecture	Y.2200–Y.2249
Service aspects: Interoperability of services and networks in NGN	Y.2250–Y.2299
Enhancements to NGN	Y.2300–Y.2399
Network management	Y.2400–Y.2499
Computing power networks	**Y.2500–Y.2599**
Packet-based Networks	Y.2600–Y.2699
Security	Y.2700–Y.2799
Generalized mobility	Y.2800–Y.2899
Carrier grade open environment	Y.2900–Y.2999
FUTURE NETWORKS	Y.3000–Y.3499
CLOUD COMPUTING	Y.3500–Y.3599
BIG DATA	Y.3600–Y.3799
QUANTUM KEY DISTRIBUTION NETWORKS	Y.3800–Y.3999
INTERNET OF THINGS AND SMART CITIES AND COMMUNITIES	
General	Y.4000–Y.4049
Definitions and terminologies	Y.4050–Y.4099
Requirements and use cases	Y.4100–Y.4249
Infrastructure, connectivity and networks	Y.4250–Y.4399
Frameworks, architectures and protocols	Y.4400–Y.4549
Services, applications, computation and data processing	Y.4550–Y.4699
Management, control and performance	Y.4700–Y.4799
Identification and security	Y.4800–Y.4899
Evaluation and assessment	Y.4900–Y.4999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T Y.2501

Computing power network – Framework and architecture

Summary

Recommendation ITU-T Y.2501 describes the framework and architecture of the computing power network (CPN). This is 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. It combines network context and user requirements to provide the optimal distribution, association, transaction and scheduling of computing, storage and network resources.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.2501	2021-09-13	13	11.1002/1000/14768

Keywords

Architecture, computing power network, framework.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T Y.2501

Computing power network – Framework and architecture

1 Scope

This Recommendation provides the framework and architecture of the computing power network, specifies its functional entities and defines the functionalities of these functional entities, in addition this Recommendation also provides general scenarios, requirements and security considerations of the computing power network (CPN).

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.2011] Recommendation ITU-T Y.2011 (2004), *General principles and general reference model for Next Generation Networks.*

[ITU-T Y.2012] Recommendation ITU-T Y.2012 (2010), *Functional requirements and architecture of next generation networks.*

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

5G	5th-Generation
AI	Artificial Intelligence
CPN	Computing Power Network
OAM	Operation Administration and Maintenance
SDN	Software Defined Network
UPF	User Plane Function
V2X	Vehicle to X
vBRAS	virtual Broadband Remote Access Server

vCPE virtual Customer Premise Equipment

VR Virtual Reality

5 Conventions

In this Recommendation:

The keywords "is required to" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this Recommendation is to be claimed.

The keywords "is recommended to" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

6 Background and motivations

The computing power network (CPN) is 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.

CPN provides technical opportunities within a certain business context.

6.1 From the perspective of computing power resource consumers

Demands for computing power are increasing day by day and ubiquitous computing power resources are needed.

The market for computing-intensive services, such as the AI-based services is growing. The large-scale application of artificial intelligence (AI) algorithms across various industries requires a great amount of computing power resources. In this the age of cloud computing, the tendency is to build centralized computing power resources pools to solve this problem. However, in some new scenarios, these centralized pools do not satisfy requirements as appropriate computing power nodes need to be selected according to the service characteristics, price, and network conditions. An access control system based on face recognition, shown in Figure 6-1, serves to explain why ubiquitous computing power resources are needed.

