CHINA COMMUNICATIONS

Deterministic Computing Power Networking: Architecture, Technologies and Prospects

Qingmin Jia[1], Yujiao Hu[2], Xiaomao Zhou[1*], Qianpiao Ma[3], Kai Guo[1], Huayu Zhang[1], Renchao Xie[1,4], Tao Huang[1,4], and Yunjie Liu[1,4]

[1] Future Network Research Center, Purple Mountain Laboratories, Nanjing, 211111, China
[2] School of Data Science and Artificial Intelligence, Chang'an University, Xi'an, 710064, China
[3] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China
[4] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[*] The corresponding author, email: zhouxiaomao@pmlabs.com.cn

Abstract: With the development of new Internet services with computation-intensive and delay-sensitive tasks, the traditional "Best Effort" network communication mode has been greatly challenged. The network system is urgently required to provide end-to-end communication determinacy and computing determinacy for new applications to ensure the safe and efficient operation of services. Based on the research of the convergence of computing and networking, a new network paradigm named deterministic computing power networking (Det-CPN) is proposed. In this article, we firstly introduce the research advance of computing power networking. And then the motivations and scenarios of Det-CPN are analyzed. Following that, we present the system architecture, technological capabilities, workflow as well as key technologies for Det-CPN. Moreover, performance evaluation and simulation results are presented to illustrate the performance of the proposed scheme. Finally, the challenges and future trends of Det-CPN are analyzed and discussed.

Keywords: Computing and Network Convergence, Computing Power Networking, Deterministic Networking, Deterministic Computing Power Networking, Det-CPN

Received: Nov. 19, 2024
Revised: Jan. 22, 2025
Editor:

I. INTRODUCTION

With the development of emerging network applications such as artificial intelligence (AI), autonomous driving, cloud virtual reality (VR) and intelligent manufacturing, these new applications have put forward higher requirements for network communication latency and computing power [1]. For example, the GPT-3 model has 175 billion parameters, and training the GPT-3 model requires 355 GPU years (a GPU V100 runs for 355 years) [2]. According the research report [3], it is estimated that by 2030, total general computing power will see a tenfold increase and reach 3.3 ZFLOPS, and AI computing power will increase by a factor of 500, to more than 100 ZFLOPS. In addition, in the filed of industrial automation, the communication between Programmable Logic Controllers (PLCs) usually has requirements regarding the upper bound of latency, and the underlying networking infrastructure must ensure a maximum end-to-end message delivery time in the range of 100us to 50ms [4]. Hence, it is necessary and significant to design a new network architecture with ultra-low latency, ultra-high bandwidth, and ultra-strong computing power.

To cope with the challenges brought by new application development, the academia and industry have been actively exploring. For example, in order to meet the challenges of new business demands for computing power, computing power networking (CPN) has been proposed, aiming at connecting distributed computing nodes, achieving rapid access to computing resources and efficient distribution of computing tasks[5][6][7][8]. Meanwhile, in order to address the challenge of latency and jitter requirements for new services, deterministic networking has also been proposed, aiming at ensuring the quality and reliability of data transmission[9][10].

However, some emerging network applications (e.g., Cloud VR, autonomous driving) have both latency-sensitive and computation-intensive characteristics, which place high demands on both latency and computing power [11] [12]. And the current research on CPN mainly focuses on how to schedule computing tasks to matching computing nodes, while neglecting the communication determinacy and computation determinacy. Therefore, the current CPN cannot solve the determinacy problems of communication and computation. How to design a new network architecture that meets latency and computing power requirements has become a major challenge at present.

Fortunately, many researchers have begun to pay attention to this issue. In [13], the authors propose a task deterministic network architecture that provides communication with bounded low latency and zero jitter for critical tasks among edge computing systems. In [14], the authors proposed a deep reinforcement learning based deterministic scheduling architecture for computing and networking convergence, achieving deterministic end-to-end transmission with bounded latency. In [15], the authors proposed a lightweight real-time scheduler named CoRaiS for multiedge cooperative computing, so as to improve decision-making efficiency and enhance transmission and processing performance. However, these works usually only considered transmission determinacy, and did not consider computing determinacy. As a result, it is still difficult to meet the demands of time-sensitive and computation-intensive computing tasks.

In this paper, we propose a new network paradigm called deterministic computing power networking (Det-CPN), which is based on computing power networking and deterministic network technology, and integrates deeply network and computing resources. The Det-CPN can provide end-to-end deterministic computing-network service capabilities for time-sensitive and computation-intensive applications, achieving the determinacy for latency, jitter, path, and computing. Thus, the Det-CPN can effectively address the challenges that traditional "Best Effort" network transmission mode and "Time Division and Sharding" computing method cannot solve, and can meet the needs of new business development. The main contributions of this article are highlighted as follows:

- The motivations and scenarios for Det-CPN are analyzed. Computation-intensive and time-sensitive application has equally important requirements for deterministic communication and deterministic computation, driving the innovation and development of DetCPN.
- An architecture of Det-CPN with network determinacy and computing determinacy is proposed for time-sensitive and computation-intensive applications. And the technological capabilities and workflow of Det-CPN are presented.
- The key technologies of Det-CPN are introduced, including network determinacy technology and computing determinacy technology.
- The performance of Det-CPN has been verified through simulation experiments. In additon, the challenges and future trends of Det-CPN are also analyzed and discussed.

The remainder of the article is organized as follows. We provide the overview of CPN, and analyze the motivations and scenarios for Det-CPN. Then, the architecture of Det-CPN is proposed, and the technological capabilities, workflow as well as the key technologies are presented. And simulation results are presented to illustrate the performance of the proposed scheme. Following that, some research challenges and future trends are discussed. Finally, we conclude the article.

II. RESEARCH ADVANCE FOR CPN

Det-CPN can be considered as the next evolution paradigm of CPN. In order to better understand Det-CPN, we summarize and analyze the research progress of CPN in this section. By improving the design of network architecture and protocols, CPN connects distributed computing nodes and coordinates scheduling, achieving performance optimization as well as efficient utilization of computing-network resources. Currently, CPN has achieved initial results in system architecture, technological innovation, and standard specifications. In terms of architecture, CPN can be typically divided into centralized architecture and dis-

