Det(Com)^2: Deterministic Communication and Computation Integration Toward AIGC Services

Weiting Zhang, Nian Tang, Dong Yang, Ruibin Guo, Hongke Zhang, and Xuemin (Sherman) Shen

Abstract

As an emerging intelligence paradigm, artificial intelligence generated content (AIGC) is envisioned to be a key technique for Internet of intelligence, which inevitably puts forward higher requirements for the network capability from both the forwarding and computing perspectives. This article proposes a novel deterministic communication and computation integration architecture, that is, Det(Com)^2, for future networks to effectively support large AI model services such as distributed training, rapid deployment, and collaborative inference. Deep reinforcement learning-based solutions are developed to achieve cross-domain computation resource orchestration and deterministic transmission scheduling. The proposed learning-based solutions can efficiently schedule computing tasks of large AI models among multiple geographically dispersed computing domains while guaranteeing bounded latency and near-zero packet loss, thus facilitating integrated resource management and supporting large AI model services across their life cycles. Finally, we present a case study on communication and computation integration and discuss open research issues.

Introduction

With the development of artificial intelligence generated content (AIGC), such as chat generative pre-trained transformer (ChatGPT), large AI models have shown remarkable advantages in various fields [1]. On the one hand, large AI models are highly intelligent and adaptive through deep learning techniques, which are able to accurately understand and process human language, images, and videos. On the other hand, by analyzing large-scale data and extracting its personalized information, users can be provided with customized services that are not limited by fields, such as healthcare, finance, or entertainment. With these advantages, AIGC services derive many new applications in wireless networks [2]. For example, a network large model enables comprehensive network monitoring and management, thereby enhancing network reliability and the capabilities for fault detection and prediction. Consequently, AIGC are expected to introduce a new wireless networking ecosystem, and yield significant societal and economic benefits.

Widely deploying and applying large AI models in wireless networks faces many challenges due to their unique characteristics. *Firstly*, the construction of large AI models is generally composed of three stages, that is, training, deployment, and inference. Each stage requires not only communication resources but computation resources to support integrated scheduling for large AI models. Furthermore, each stage has specific requirements for the multi-dimensional resources, which may lead to the conflict of resource requirements in different stages and the competition for the same resource blocks. In addition, the training stage focuses more on improving learning speed and efficiency through parallel computing, while the inference stage focuses more on response time and quality of service (e.g., accuracy). Therefore, customized resource scheduling strategies need to be adopted to satisfy the diversified requirements [3]. *Secondly*, large AI models generate distinct data traffic at different stages. In specific, model parameters, binary text files, and user raw data are generated during the training, deployment, and inference stages, respectively. To avoid degradation of model performance due to incontinuity and inconformity transmission for the data traffic of large AI models, deterministic transmission should be considered at each stage of the large models [4]. Although time-sensitive networking (TSN) and deterministic networking (DetNet) can achieve deterministic data transmission with guaranteed bounded latency, jitter, and packet loss, the new data traffic complicates transmission scheduling [5]. *Thirdly*, constructing large AI models requires the collaboration of edge computing, cloud computing, and in-network computing to support large AI model services with a higher demand for computation resources[6]. In particular, to train a large AI model, high-dimensional (millions to billions) model parameters need to be transmitted repeatedly and asynchronously among multiple powerful computing centers [7]. Hence, it is paramount to design a communication and computation integrated network architecture to support the emerging large AI models in future wireless networks.

As illustrated in Table 1, a number of state-of-the-art deep reinforcement learning (DRL)-based solutions have been proposed, to realize flexible and intelligent resource orchestration and transmission scheduling [8]. Due to the complex coupled constraints between cross-domain computation resource orchestration and deterministic trans-

Weiting Zhang, Nian Tang, Dong Yang (corresponding author), Ruibin Guo, and Hongke Zhang are with Beijing Jiaotong University, China; Xuemin (Sherman) Shen is with University of Waterloo, Canada.