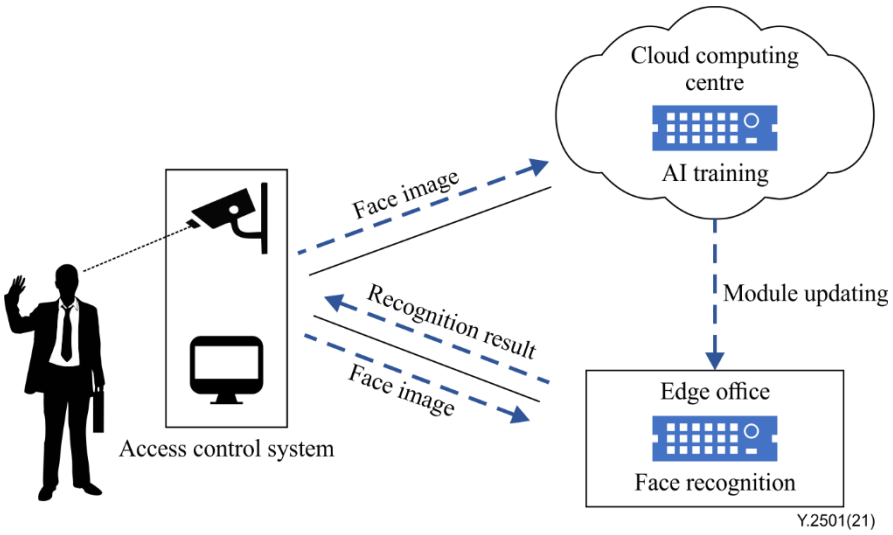

Figure 6-1 – Access control system based on face recognition

In this example, an AI training part can be deployed in centralized computing power pools, such as a cloud computing centre, to perform complex computing processes offline. However, in the reasoning phase, such as for instance, when recognizing a face to open a door, the latency factor should be considered. If the image or video information is sent to a cloud computing centre for processing, the waiting time may be too long to be practical. Consequently, services need to consider not only computing power resources but also the time for processing, the time for transmitting and the cost. For fully-centralized computing power resource pools, they cannot satisfy all requirements, while distributed computing power resource pools, such as edge computing nodes, have obvious advantages in bandwidth and latency, thus meeting the development requirements of future services.

6.2 From the perspective of computing power resource providers

The construction costs of computing power resources per unit are decreasing.

In accordance with Moore's Law, the performance of chips is improving, while the costs of computing power resources per unit are decreasing. Compared with the provision of electricity, the provision of computing power is relatively easy. In theory, as long as there is space and electricity and due compliance with fire safety regulations, all units or even individuals can build a computing power resource pool to provide computing resources for various services. However, the existing computing resources pools, self-built by enterprises, are geographically scattered, and their utilization is quite low. If all these resources could be used, it would generate more value and create more wealth for users and society. It is also noted that super-computing centres own a large amount of computing power resources, while users do not have a way to apply for these resources. Consequently, there is a need to find a new solution and technology system to realize computing power resources sharing.

6.3 From the perspective of network technology

Network development has laid the foundation for the flexible provision of multi-party computing resources.

For sharing of computing power resources, the accessibility of computing power resources needs to be considered first. With the development of network technologies such as 5G and software defined networks (SDNs) the network is now no longer a cause of bottlenecks. Users and computing power resources can be efficiently connected through networks.

The isolated computing power pool is difficult to transform into a measurable and saleable commodity that can promote a sustainable development of the industry chain. With the development of new network technologies such as 5G, computing power resources can be connected in a simple, efficient, and low-cost way, and can be sold and used through a new business model.

6.4 From the perspective of a business model

A new business model is necessary.

How to integrate the resources of all parties, so that users can flexibly choose appropriate computing resources according to service requirements is problematic. Cloud computing providers need to employ a lot of people to negotiate with owners of computing power resources, sign and then execute service contracts in order to integrate the discrete computing power resources. In addition, operation and maintenance operations also require a great number of staff. This goes against their principles of minimalist operation and maintenance. Consequently, cloud service providers prefer to cooperate with enterprises with large amounts of resources rather than small companies or even individuals.

It is hard for the owners of small and mid-sized computing resources pools to compete with large-scale cloud computing service providers. First of all, due to the limitations of their technical capabilities, the products and services that can be provided are quite limited. Secondly, limited by the size of their enterprises, their marketing efforts are limited, and it is therefore difficult to inform potential users of their resource information.