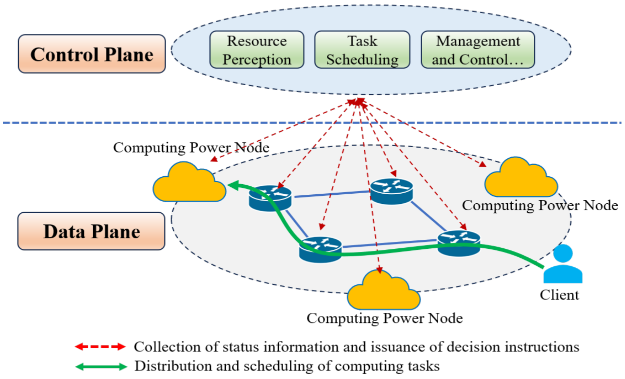

Figure 1. *The centralized architecture scheme of CPN*

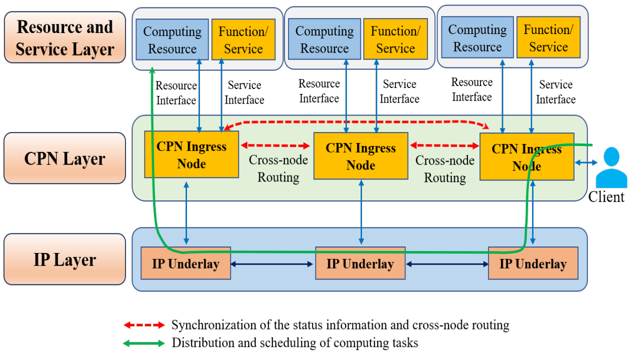

Figure 2. *The distributed architecture scheme of CPN*

tributed architecture.

In the centralized architecture scheme, as shown in Fig.1, the CPN system is divided into control plane and data plane. The control plane has a full resource view of the CPN, and makes unified computing power scheduling decisions. Through mechanisms such as centralized computing-network scheduling, it achieves routing addressing, distribution scheduling, and resource allocation of computing network resources.

Based on a centralized approach, a large amount of research has been conducted in both academia and industry. In [16], the authors proposed the concept of Sky Computing, which obtains resource service status information of distributed computing nodes through centralized methods, and then performs global unified task scheduling. In [17], the authors proposed a computing power networking framework for ubiquitous AI by establishing networking in AI computing-power pool. And this framework were designed to enable the adaptability for computing-power users, the flexibility for networking, and the profitability for computing-power providers. Moreover, the authors in [18] proposed a Service Intent-aware Task Scheduling framework for CPN to achieve the optimal matching of task intent and computing-networking resources.

In the distributed architecture scheme, as shown in Fig.2, the CPN achieves synchronization of the status information through the interaction between adjacent routing nodes. And the routing and forwarding of computing tasks are also completed during the decision-making of network nodes. And this scheme is usually implemented through network layer protocol extension.

Based on a distributed approach, the academic community has also conducted extensive research on CPN. In [19], the authors proposed the Compute First Networking technology solution, which is implemented using the Border Gateway Protocol (BGP) extension method. And it implements the synchronization of computing-network status information, routing and forwarding decisions of computing tasks at the network layer. In [20], the authors designed a scheduling strategy based on load balancing to allocate users' computing power tasks to an optimal computing power site by sensing the load and network status of each computing site. In [21], the authors explored the problem of network slicing and resource scheduling in compute first networking, and proposed a cooperative game model on the networking and computation resource allocation to optimize the system performance. In [22], the authors proposed a time-sensitive service coverage concept in the compute first networking, and then propose a novel framework to distributively shape the service coverage. In addition, the research on computing power networking, combined with technologies such as cloud native [23] and named data networking [24], has also attracted industry attention.

The standardization of CPN has also made significant progress. The International Telecommunication Union (ITU) in July 2021 approved CPN standards such as Y.2501 "Computing Power Network framework and architecture" [25], marking new progress in the internationalization of CPN standards. The Internet Engineering Task Force (IETF) has also established the Computing in the Network Research Group (COINRG) working group to carry out research and standardization work on intra network computing[26].

III. MOTIVATIONS AND SCENARIOS FOR DET-CPN

Det-CPN has deterministic guarantee capabilities in transmission and computation, which can achieve deterministic transmission and computation of computing tasks within constrained time. Compared to traditional computing power networking, Det-CPN can better meet the requirements of time-sensitive and computation-intensive applications. Det-CPN has important application value in emerging business fields that are time-sensitive and computation-intensive. In this section, we take intelligent driving, Cloud VR and intelligent manufacturing as examples to analyze the motivations and typical application scenarios of Det-CPN.

3.1 Intelligent Driving

Intelligent driving is a core application service in future intelligent society[27][28]. It relies on the vehicle's own cameras, millimeter wave radar, LiDAR, inertial navigation and other sensors for environmental perception, and then performs calculations, decisions, and control execution, which requires strong computing power support and strict low latency communication guarantees. However, massive perceptual data and complex computing tasks make it difficult to process at a lower cost on vehicle computing platforms. Therefore, the intelligent driving will move towards the trend of "cloud-network-edge-end" integrated development. The interconnection and integration of "cloud-network-edge-end" require a low latency, low jitter, and zero packet loss computing-network environment. Correspondingly, Det-CPN can provide a flexible, agile, accurate and deterministic computing-network resource service capability. Based on the latency requirements of intelligent driving applications, Det-CPN can achieve collaborative scheduling of computing tasks. Therefore, Det-CPN can empower the development of intelligent driving.

3.2 Cloud VR

Cloud VR refers to the introduction of the concepts and technologies of cloud computing and cloud rendering into VR business applications. Compared to traditional VR, Cloud VR has the advantages of reducing user terminal costs, improving VR resource utilization efficiency, and facilitating centralized content management in the cloud. However, Cloud VR has relatively high requirements for computing and network performance. Usually, the motion-to-photon (MTP) latency cannot exceed 20 ms. At the same time, the typical computing requirements include 8K H.265 real-time hard decoding and multi-channel parallel computing capabilities [29][30]. Users may experience dizziness if their viewing experience is repeatedly hindered by excessive latency. In order to avoid dizziness, the strong interactive services of Cloud VR require deterministic latency and jitter. Therefore, Cloud VR requires the support of Det-CPN.

3.3 Intelligent Manufacturing

With the development of the manufacturing industry towards intelligent and digital transformation, industrial control systems are gradually moving towards cloud deployment, and the process operations on the production site can be remotely controlled and processed to ensure production flexibility and safety[31]. At the same time, the cloud deployment of intelligent manufacturing also allows large enterprises to achieve production factor allocation and optimization between headquarters and multiple bases on a larger scale, achieving cost reduction and efficiency increase for enterprises. Therefore, in response to the trend of industrial control systems towards wide area and cloud development, Det-CPN can provide real-time computing power and real-time transmission guarantee for the next generation of industrial control systems. For example, deploying the factory control system in the form of cloud services on the cloud, transmitting the information collected by sensing devices with ultra-low latency and ultra-high reliability to edge computing nodes. Through rapid identification and decision-making, control instructions are quickly fed back to terminal devices and actions are executed.