Digital Object Identifier: 10.1109/MWC.003.2300475

Ref	Problems	Performance Metrics	Decision Variables	Computation and Communication Integration	Deterministic Transmission	AI-assisted
[3]	Address the interactivity requirements of augmented information services	Minimize overall cloud network operation costs	A queuing system is designed to track the packet lifetime	×	✓	×
[4]	Delay-sensitive medical data transfer scheduling	Meet medical quality of service requirements including priority awareness and latency constraints	A truthful scheduling method is proposed to meet medical quality of service requirements with priority awareness and latency constraints	×	✓	×
[5]	Enhance collaborative learning for IoT to satisfy ultra-reliable and low latency requirements	Maximize scheduling success and learning accuracy	The cyclic queuing and forwarding (CQF) is used to achieve microsecond latency and near-zero packet loss in model parameter transmission	×	✓	✓
[6]	The collaboration between edge computing and cloud computing	Support edge computing services that require strong computing power to optimize resource utilization	A flexible scheduling scheme is designed to achieve computing, storage, and network integration	✓	×	×
[7]	Solve uneven resource utilization in cellular networks	More balanced cell load distribution than centralized DRL	A distributed DRL-based MLB method is proposed to improve the computational efficiency by dividing action space when the number of cells is increased	✓	×	✓
[8]	Reduce energy consumption and ensure network throughput in path management of MPTCP	Ensure the optimal path selection without additional delay and overhead	A multipath scheduler is introduced to select the best path adaptively through DRL and MPTCP model	×	×	✓
[9]	Optimize resource utilization for different QoS requirements of IoV services	Reduce the system cost while meeting QoS requirements	A two-layer constrained RL algorithm is proposed to effectively make resource and workload allocation decisions	✓	×	✓
Ours	Improve the overall performance of large AI models	Joint optimizing resource utilization and deterministic transmission	DRL-based algorithms are designed to optimize cross-domain computation orchestration and deterministic transmission scheduling	✓	✓	✓

TABLE 1. AI-assisted resource scheduling for deterministic communication and computation integration.

mission scheduling, it is challenging for a global DRL-based scheme to satisfy the rapidly changing network environment and diverse service requirements [9]. To this end, the original complex problems can be decoupled, and two DRL-based algorithms are designed separately to achieve efficient resource orchestration and deterministic transmission scheduling while ensuring the optimal decision performance.

In this article, we investigate how to effectively support large AI model services across life cycles while guaranteeing bounded latency, jitter, and packet loss, and improving computation resource utilization. The main contributions of this article are summarized as follows:

- We propose a deterministic communication and computation integrated network architecture for large AI models, which supports cross-domain computation resource orchestration and deterministic transmission scheduling via intelligent resource management.
- We present a multi-agent proximal policy optimization (MAPPO)-based resource orchestration algorithm, which enables efficiently cross-domain resource scheduling via multiple cooperative agents while optimizing the overall utilization of dispersed computation resources.
- We present a dueling double deep Q-network (D3QN)-based end-to-end deterministic transmission scheduling algorithm to accommodate the data traffic of large AI models at training, deployment, and inference stages while satisfying their diverse requirements.

The remainder of this article is organized as follows. A novel network architecture for large model construction is introduced in the next section. Learning-based resource orchestration and deterministic transmission scheduling algorithms are