With the support of new technologies, computing power resources information can be distributed through networks. A computing power transaction platform for resource providers and resource consumers could be provided. Through this computing power transaction platform, computing power providers and consumers do not need to be concerned about the transaction partners. The transaction platform shields the difference between the providers and the consumers. Providers and consumers do not perceive each other, and they can pay according to unified rules, forming a new resource transaction model.

7 Scenarios

Computing power networks mainly work in two types of scenarios: the first type are low-latency scenarios represented by edge computing. In this kind of scenario, the computing power network needs to provide coordination between computing resources; the second type are high-mobility scenarios represented by live broadcast and vehicle to X (V2X). In this kind of scenario, the computing power network needs to provide flexible scheduling for services.

7.1 Scenarios of low latency

The low latency requirement of edge computing needs the interworking of computing and network capabilities. Low latency is one of the most critical characteristics of edge computing, and it is also the key factor that distinguishes edge computing from cloud computing. In the edge computing scenario, clients have an urgent need for low latency, high computing capabilities and networking routing capabilities, which is already beyond the capabilities of traditional cloud service providers, and network capabilities have become a part of the cloud computing capabilities. Therefore, in addition to deploying computing nodes at the edge of the network, it also needs a 'computing-network interworking' reconstruction of the substrate network architecture. This reconstruction evolves the traditional network architecture mainly carrying north-south traffic to the new network architecture which can schedule flexibly taking into account latency and carry east-west traffic as well.

The new application of edge computing needs 'cloud-network interworking' network architecture. For a typical AI application for example, AI reasoning needs low latency, therefore it should be deployed in edge computing nodes. While AI training can be deployed in a centralized cloud computing platform at low cost as it needs high computing capabilities but has no requirements on latency. However, the current edge computing solution generally follows the network construction scheme of traditional data centres. This network construction solution consists of two layers of switches (aggregation + access or Spine-Leaf architecture) and one layer of egress routers [ITU-T Y.2011] and [ITU-T Y.2012]. This kind of networking architecture increases the number of routes and devices of the edge computing nodes. It also increases the latency, and the latency is uncontrollable which is quite different from the expected indicators.

Therefore, for AI applications, computing power network technology is needed to collect the resources and network information about each resource node, then, according to the service requirement, the AI training task with the heaviest computing needs can be deployed on the central cloud. The computing tasks such as feature extraction, template matching, and object recognition are dynamically offloaded to appropriate edge nodes. The terminal only needs to be responsible for target tracking and screen rendering and display to reduce the end-to-end latency.

7.2 Scenarios of high mobility

Services such as live broadcast for travel and V2X often have high mobility. In an example scenario of live broadcast, anchor X is conducting a travel live broadcast. During the 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 scenery features, real-time interaction with the audience (such as barrage communication, voice and video connections, live streaming commerce, etc.). In the future, VR live broadcast 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 the change will increase as will the physical distance between the service and the computing resources, leading to increased latency. Therefore, the computing power network may select matching resources for users in real-time based on the location and resource conditions, so as to provide the accompanying resources and improve the user experience. As shown in Figure 7-1, when anchor X is in city A, resource pool N1 provides services, when anchor Xi is in city B, resource pool N2 provides services.