IV. THE ARCHITECTURE, TECHNOLOGICAL CAPABILITIES AND WORKFLOW OF DET-CPN

Based on the introduction above, the definition and concept of Det-CPN are further clarified here. In fact, Det-CPN is an advanced stage in the development of CPN. Based on the deep integration of computing

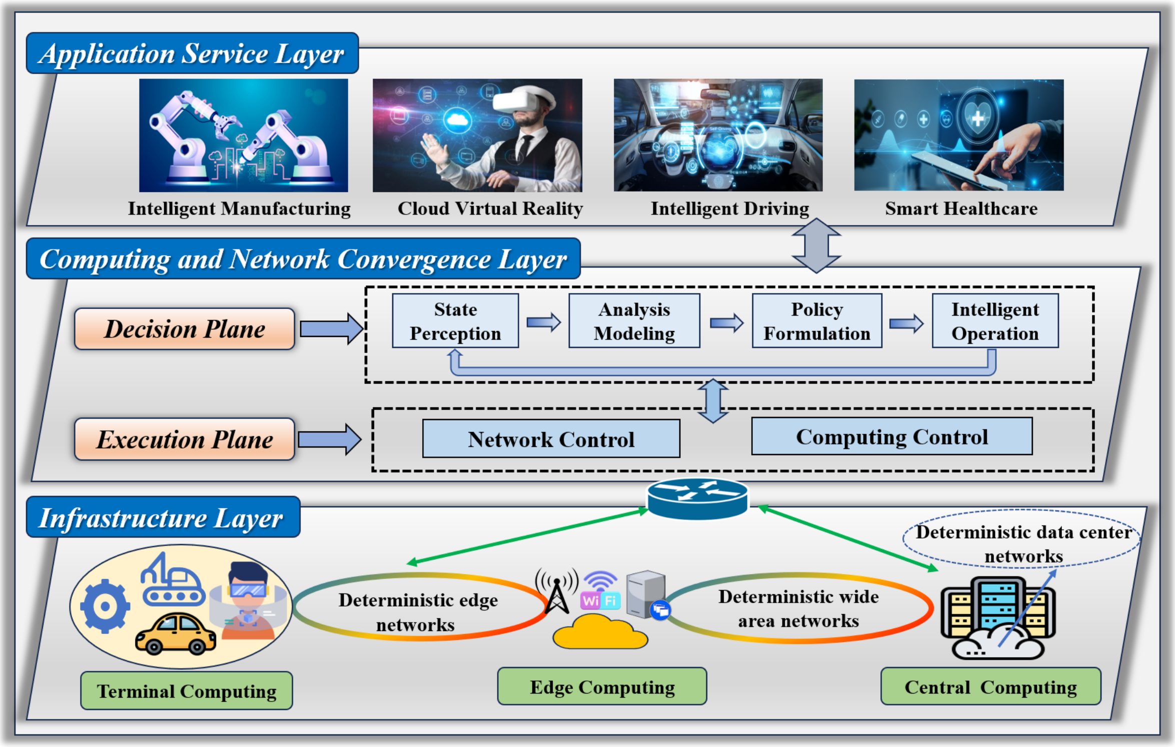

Figure 3. *The architecture of Det-CPN*

power and communication networks, Det-CPN fully considers the time constraints of new services in terms of computation and transmission. By adopting deterministic mechanism methods for transmission and computation, including task grading, resource reservation, resource pre-adjustment, etc., Det-CPN can achieve deterministic transmission and computation of computing tasks within the constrained time. In this section, we present the architecture, analyze the technological capabilities and introduce the workflow for Det-CPN system.

4.1 Architecture Design of Det-CPN

In this section, we present the architecture of Det-CPN. As shown in Fig.3, the architecture mainly consists of three parts, namely the infrastructure layer, computing and network convergence layer, and application service layer.

4.1.1 Infrastructure Layer

The infrastructure layer includes heterogeneous multi-level computing infrastructure and heterogeneous ubiquitous network infrastructure. Computing infrastructure provides heterogeneous computing power resources such as basic computing power, intelligent computing power and super computing power, usually including terminal computing nodes, edge computing nodes and central cloud computing nodes. Network infrastructure provides end-to-end network connectivity, including deterministic edge networks, deterministic wide area networks, and deterministic data center networks. The infrastructure layer provides basic computing and network resource capabilities for the upper layer.

4.1.2 Computing and Network Convergence Layer

The computing and network convergence layer is the core of the Det-CPN, including the computing and network decision plane (CNDP), the computing and network execution plane (CNEP).

The CNDP consists of state perception, analysis modeling, policy formulation, and intelligent operation. The CNDP obtains computing-network information through the functions of the state perception module, such as state detection, task perception, network perception, and intention perception. And then based on the state information, the analysis modeling module can achieve computing measurement, computing modeling, task deconstruction, and knowledge modeling. On the basis of analysis and modeling, policies such as network configuration, application orchestration, swarm intelligence learning, and intelligent distribution are formulated through the policy formulation module. At the same time, the CNDP achieves integrated services, intention driven decision-making, adaptive optimization, and autonomous operation and maintenance through the intelligent operation module. The CNDP will hand over the formulated strategies to the CNEP for execution.

The CNEP is divided into network control module and computing control module. The network control module mainly controls the network infrastructure, including edge networks, wide area networks, and data center networks, so as to achieve end-to-end deterministic and high-quality network control. The computing control module mainly manages and orchestrates the multi-level computing resources, achieving functions such as hierarchical and domain based computing, heterogeneous computing, computation offloading, and serverless scheduling. Its core capability is the deterministic computing processing of computing tasks. Namely, it constrains the processing delay of computing tasks to a fixed range, thereby ensuring that user terminal can obtain the results of computing task completion and return within the constraint time.

4.1.3 Application Service Layer

The application service layer provides users with various application services and is responsible for service operation. Application services mainly include computation-intensive and time-sensitive applications such as Cloud VR, intelligent driving, intelligent manufacturing, and smart healthcare.

4.2 The Technological Capabilities of Det-CPN

In order to achieve the design goals of Det-CPN, it is necessary to possess three core capabilities, namely, deterministic communication capability, deterministic computing capability and intelligent decision-making capability. In this subsection, we present these three technical capabilities.

4.2.1 Deterministic Communication Capability

In Det-CPN, the transmission of computing tasks usually has strict limitations on latency and jitter. Therefore, providing end-to-end deterministic communication guarantee for computing tasks is one of the core capabilities of Det-CPN. On the edge network side, the deterministic edge network can be constructed by introducing time sensitive networking (TSN) and "5G+TSN" technology. On the wide area network side, the deterministic wide area network can be implemented by introducing deterministic networking (DetNet) and segment routing (SR) technology. On the network side of the data center, the deterministic data center network introduces technologies such as intelligent lossless networks. And combined with software defined network (SDN) technology, the Det-CPN can achieve end-to-end deterministic communication, ensuring latency determinacy, jitter determinacy, and path determinacy.