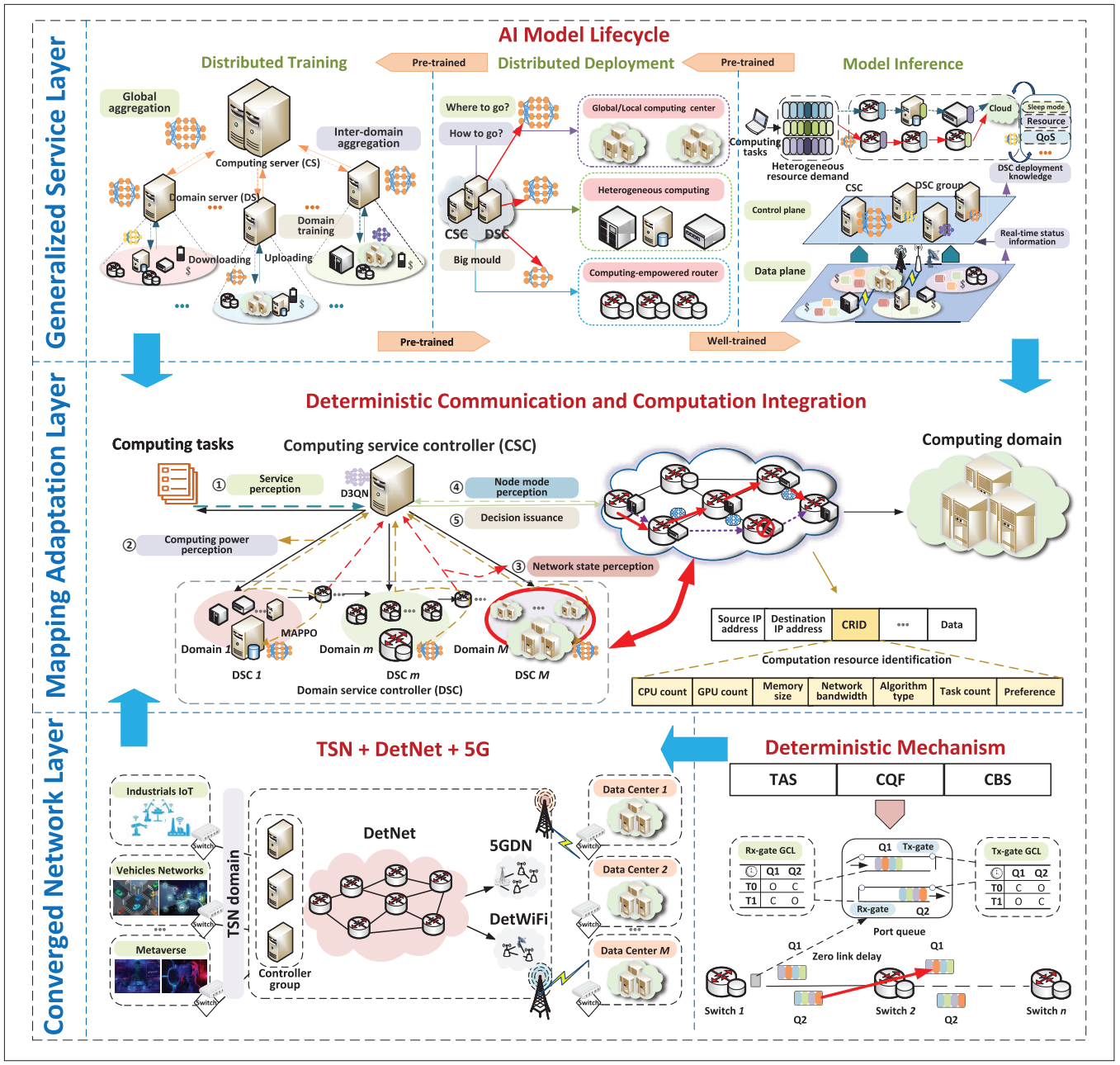

FIGURE 1. An illustration of the communication and computation integration architecture for large AI models.

presented following that. Then we give the simulation results. Following that we illustrate the future research directions, followed by the conclusion.

Novel Network Architecture toward Foundation Models

In this section, we propose a novel network architecture as shown in Fig. 1, which consists of three parts. The detailed functions of each part are described below.

Communication and Computation Integration

In traditional communication-dedicated network architecture, communication and computation are separated, and computing domains with different types of computation resources are managed independently. Hence, it is difficult to support cooperative resource scheduling cross multiple computing domains, which results in certain limitations and drawbacks in the construction and application of large AI models.

On the one hand, in the training and deployment of large models, the amount of data transmission is huge. If the data transmission only relies on the network, it faces long transmission time, bandwidth limitation, and network congestion, which cannot meet the demands for efficient data transmission and thus affecting the execution efficiency of computing tasks. On the other hand, the training and inference of large AI models require highly parallel computing to deal with complex computing tasks, and a single computing domain may not be able to meet such demands. However, the communication-dedicated network architecture lacks flexibility in the allocation of computation resources, resulting in uneven distribution of geographically dispersed computation resources.

Therefore, for the large AI models with huge data amount and low-latency requirements, the integration between communication and computation is crucial. To address these challenges, we propose a three-layer network architecture, which is composed of three layers, that is, generalized service layer, mapping adaptation layer, and integrated network layer. The specific functions of each layer are as follows.

Generalized Service Layer: This layer implements the description and representation of computing tasks generated by the three stages of large AI models. Taking the large model training as an example, a model parameter transmission task generated in the training stage can be described and expressed in terms of the data amount of the model parameters (e.g., dimension and total bytes), transmission speed (e.g., bytes or bits transmitted per second), transmission time (e.g., the ratio of transmission speed to data volume), and cost (e.g., network cost and bandwidth cost). Therefore, the scale, demand, and cost of model parameter transmission tasks can be accurately obtained, and the corresponding resource orchestration and transmission scheduling can be further carried out. A computing server (CS) and multiple domain servers (DS) are deployed in the generalized service layer to support efficient distributed training of large models. The CS initializes model parameters according to training task requirements and aggregates the updated model parameters from the distributed servers, and the DS receives global parameters from CS and performs local training to further update the parameters.