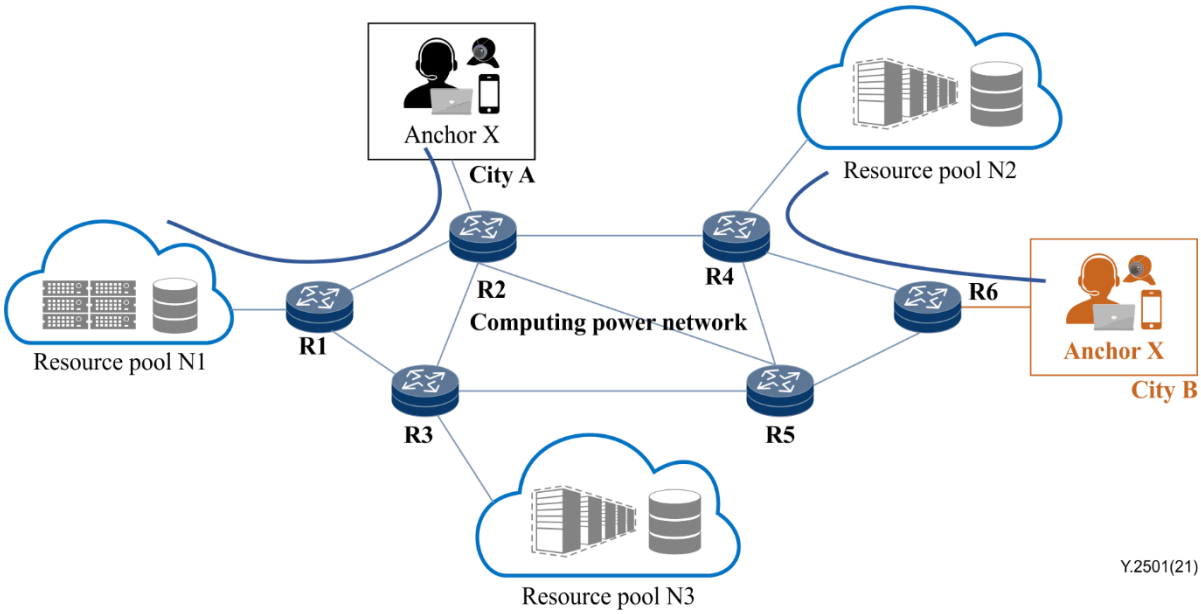

Figure 7-1 – Live broadcast scenario

8 Requirements of computing power network

Computing power networks are required to have the ability to allocate the computing power resources according to user requirements and adjust the allocation strategy with the change of requirements in a timely manner. Thus, the following requirements must be satisfied:

– CPN is required to support computing power measurement for the quantification of computing power resources. In other words, CPN is required to provide unified description for computing power resources.

– CPN is required to have the capability to collect resources information.

– CPN is required to make resource allocation strategies for users according to user requirements.

– CPN is required to establish network connection between resource users and resource providers according to the resource allocation strategies.

– CPN is required to support resource transaction between resource users and resource providers.

– CPN is required to support resource monitoring to update resource information in a timely manner.

NOTE – The unified description refers to a set of measurement rules, according to which we can compare computing resources under certain conditions. It does not mean we need to give only one unit for all computing power resources.

9 General framework of the computing power network

A computing power network framework includes a computing power network consumer, a computing power network provider, a computing power network transaction platform, a computing power network control plane, a network operator, etc. An AI empowered platform can be linked according to services development requirements. The computing power network framework is shown in Figure 9-1.