4.2.2 Deterministic Computing Capability

In computation-intensive and delay-sensitive applications, the total communication and processing time of computing tasks are constrained, only ensuring communication determinacy cannot meet the delay requirements of some applications. Therefore, in order to ensure ultra-low latency in end-to-end communication and computation, and prevent incoming computing tasks from queuing up, it is necessary to promptly process computing tasks that arrive at computing processing units (such as CPUs and GPUs). If the traditional "time division and sharding" computing method is used, it will be difficult to ensure the processing delay of the computing tasks. Therefore, by designing mechanisms such as prioritization of computing tasks, preemption of high priority tasks, reservation and locking of computing resources, and elastic scaling of computing resources, deterministic computing can be achieved. It should be emphasized that computational determinacy refers to the time required to complete computing tasks within a bounded time range.

4.2.3 Intelligent Decision-making Capability

Intelligentization is the future development trend of CPN. Det-CPN also needs to possess the ability of network intelligence. Based on AI strategies and approaches, Det-CPN can achieve self-perception, self-configuration, self-optimization, self-decision, and self-maintenance. Specifically, the CNDP of Det-CPN needs to be based on intelligent strategy methods to achieve functions such as optimizing network service perception, computing-network task scheduling, and computing-network resource orchestration. On the other hand, considering that single point intelligence cannot meet the development of future computing-network intelligence services, Det-CPN also needs to have intelligent networking capabilities[32][33], that is, by interconnecting and sharing intelligent models, intelligent resources, etc., to improve the overall intelligence level of the computing power networks.

4.3 The Workflow of Det-CPN

In order to further understand the working mechanism of Det-CPN, we present the basic workflow of Det-CPN in this subsection, as shown in Fig.4.

Step1: Network state perception. The CNDP system collects the status information of network resources, then stores them in the status information database and updates them regularly.

Step2: Computing state perception. The CNDP system collects the status information of computing resources and computing services, then stores them in the status information database and updates them regularly.

Step3: User intention perception. The CNDP system also needs to collect user intention information, including latency, jitter, packet loss rate, computing power type, etc., so as to better provide customized computing-network services.

Step4: Analysis modeling and policy formulation. Based on computing-network state information and user intention information, the CNDP system will conduct analysis, modeling, and policy formulation.

Step5: Computation tasks scheduling. The CNDP system obtains the optimal computing node and transmission path through perception and analysis, and then makes the decisions for computing tasks scheduling.

Step6: Network control. According to the decision results of CNDP, the network controller in CNEP needs configure network devices to achieve routing and forwarding of computing tasks.

Step7: Computing control. The computing controller in CNEP adopts deterministic computing technology to efficiently handle target computation tasks.

Step8: Computing results return. After the computing node completes the task processing, the computing results are returned to the end user.

V. KEY TECHNOLOGIES OF DET-CPN

Det-CPN needs to integrate multiple key network technologies to achieve its goals, including edge network determinacy, wide area network determinacy, data center network determinacy, as well as computing determinacy on computing nodes. Therefore, in this section, we analyze and discuss the key technologies of the Det-CPN in detail.

5.1 Network Determinacy Technology

Network determinacy technology adopts technical mechanisms such as delay determinacy, jitter determinacy, and path determinacy, to construct an end-to-end deterministic network system, which is a key enabling technology for Det-CPN. In this paper, we divide the network determinacy technology into edge network determinacy, wide area network determinacy, and data center network determinacy.

5.1.1 Edge Network Determinacy

Edge network determinacy mainly refers to the access side network determinacy, and is the foundation for achieving deterministic computing power network access. It usually includes wireless edge networks and wired edge networks.

The determinacy of wireless edge networks usually requires 5G access network technology, including gigabit bandwidth access capability and millisecond level highly reliable transmission capability. In addition, Det-WiFi can also be applied to wireless edge networks [34], which support high-speed applications and provide better deterministic services in practical multi-hop edge environments. The determinacy of wired edge networks is mainly based on TSN technology [35]. TSN technology mainly focuses on the network link layer, with mechanisms such as clock synchronization, traffic shaping, resource reservation, path selection, and fault tolerance to ensure deterministic latency. Moreover, there is a demand for wireless and wired hybrid deterministic networking between

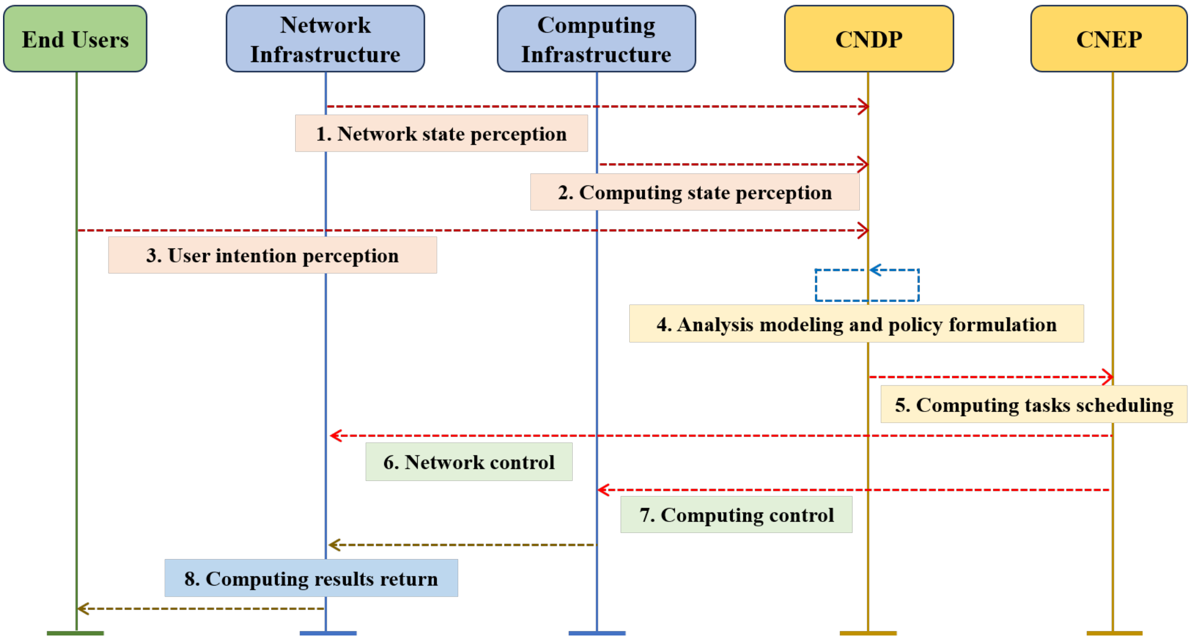
End Users
Network Infrastructure
Computing Infrastructure
CNDP
CNEP
1. Network state perception
2. Computing state perception
3. User intention perception
4. Analysis modeling and policy formulation
5. Computing tasks scheduling
6. Network control
7. Computing control
8. Computing results return

Figure 4. *The basic workflow of Det-CPN*

edge computing nodes, and “5G/B5G+TSN” is considered as an effective candidate for deterministic edge network [36]. Combined with the precise clock synchronization ability and deterministic traffic scheduling ability of TSN technology, edge network can ensure low latency and highly reliable transmission of various business flows, thereby providing high-quality and highly reliable edge network support.