Mapping Adaptation Layer: This layer realizes the dynamic resource orchestration and transmission scheduling via communication and computation integration to support the three stages of large AI models while satisfying the diverse requirements of computing tasks. To achieve efficient resource orchestration and transmission scheduling, multiple domain service controllers (DSC) and a computing service controller (CSC) are deployed in this layer. Based on the quantitative description of computing tasks, the mapping adaptation layer can dynamically allocate communication and computation resources to computing tasks via the DSC and CSC. Specifically, the DSC aggregates resource information from the corresponding local computing domain and executes resource orchestration decisions. Meanwhile, the CSC executes transmission scheduling decisions and implements resource cooperative management across geographically dispersed computing domains to satisfy the diverse requirements of computing tasks.

Integrated Network Layer: This layer provides a stable and deterministic network environment for the entire architecture. According to the resource orchestration and transmission scheduling strategies of the mapping adaptation layer, the computing tasks are deterministically transmitted to the specified computing domain for processing with bounded latency, jitter, and packet loss. In this layer, the deterministic transmission mechanisms, for example, time aware shaper (TAS), CQF, or credit-based shaper (CBS), are deployed in the switches or routers of TSN and DetNet, and the 5G technology is used to provide high-bandwidth and large-capacity data transmission capabilities, which jointly satisfy the deterministic transmission requirements of computing tasks.

With these designs, the synergistic of communication and computation is fully utilized to enhance computing efficiency and guarantee transmission latency. Furthermore, it facilitates the unified management and collaborative scheduling of computation resources both inside and outside the network, thus promoting the construction and application of large AI models.

Large Model Training, Deployment, and Inference

The training, deployment, and inference stages of large AI models have different demands for communication and computation resources, and there are complex correlations among each stage. To address the above issues, the DRL-based method is adopted to realize customized and dynamic resource orchestration and transmission scheduling strategies.

Firstly, for the large model training, a huge amount of computation resources is essential to support highly parallel training tasks. A single computing domain (e.g., a data center) faces long training latency and low training efficiency issues caused by the limited computation resources. Secondly, for the large model deployment, the model scalability requires more consideration to achieve dynamic resource allocation, thus ensuring that the model can efficiently operate in a robust network environment. Due to the independent management among computing domains, resource allocation is generally uneven and it is difficult to perform collaborative scheduling. Thirdly, for the large model inference, it requires less latency and quick respond to user requests. Meanwhile, the elastic computation resources are needed to meet the peak demands. Hence, to address the above issues, it is necessary to achieve resource orchestration among multiple computing domains.

To solve the resource orchestration problem in different stages of large AI models, an MAPPO-based cross-domain computation resource orchestration algorithm is proposed, which allows multiple agents to learn in their associated computing domain while collaborating to achieve a common goal [10]. Through distributed learning and strategic collaboration among agents, the proposed MAPPO-based algorithm can dynamically optimize computation resource allocation in the computing domains according to the requirements of different stages, thereby ensuring the efficient operation of training, deployment, and inference [11].

Deterministic Transmission Scheduling

In the application scenarios of large AI models, such as Industrial Internet of Things (IoT), Internet of vehicles, and metaverse, the demand for data transmission and processing is stringent, which needs to satisfy the requirements of ultra-low latency and ultra-high reliability. For example, in industrial IoT scenarios, multi-dimensional sensor data needs to be transmitted with guaranteed bounded latency, jitter, and packet loss, and processed via a large AI model to accurately and instantly identify the machine conditions (e.g., normal or faulty). In this complex environment, efficient transmission scheduling becomes a key issue, especially for time-sensitive and critical data processing tasks.

To address the above challenges, a DRL-based deterministic resource transmission scheduling

The training, deployment, and inference stages of large AI models have different demands for communication and computation resources, and there are complex correlations among each stage. To address the above issues, the DRL-based method is adopted to realize customized and dynamic resource orchestration and transmission scheduling strategies.