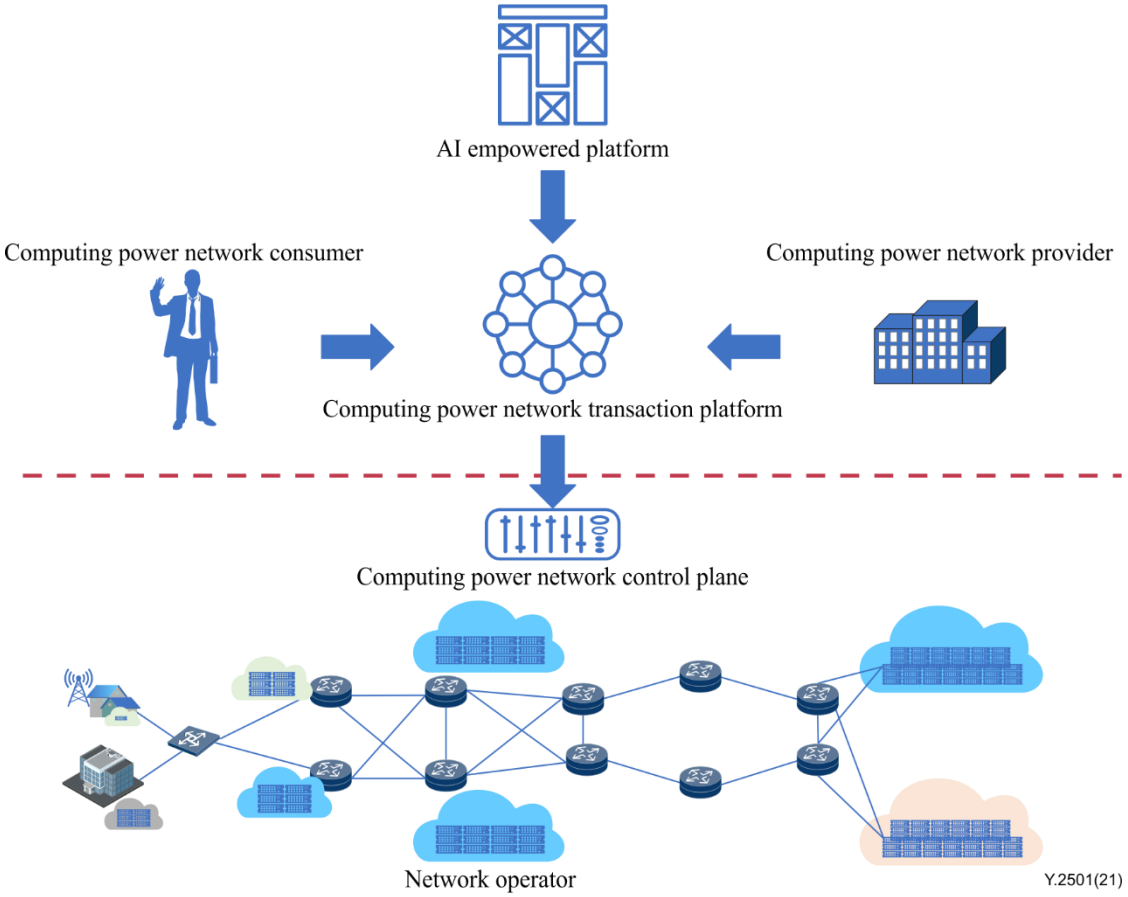

Figure 9-1 – Computing power network framework

– 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: 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 supercomputing 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 provides the computing power resources. It can also be an anonymous transaction meaning that the computing power network consumers do not need to know who the computing power network providers are, 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: It collects computing power information, network information, etc., and sends this information to the computing power network transaction platform for computing power network consumers to select appropriate computing power resources and network resources, and then provides the computing power network consumers with the optimal computing power resource allocation and network connection solution.

– Network operator: An operator that provides connection services which connect users and computing power resources, and 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.

10 Functional architecture of computing power network

The functional architecture of the computing power network is depicted in Figure 10-1. It consists of four layers which provide the aforementioned functionalities by interacting with each other.