5.1.2 Wide Area Network Determinacy

Wide area network determinacy technology is a new network technology aimed at large-scale and long-distance transmission, which provides deterministic service quality such as low latency, low jitter, low packet loss rate, high bandwidth, and high reliability. This technology is a collection of a series of protocols and mechanisms, achieving deterministic latency through mechanisms such as clock synchronization, frequency synchronization, scheduling shaping, and resource reservation. The deterministic jitter and packet loss rate can be implemented through mechanisms such as priority division, jitter reduction, and buffer absorption. And the deterministic reliability is achieved through technologies such as multiplexing, packet replication and elimination, and redundant backup. The representative technology for wide area network determinacy includes deterministic networking (DetNet) [10] and Segment Routing (SR) [37].

DetNet focuses on the network layer and defines deterministic algorithms for traffic queuing, shaping, scheduling, and preemption. By defining traffic control rules, it achieves deterministic jitter and packet loss rate. In Det-CPN, the use of deterministic networking technology enables the determination of latency and jitter in computing task transmission. The deterministic and non-deterministic services can be flexibly switched, and the level of deterministic service quality can be autonomously controlled. It can provide deterministic data transmission channels for computing task distribution and enable computing power collaboration across nodes, clusters, and regions.

SR is a source packet routing technique in which network nodes forward data packets based on an ordered list of instructions called segments. The segmented routing mechanism simplifies traffic engineering and management across network domains. Applying SR technology in Det-CPN can monitor the entire network topology and its traffic in real-time, and based on these data, determine the network transmission paths that computing tasks should pass through, and allocate bandwidth to these paths. Therefore, in Det-CPN, SR technology is used to achieve path determinacy, ensuring accurate distribution and transmission of computing tasks.

5.1.3 Data Center Network Determinacy

The data center network (DCN) is usually the “Last Mile” of the end-to-end deterministic networks, which plays an important role in Det-CPN. DCN determinacy refers to that the DCN has deterministic low latency capabilities, and it usually requires intelligent lossless network technology to provide a low latency and high throughput network environment for Det-CPN, thereby accelerating the efficiency of computing and storage, and greatly improving user experience [38]. Intelligent lossless network is a new type of low latency network that deeply integrates computing, storage, and network to improve and innovate in congestion control, flow control, packet forwarding, routing, and other aspects of the network [39]. Intelligent lossless network can provide a “zero packet loss, low latency and high throughput” network environment for Det-CPN, combined with intelligent lossless algorithms, thereby improving the performance of DCN. In addition, intelligent lossless network can be combined with other technologies to further expand their role in Det-CPN. For example, in combination with Remote Direct Memory Access (RDMA) technology, applications are used to directly read or write to remote memory, avoiding the intervention of operating systems and protocol stacks, achieving more direct, simple, and efficient data transmission, significantly reducing the time required in the data transmission process.

5.2 Computing Determinacy Technology

With the deployment of time-sensitive and computation-intensive applications in the cloud, CPN has increasingly high requirements for the computing determinacy. Det-CPN needs to ensure that computing tasks return execution results within the required latency of user terminals, subject to time constraints of end user business. Therefore, the traditional “time division and sharding” cloud computing method will no longer be applicable to Det-CPN scenarios. Deterministic computing technology needs to provide deterministic service quality with low computing latency, low latency jitter, high reliability, etc. It needs to provide stable and controllable deterministic service quality assurance for different businesses, and have the ability to respond to sudden large amounts of computing requests. To achieve deterministic computing, it is necessary to deeply improve the current CPN based on the key technologies such as task grading, resource reservation, and resource pre-adjustment.

5.2.1 Task Grading Mechanism

In computing determinacy, computing task requests can be divided into multiple priorities based on application latency requirements and computing load characteristics. Computing task requests with the requirements of lower latency and higher computational load should have higher processing priority. When task requests with different priorities arrive, high priority requests should be given priority in allocating computing resources to meet their latency requirements. When allocating resources for higher priority requests, it should be allowed to adjust the resource allocation strategy for lower priority requests and allow higher priority task requests to preempt computing resources.

5.2.2 Resource Reservation Mechanism

In computing determinacy, the computing resources of nodes should be reserved to respond to sudden computing requests. When computation-intensive and delay-sensitive task requests arrive at the computing node, task requests can be immediately allocated computing resources without the need to queue and wait. Due to the existence of redundant computing resources in computing power nodes, the performance of real-time processing for computing tasks will be greatly improved.

5.2.3 Resource Pre-adjustment Mechanism

In computing determinacy, serverless computing technology can be applied for server scaling to respond to sudden computing requests [40]. However, due to the current serverless computing technology, which is mostly based on server load level for dynamic scaling, the additional delay caused by cold start during scaling can affect real-time computing. The lagging scaling strategy during scaling will lead to unnecessary waste of network computing resources. Therefore, pre-adjustment of computing resources can be based on network resource situational awareness and task request prediction technology. Namely, when an increase in task requests is predicted, computing resources can be pre-deployed to the hot pool through cold start. When task requests increase, hot start can be used to avoid startup delay. When task requests are reduced, a pre-shrinking mechanism can be established to optimize computing resources. When the number of task requests drops to a certain threshold, timely recycling of partial computing resource allocation.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed deterministic computing power networking scheme using the numerical simulation method. We consider a simplified network topology consisting of terminal devices, several network nodes, and computing power node. Meanwhile, the communication path of computing tasks is determined. Especially, this article uses the network communication delay and processing delay of computing tasks as indicators to verify system performance. In this simulation experiment, we mainly compare deterministic computing power networking (Det-CPN) with ordinary computing power networking (Ord-CPN). Among them, ordinary computing power networking do not have the deterministic guarantee ability of network and computing power resources.

Here, we first introduce the model of computing delay. In this model, the computing tasks can be denoted as $\Gamma = \{\tau_1, \tau_2, \tau_3, \ldots \tau_N\}$. And each task τ_n is modeled by two parameters (S_n, C_n), where S_n is the input data size of τ_n, and C_n is the number of required CPU cycles to complete computing task τ_n. Namely, the computational workload of a computing task is represented by the number of CPU cycles. In addition, we assume there are $M = \{N_1, N_2, N_3, \ldots N_M\}$ computing power nodes, and the computation capability of computing power node N_m can be denoted as f_m, which is usually expressed by the computing frequency of the CPU. Hence, the time for completing computing task τ_n depends on the number of required CPU cycles C_n and the CPU frequency f_m of computing power node N_m [41][42]. Thus, the time of completing computing task τ_n can denoted as C_n/f_m.