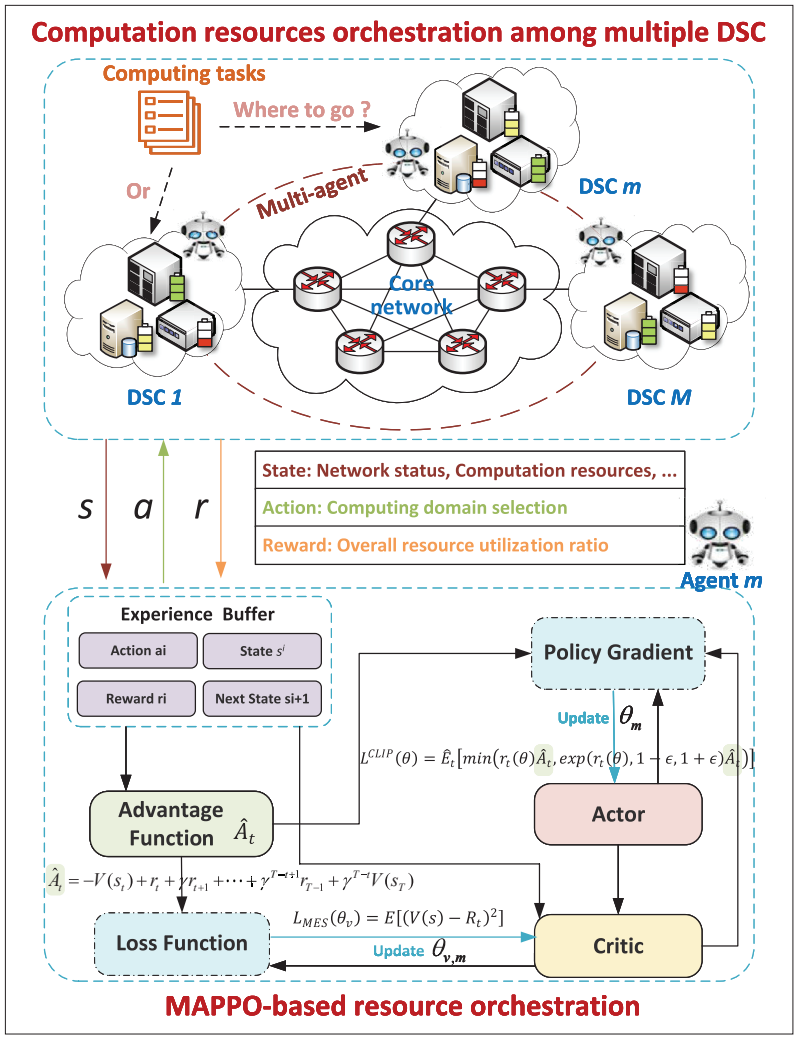

FIGURE 2. MAPPO-based resource collaboration orchestration algorithm.

algorithm is proposed by combining deterministic resource scheduling mechanisms and the D3QN algorithm [12]. Firstly, the integrated network architecture of TSN, DetNet, and 5G technologies supports both wireless and wired deterministic transmission. TSN ensures that all devices send and receive data packets in the local-area network according to the same schedule list, thus ensuring the predictable transmission latency of data packets. DetNet allows specific resources (e.g., bandwidth and queues) to be allocated for critical traffic (e.g., time-sensitive flow) in the wide-area network, thereby reducing latency and improving reliability. In addition, 5G networks provide high bandwidth and large capacity data transmission capabilities by using higher frequency bands, more antennas, and advanced modulation technologies. These technologies are combined to satisfy the strict requirements of large AI model services. Secondly, deterministic transmission mechanisms, for example, CQF, are deployed in the switches or routers to construct a transmission queuing model, thus achieving prioritized task transmission. Taking the CQF mechanism as an example, it consists of a cyclic timer and two transmission queues, and the queue states (i.e., open or close) can switch according to parity time slots. For each time slot, one queue can transmit packets and the other queue can receive packets. By adjusting the packet queues, the latency of data packets can be guaranteed during the transmission. Thirdly, the D3QN-based transmission scheduling algorithm is proposed to realize intelligent and flexible computing task scheduling by dynamically controlling the transmission sequencing of computing tasks. As such, the computing tasks of large AI models can be preferentially satisfied with deterministic requirements, therefore achieving low-latency and high-reliability transmission for computing tasks [13].

AI-Assisted Resource Orchestration and Transmission Scheduling for Foundation Models

In this section, the DRL-based cross-domain computation resource orchestration and deterministic transmission algorithms are introduced in detail.