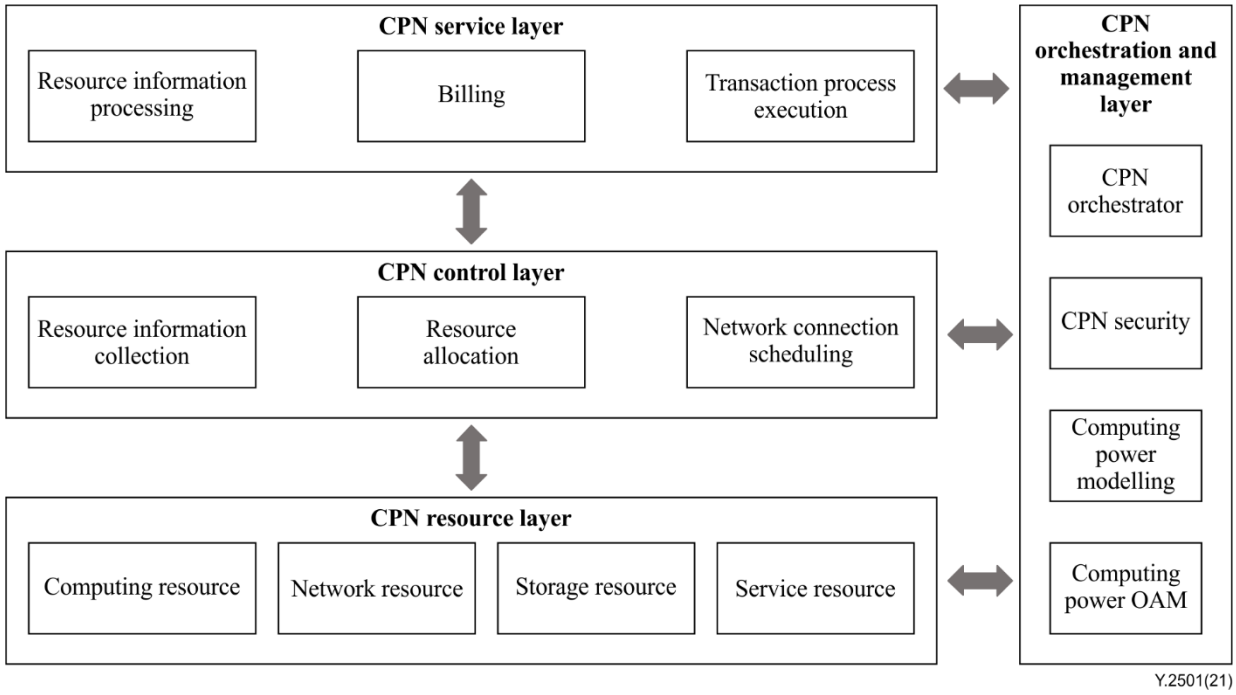

Figure 10-1 – Computing power network functional architecture

10.1 CPN resource layer

CPN resource layer is where the resources reside that are 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.

10.2 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 a service layer for further processing. After receiving the processing results from the CPN service layer, the CPN control layer will pre-occupy 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 a variety of 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 consumers 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 A resource allocation strategy could be 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. Network connection requirements could include 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 just traditional communication pipeline, it may also require the deployment of corresponding network elements, such as 5G user plane function (UPF), virtual broadband remote access server (vBRAS), virtual customer premise equipment (vCPE) and other access control network elements, according to network connection requirements.

10.3 CPN service layer

The CPN service layer can realize the functions of the CPN transaction platform mentioned in clause 9. The service layer supports the following functionalities:

– **Resource information processing**: 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 pricing based on the construction cost, maintenance cost, scarcity, and their competition relationships.

– **Billing**: 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**: The transaction process is as follows:

  • First, computing power network consumers input resource requirements or service requirements, such as resource location, latency, bandwidth, resource quantity.

- Second, the computing power network transaction platform generates a resource view according to the user requirements and resource information received from the CPN control layer. In the resource view, it contains the optional resources as well as their price.
- Third, users choose the most suitable resources according to the resource view, and then make a transaction contract with providers.
- Then, the computing power network transaction platform sends the transaction information to the computing power network control layer, and updates the corresponding resource information.
- Finally, the computing power network transaction platform will monitor the resource usage until the end of the transaction contact. The computing power network transaction platform will then terminate the service and release the associated resources.

Emerging technologies such as blockchain are recommended to be used in the service layer to realize new functions such as distributed ledgers and anonymous transactions.

10.4 CPN orchestration and management layer

The CPN orchestration and management layer can realize orchestration, security, modelling, and operation administration and maintenance (OAM) functions for CPN:

– The CPN orchestrator is in charge of the orchestration and management of CPN resources and services.

– CPN security module is responsible for applying security related controls to mitigate the security threats in CPN environments.

– Computing power modelling module is used to computing power modelling according to various services.

– Computing power OAM realizes the operation, administration and maintenance.

11 Security considerations

The following are the security considerations for CPN:

1) CPN is required to provide mechanisms to support trusted transactions.

2) CPN is required to provide mechanisms to ensure the security of applications published by third-parties.

3) CPN is required to provide security mechanisms to authorize and authenticate computing power providers.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	**Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities**
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2021