Then, we introduce the model of network communication delay. According to [43], The delay at each hop can be subdivided into two main parts: fixed part and variable part. The fixed part usually includes the transmission delay at the sender and the propagation delay over the link. And the variable part usually includes the processing delay and queuing delay at the sender. In addition, the processing delay is dependant on the hardware's core processing power and memory access speed. In this paper, we assume that ordinary CPN and DetCPN have the same hardware's core processing power and memory access speed. Hence, queuing delay has become a key factor in evaluating network communication latency. In fact, compared to ordinary CPN, the advantage of DetCPN lies in optimizing queue latency. Therefore, this article mainly conducts comparative analysis through queue latency.

In DetCPN, the communication mode adopts the technical mechanism of deterministic networks, such as TSN. According to [44][45], the scheduling models of queuing delay can be classified into no-wait model and wait-allowed model. Here, we adopt the "no-wait" model so as to simplify the complexity of the simulation evaluation. Hence, the computing task data will not be queued and waited on network node devices, thus ensuring the "on-time and accurate" communication of data. On the other hand, the ordinary CPN adopts "Best-Effort" communication mode, which can lead to random occurrence of task data queuing and waiting, resulting in significant uncertainty in network latency.

Here, we introduce the parameter settings of simulation evaluation. In the model of network communication delay, we consider the propagation delay of one-hop is bounded between 2ns and 6ns. The processing delay of one-hop is bounded within 2us. In addition, we consider the network bandwidth is denoted as B, and the data size of computing task τ_n is donated as S_n, thus the transmission delay can be determined and denoted as S_n/B. And we assume the $B = 1000Mbps$, and the data size of computing task is $S_n = 10Mbit$. In the model of computing delay, the amount of computing task can be set to $C_n = [100, 200, 300, 400, 500, 600]Mcycles$, and we assume the computing node is single core processor and the CPU clock speed is set to 1.5GHz.

In Fig.5, we evaluate the performance of network communication delay versus network hops. From this figure, we can observe that the network communication delay increases with the network hops increases. This is because the latency of end-to-end communication for computing task increases with the number of network hops increases, including transmission delay,

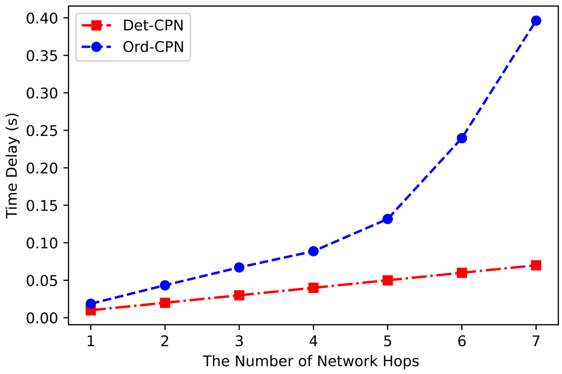

Figure 5. *Performance evaluation for network communication time delay versus network hops*

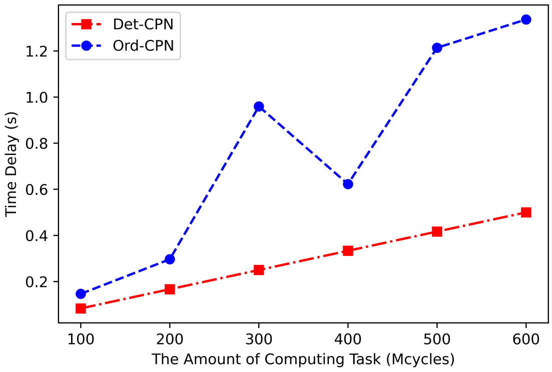

Figure 6. *Performance evaluation for computing time delay versus the amount of computing task*

propagation delay, processing delay and queuing delay. Moreover, under the same number of hops, the network communication latency of DetCPN is significantly lower than the ordinary CPN. This is because ordinary CPN adopts a “Best Effort” network communication mode, where computing task data has to queue and wait at network nodes, resulting in unstable and significantly increased of network communication latency. And the Det-CPN adopts a deterministic communication mode, which greatly improves network communication performance and significantly reduces communication latency.

In Fig.6, we evaluate the performance of computing time delay versus the amount of computing task. From this figure, we can observe that the computing delay generally increases with the increase of the amount of computing task. This is because with a certain amount of computing resources, the larger the amount of computation task, the more time is required. In Det-CPN, the computing resources can be specifically reserved for the target computing task, fully ensuring the computational processing requirements of the task. However, in ordinary CPN, computing resources are shared, and the target computing task can only obtain computing resources periodically or randomly, which cannot guarantee real-time processing of computing task. Hence, in ordinary CPN, the processing latency of computing task is not only relatively high but also fluctuating.

VII. THE CHALLENGES AND FUTURE TRENDS OF DET-CPN

Det-CPN paves the way for end-to-end deterministic communication and deterministic computing, while leaving some challenges to be discussed. In this section, we analyze some potential research issues that need to be discussed for future research.

7.1 Integrated Modeling and Control for Computing and Networking

Det-CPN is a new paradigm that combines network communication and computing processing, requiring unified scheduling and control for the networks and computing. However, due to the fact that network infrastructures typically belong to communication network operators and computing infrastructures typically belong to cloud service providers, the management and control of networks and computing power are separated, making it difficult to achieve integrated scheduling and control of computing-network resources. Therefore, it is necessary to build a unified Det-CPN operating system on top of communication network operators and cloud service providers, jointly model and control the network and computing resources, plan the total latency of task transmission and task calculation, so as to meet the communication and computing latency requirements of service businesses.

7.2 Integration of Forwarding, Computing and Caching

Currently, Det-CPN solution adopts an overlay approach, which controls the computing and network resources of the infrastructure layer to achieve deterministic communication and processing of computing tasks, without making fundamental modifications to existing network devices. Facing the future, in order to reduce the communication and computing latency of tasks, processing computing tasks in network devices may become a trend. The devices in Det-CPN will have forwarding, computing, and caching capabilities, namely, the ability of integration of forwarding, computing and caching [1]. When the network device receives a computing task, it will use the computing resources integrated by the network device for task processing and cache the results. The cached results can be used to simplify the computational cost of similar computing tasks. By deploying network devices that integrate computing and storage in network edge environments, the processing and response latency of computing tasks can be greatly reduced, meeting the new business requirements in Det-CPN.