Dynamic Resource Awareness

Dynamic resource awareness plays an important role in computation resource orchestration and deterministic transmission scheduling. It is to awareness the communication and computation resources by acquiring and analyzing the underlying heterogeneous resource information. In complex network environment, the available capacities of computation and communication resources dynamically change. The network should have the ability to sense these resources in real time to ensure that the communication and computation integrated network can obtain accurate resource information. The goal is to build a matching system that supports the synergy of resource orchestration and transmission scheduling. In specific, resource awareness involves the following steps:

- *Information collection:* The CSC and DSC sense a variety of real-time resource information from the servers deployed in multiple computing domains, such as CPU frequency, network bandwidth, available resources, remaining energy, and utilization price.
- *Feature analysis:* The CSC and DSC extract resource features (e.g., CPU or GPU is preferred) from the collected information by analyzing long-term statistical characteristics, and implement real-time monitoring (e.g., status, load, and availability) to update the information.
- *Automatic decision*: The CSC and DSC select the optimal resource supply, and make and execute resource orchestration and transmission scheduling decisions via DRL-based algorithms according to the dynamic matching between resource characteristic and task requirement.
- *Continuous optimization:* The proposed MAPPO-based and D3QN-based algorithms are deployed at the DSC and CSC, respectively, which gradually improve their ability to scheduling resources and optimize the efficiency of resource allocation through continuous learning.

Computation Resource Collaborative Orchestration

As shown in Fig. 2, to implement cross-domain computation resource scheduling, an MAPPO-based resource orchestration algorithm is proposed, in which multiple learning agents are deployed at the DSC to jointly perceive the resource states of the current network environ-

ment. Through iterative interaction among the learning agents, the global optimal resource orchestration policy can be obtained. In the proposed MAPPO-based algorithm, each agent observes the current resource state, for example, task requirements and available dispersed computation resources. Then, the agents allocate the multi-dimensional resources to the corresponding tasks, denoted by $\mathcal{I} = \{1, 2, ..., I\}$, through actor decision making, for example, which computing domain should be scheduled to and how much resources in the domain are utilized, and execute the resource orchestration decision to process the tasks. Finally, a reward is obtained based on the overall resource utilization ratio, where an orchestration policy that utilizes the dispersed computation resources more reasonable results in a higher reward. The core elements of the algorithm are defined as follows.

State: The state observed by the agent reflects the environment situation consisting of the aggregated information of DSCs, denoted by $\mathcal{M} = \{1, 2, ..., M\}$. We adopt the multi-agent learning framework, where each DSC can be considered as a single agent, and the state can be defined as $S = \{s_m\}_{\forall m \in \mathcal{M}}$. Here, $s_m = \{c_m, o_m, b_m, e_i\}_{\forall m \in \mathcal{M}, \forall i \in \mathcal{I}}$ is the state information perceived by DSC_m, $\forall m \in \mathcal{M}$, where c_m, o_m, and b_m denote the computation resources, storage resources, and communication resources of the m-th computing domain, and e_i denote the number of computation resources (e.g., processor cores) required for the computing tasks.

Action: The action corresponds to the DSC decision on resource orchestration of the computing tasks, which can be defined as $A = \{(a_{m,i}, d_{m,i}\}\backslash\backslash_{\forall m \in \mathcal{M}, \forall i \in \mathcal{I}}$. Here, $a_{m,i} \in [a_{1,i}, a_{M,i}]$ represents the scheduling decision of DSC_m, that is, whether to schedule the tasks to computing domain m. In addition, $d_{m,i} = \{p_{m,i}, g_{m,i}, n_{m,i}\}$ represents the resource allocation decision, where $p_{m,i}$, $g_{m,i}$ and $n_{m,i}$ is the allocated amount of computation, storage, and communication resources in computing domain m.

Reward: The agents obtain a reward from the integrated network environment to evaluate their resource orchestration decision under a specific state, respectively. In the algorithm, the reward function is to guide the optimization of computation resource orchestration policy of DSCs. Aiming at maximizing the overall resource utilization ratio of computing domains, the reward function can be defined as $r = \sum_m r_m$, where $r_m = \| \bar{u} \|_2 - \| \bar{u} - u_m \|_2$. Here, $\bar{u}$ and u_m are the target resource utilization ratio and the actual resource utilization ratio of the m-th computing domain.