7.3 Transmission Control Optimization

In Det-CPN, both network and computation adopt deterministic related technologies, greatly improving the reliability of data transmission and processing for computing tasks. Due to the current network adopting a layered design approach, the transport layer is decoupled from the network layer and link layer, and the transport layer is insensitive and untrustworthy of the underlying network conditions. Therefore, Transmission Control Protocol (TCP) is commonly used in the transmission layer to ensure the reliability of data transmission by increasing the complexity of the network protocol stack. With the adoption of technologies such as deterministic networks in Det-CPN, data transmission has highly reliable, high-quality, and predictable capabilities, and the traditional transmission control protocol mechanisms becomes redundant. Therefore, it is necessary to design new transmission control optimization methods, cut out complex and redundant traditional transmission control mechanisms or redesign them to ensure efficiency, simplicity and lightweighting. The transmission control optimization of Det-CPN can be achieved through technologies such as fine-grained congestion detection, network traffic prediction.

VIII. CONCLUSION

Providing end-to-end transmission and computing determinacy for computation-intensive and time-sensitive applications is of great significance. In this article, on the basis of research on CPN, we presented the architecture, technological capabilities and workflow of Det-CPN, and analyzed its application scenarios, key technologies. Moreover, the performance evaluation of Det-CPN are presented in comparison with ordinary CPN. Simulation results demonstrate that the proposed scheme can achieve better performance. Finally, the potential technical research challenges and future trends were discussed. In the future, we will study integrated modeling and control, integration of forwarding, computing and caching, as well as transmission control optimization in Det-CPN, and we will conduct application demonstrations based on China Environment for Network Innovations (CENI).

REFERENCES

[1] ZHOU Y, LIU L, WANG L, et al. Service-aware 6g: An intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.

[2] LU L, JIN P, PANG G, et al. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators[J]. Nature machine intelligence, 2021, 3(3): 218-229.

[3] Huawei Technology Report. Computing 2030 [R]. 2023.

[4] GROSSMAN E. Deterministic networking use cases[J]. IETF RFC 8578, 2019.

[5] SUN Y, LEI B, LIU J, et al. Computing power network: A survey[J/OL]. China Communications, 2024, 21(9): 109-145. DOI: 10.23919/JCC.ja.2021-0776.

[6] TANG S, YU Y, WANG H, et al. A survey on scheduling techniques in computing and network convergence[J/OL]. IEEE Communications Surveys & Tutorials, 2024, 26(1): 160-195. DOI: 10.1109/COMST.2023.3329027.

[7] TANG X, CAO C, WANG Y, et al. Computing power network: The architecture of convergence of computing and networking towards 6g

requirement[J]. China communications, 2021, 18 (2): 175-185.

[8] TANG Q, XIE R, FANG Z, et al. Joint service deployment and task scheduling for satellite edge computing: A two-timescale hierarchical approach[J]. IEEE Journal on Selected Areas in Communications, 2024, 42(5): 1063-1079.

[9] HUANG Y, WANG S, HUANG T, et al. Cycle-based time-sensitive and deterministic networks: Architecture, challenges, and open issues[J]. IEEE Communications Magazine, 2022, 60(6): 81-87.

[10] NASRALLAH A, THYAGATURU A S, ALHARBI Z, et al. Ultra-low latency (ull) networks: The ieee tsn and ietf detnet standards and related 5g ull research[J]. IEEE Communications Surveys & Tutorials, 2019, 21(1): 88-145.

[11] CAI Q, ZHOU Y, LIU L, et al. Collaboration of heterogeneous edge computing paradigms: How to fill the gap between theory and practice[J]. IEEE Wireless Communications, 2023, 31(1): 110-117.

[12] ZHOU Y, TIAN L, LIU L, et al. Fog computing enabled future mobile communication networks: A convergence of communication and computing [J]. IEEE Communications Magazine, 2019, 57 (5): 20-27.

[13] PENG G, WANG S, HUANG Y, et al. Enabling deterministic tasks with multi-access edge computing in 5g networks[J]. IEEE Communications Magazine, 2022, 60(8): 36-42.

[14] ZHANG W, GUO R, YANG D, et al. Detcncs: Deterministic computing and networking convergence scheduling[C]//Proc. the ACM Turing Award Celebration Conference-China 2023. 2023: 59-60.

[15] HU Y, JIA Q, CHEN J, et al. CoRaiS: Lightweight real-time scheduler for multiedge cooperative computing[J/OL]. IEEE Internet of Things Journal, 2024, 11(17): 28649-28666. DOI: 10.1109/JIOT.2024.3402257.

[16] YANG Z, WU Z, LUO M, et al. Skypilot: An intercloud broker for sky computing[C]//Proc. USENIX NSDI. 2023: 437-455.

[17] WANG X, REN X, QIU C, et al. Net-in-AI: A computing-power networking framework with adaptability, flexibility, and profitability for ubiquitous ai[J/OL]. IEEE Network, 2021, 35(1): 280-288. DOI: 10.1109/MNET.011.2000319.

[18] TANG Q, XIE R, FENG L, et al. SIaTS: A service intent-aware task scheduling framework for computing power networks[J/OL]. IEEE Network, 2023: 1-1. DOI: 10.1109/MNET.2023.3326239.

[19] KRÓL M, MASTORAKIS S, ORAN D, et al. Compute first networking: Distributed computing meets icn[C]//Proc. ACM ICN. 2019: 67-77.

[20] LIU B, MAO J, XU L, et al. CFN-dyncast: Load balancing the edges via the network[C]//Proc. IEEE WCNC Workshops. 2021: 1-6.

[21] WANG Z, ZENG D, GU L, et al. A game-based network slicing and resource scheduling for compute first networking[C]//Proc. IEEE GLOBECOM. IEEE, 2020: 1-6.

[22] QI J, SU X, WANG R. Toward distributively build time-sensitive-service coverage in compute first networking[J]. IEEE/ACM Transactions on Networking, 2023.

[23] YEHEZKEL A, ELYASHIV E, BARKAI S. Using CFN for uniform sampling of cloud-native datacenters[C]//Proc. IEEE CCNC. 2022: 967-968.

[24] CHEN H, TAO Y, ZHU Y. Nsacs-ps: A named service access control scheme based on proxy signature in named computing first networking [C]//Proc. IEEE HotICN. 2021: 81-85.

[25] ITU. Computing power network- framework and architecture: Y.2501[C]//ITU, 2021.

[26] ZENG D, ANSARI N, MONTPETIT M J, et al. Guest editorial: In-network computing: Emerging trends for the edge-cloud continuum[J/OL]. IEEE Network, 2021, 35(5): 12-13. DOI: 10.1109/MNET.2021.9606835.

[27] XIE G, XIONG Z, ZHANG X, et al. Gaiiov: Bridging generative ai and vehicular networks for ubiquitous edge intelligence[J/OL]. IEEE Transactions on Wireless Communications, 2024, 23(10): 12799-12814. DOI: 10.1109/TWC.2024.3396276.

[28] MA Q, XU H, WANG H, et al. Fully distributed task offloading in vehicular edge computing[J/OL]. IEEE Transactions on Vehicular Technology, 2024, 73(4): 5630-5646. DOI: 10.1109/TVT.2023.3331344.

[29] CHEN L, TANG Y, XIA J, et al. Multi-MEC collaboration for VR video transmission: Archi-

tecture and cache algorithm design[J]. Computer Networks, 2023, 234: 109864.

[30] JIA Q, XIE R, LU H, et al. Joint optimization scheme for caching, transcoding and bandwidth in 5g networks with mobile edge computing[C]// Proc. 2019 IEEE 5th International Conference on Computer and Communications (ICCC). IEEE, 2019: 999-1004.

[31] HU Y, JIA Q, YAO Y, et al. Industrial internet of things intelligence empowering smart manufacturing: A literature review[J/OL]. IEEE Internet of Things Journal, 2024, 11(11): 19143-19167. DOI: 10.1109/JIOT.2024.3367692.

[32] YU F R. From information networking to intelligence networking: Motivations, scenarios, and challenges[J]. IEEE Network, 2021, 35(6): 209-216.

[33] TANG Q, XIE R, YU F R, et al. Collective deep reinforcement learning for intelligence sharing in the internet of intelligence-empowered edge computing[J]. IEEE Transactions on Mobile Computing, 2023, 22(11): 6327-6342.

[34] CHENG Y, YANG D, ZHOU H. Det-wifi: A multihop tdma mac implementation for industrial deterministic applications based on commodity 802.11 hardware[J]. Wireless Communications and Mobile Computing, 2017, 2017(1): 4943691.

[35] ZHANG T, WANG G, XUE C, et al. Time-sensitive networking (tsn) for industrial automation: Current advances and future directions[J]. ACM Computing Surveys, 2023.

[36] SASIAIN J, FRANCO D, ATUTXA A, et al. Towards the integration and convergence between 5g and tsn technologies and architectures for industrial communications: A survey[J]. IEEE Communications Surveys & Tutorials, 2024.

[37] VENTRE P L, SALSANO S, POLVERINI M, et al. Segment routing: A comprehensive survey of research activities, standardization efforts, and implementation results[J]. IEEE Communications Surveys & Tutorials, 2020, 23(1): 182-221.

[38] HAN F, WANG M, CUI Y, et al. Future data center networking: From low latency to deterministic latency[J/OL]. IEEE Network, 2022, 36(1): 52-58. DOI: 10.1109/MNET.102.2000622.

[39] New H3C Technology White Paper. Intelligent lossless network technology white paper[R]. 2024.

[40] SHAFIEI H, KHONSARI A, MOUSAVI P. Serverless computing: a survey of opportunities, challenges, and applications[J]. ACM Computing Surveys, 2022, 54(11s): 1-32.

[41] SADATDIYNOV K, CUI L, ZHANG L, et al. A review of optimization methods for computation offloading in edge computing networks[J]. Digital Communications and Networks, 2023, 9(2): 450-461.

[42] JIA Q, XIE R, TANG Q, et al. Energy-efficient computation offloading in 5g cellular networks with edge computing and d2d communications [J]. IET Communications, 2019, 13(8): 1122-1130.

[43] ABBOU A N, TALEB T, SONG J. Towards sdn-based deterministic networking: Deterministic e2e delay case[C]//Proc. 2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021: 1-6.

[44] XUE C, ZHANG T, ZHOU Y, et al. Real-time scheduling for 802.1qbv time-sensitive networking (tsn): A systematic review and experimental study[C/OL]//Proc. 2024 IEEE 30th Real-Time and Embedded Technology and Applications Symposium (RTAS). 2024: 108-121. DOI: 10.1109/RTAS61025.2024.00017.

[45] STÜBER T, OSSWALD L, LINDNER S, et al. A survey of scheduling algorithms for the time-aware shaper in time-sensitive networking (tsn) [J]. IEEE Access, 2023, 11: 61192-61233.

BIOGRAPHIES

Qingmin Jia received the B.S. degree from Qingdao University of Technology, Qingdao, China, in 2014, and the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2019. He is currently a Researcher with the Future Network Research Center, Purple Mountain Laboratories, Nanjing, China. His current research interests include future network architecture, computing and network convergence, and deterministic networks, and edge intelligence.

Yujiao Hu is currently an associate professor in the School of Data Science and Artificial Intelligence, Chang'an University. She received the B.S. and Ph.D. degrees from the School of Computer Science, Northwestern Polytechnical University, in 2016 and 2021 respectively. From July 2021 to April 2025, she was a faculty member in Purple Mountain Laboratories, Nanjing, China. Her research focuses on computing power networking, multi-agent cooperation and UAV-oriented digital twins.

Xiaomao Zhou received the bachelor's and Ph.D. degrees from Harbin Engineering University, Harbin, China, in 2014 and 2020, respectively. From October 2016 to October 2018, he was a visiting Ph.D. student with Hamburg University, Hamburg, Germany. He is currently a Faculty Member with Purple Mountain Laboratories. He focuses on deep learning, edge intelligence, and computing power networking.

Qianpiao Ma received the B.S. degree in computer science and the Ph.D. degree in computer software and theory from the University of Science and Technology of China, Hefei, China, in 2014 and 2022, respectively. He is currently an Associate Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology (NJUST). His primary research interests include federated learning, mobile-edge computing, and distributed machine learning.

Kai Guo received his Ph.D. degree from the University of Tsukuba, Japan, in 2020. He is currently a researcher at Purple Mountain Laboratories, China. His research interests include computing and network convergence, mobile computing, and wireless networks.

Huayu Zhang received the B.E. degree from the Huazhong University of Science and Technology, China, in 2010, and the Ph.D. degree from Peking University, China, in 2017. He is currently a Researcher with Purple Mountain Labs, Nanjing, China. His research interests include distributed systems, graph theory, and next generation networks.

Renchao Xie received the Ph.D. degree in electrical engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2012. Currently, he is a professor in the School of Information and Communication Engineering at BUPT. His research interests include edge computing, information centric networking, and industrial Internet of Things. He has served as the Technical Program Committee (TPC) Co-Chair of numerous conferences. He has also served for several journals as a reviewer, including IEEE Network, IEEE Transactions on Communications, and so on.

Tao Huang received the BS degree in communication engineering from Nankai University, in 2002, and the MS and PhD degrees in communication and information system from the Beijing University of Posts and Telecommunications, in 2004 and 2007, respectively. He is currently a professor with the Beijing University of Posts and Telecommunications. His research interests include network architecture, and deterministic networks.

Yunjie Liu received the B.S. degree in technical physics from Peking University, Beijing, China, in 1968. He is currently the Academician of China Academy of Engineering, Chief Scientist of the Purple Mountain Laboratories. His current research interests include next-generation network, network architecture and management, software-defined networking, and deterministic networks.