Deterministic Transmission Scheduling

As shown in Fig. 3, to implement deterministic transmission for computing tasks, a D3QN-based end-to-end transmission scheduling algorithm is proposed. A centralized learning architecture is adopted, in which a learning agent is deployed at the CSC to perceive the states of TSN switches and communication link resources in the deterministic network environment. First, the agent observes the current network states (e.g., the source address and destination address of a computing task flow). Then, the task transmission scheduling decision (e.g., CQF queue selection and spectrum resource allocation) is made, and the decision is estimated via a current value function and a target value

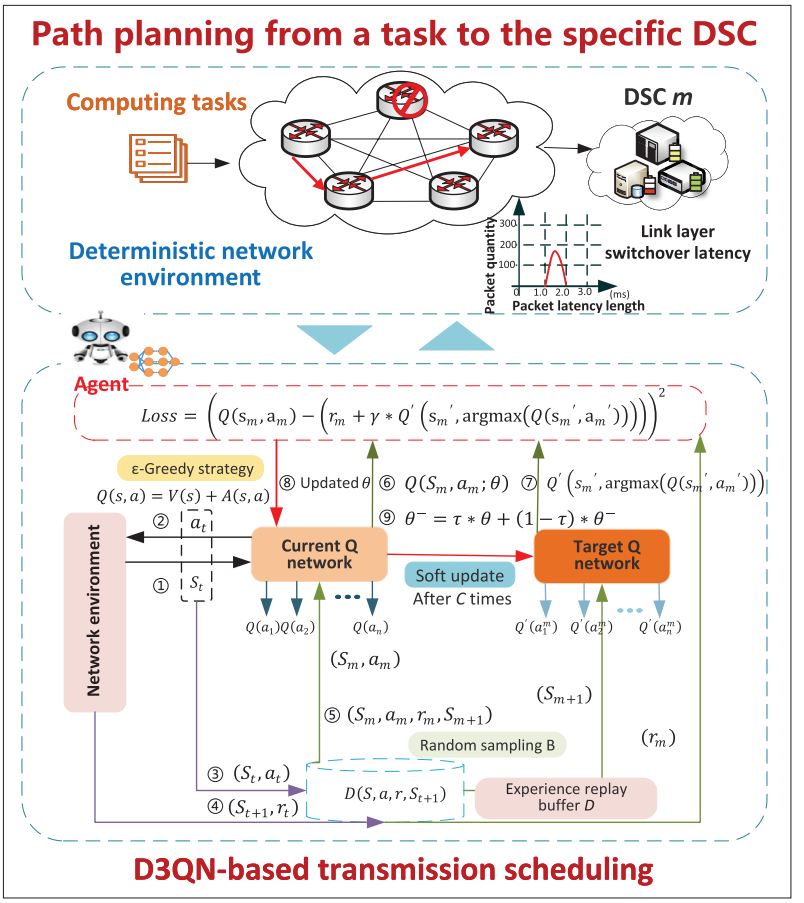

FIGURE 3. D3QN-based deterministic transmission scheduling algorithm.

function operated by two neural networks. Finally, a reward is obtained according to the system cost, for which the transmission scheduling decision with satisfied deterministic requirements of latency, jitter, and packet loss will obtain a higher reward. To obtain the optimal deterministic scheduling policy, the D3QN-based algorithm is iteratively updated through experience playback techniques. The core elements of the algorithm are defined as follows.

State: The state can be defined as $S = \{s_k\}_{\forall k \in \mathcal{K}}$, where $s_k = \{C_{TSN}, C_{5G}, F_i^s, F_i^d, T_i\}_{\forall i \in \mathcal{I}}$ is the state information perceived by the CSC in time slot k. Here, C_{TSN} is the TSN link capacity (i.e., the capacity of two CQF queues), and C_{5G} is the 5G link capacity. In addition, F_i^s, F_i^d, and T_i denote the source address, destination address, and the acceptable latency of the i-th computing task.

Action: The action corresponds to the CSC decision on deterministic transmission scheduling and path optimization for the computing tasks, which can be defined as $A = \{(q_{k,i}, b_{k,i})\}_{\forall k \in \mathcal{K}, \forall i \in \mathcal{I}}$. Here, $q_{k,i} \in \{0, 1\}$ is the queue state of CQF. If $q_{k,i} = 0$, the first CQF queue is close in time slot k, otherwise $q_{k,i} = 1$. In addition, $b_{k,i} \in [b_{min}, b_{max}]$ is the amount of 5G spectrum resources allocated to the current time slot.

Reward: As previously described, the reward function of the D3QN-based algorithm is to guide the policy optimization of deterministic transmission scheduling in CSC, which is defined as $r = \sum_k r_k$, where $r_k = \mu_s D - \mu_b B$. Here, D is the packet size, and B is the amount of occupied bandwidth resources for retransmission. In addition, μ_s and μ_b are the weight factors, respectively.

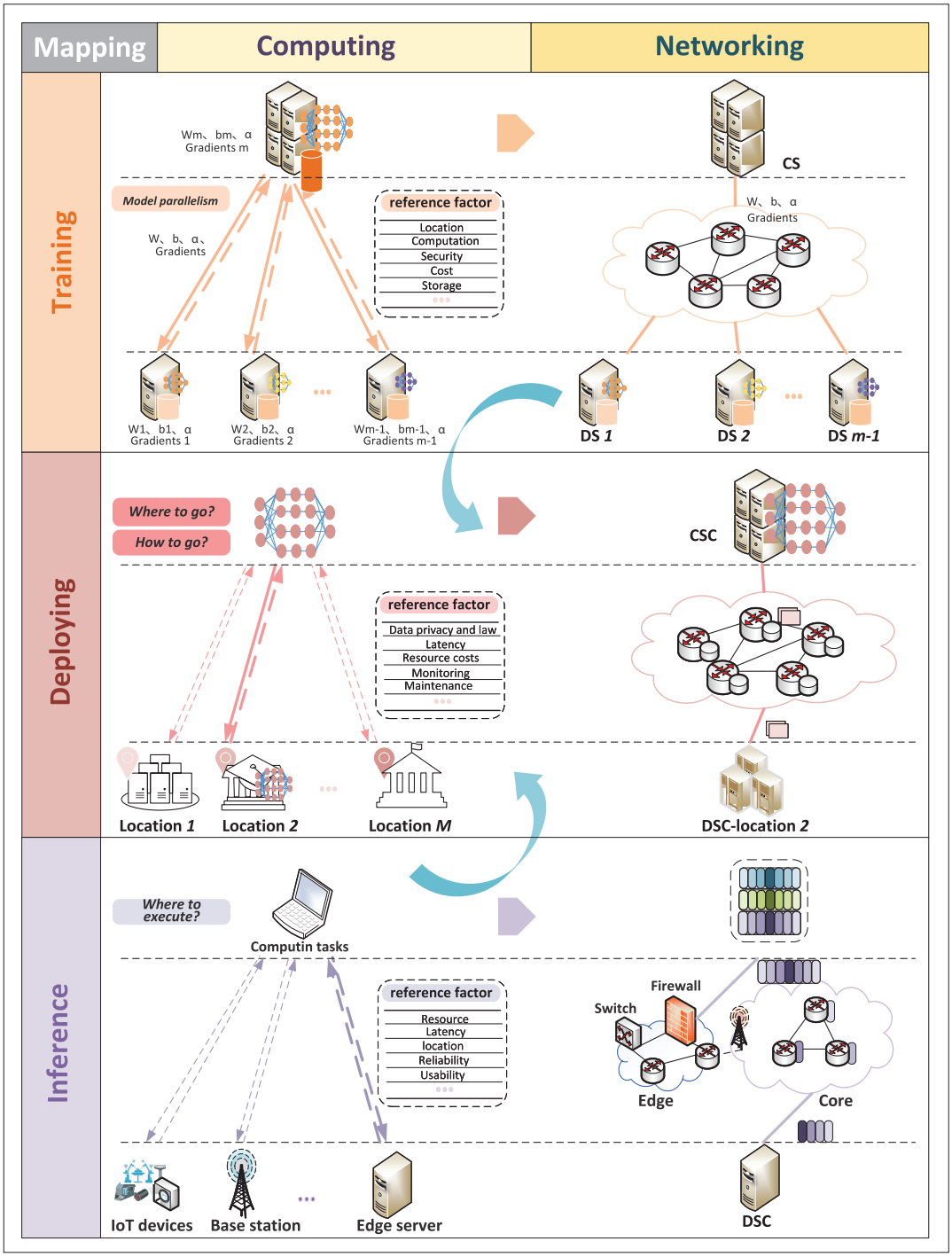

FIGURE 4. Mapping Mechanism between Computing and Networking.

Mapping Mechanism between Computing and Networking

To support efficient large model training, deployment, and inference, the mapping mechanism between the computing and networking is necessary, which is to achieve cooperative adaptation among diverse computing tasks, multiple computing service providers (edge servers and data centers), and heterogeneous network resources.

As shown in Fig. 4, the mapping mechanism is to transmit different computing tasks generated at the three stages of large AI models to the appropriate computing domain for processing through appropriate transmission paths. Firstly, the MAPPO-based resource orchestration algorithm is responsible for selecting certain computing domains to support the processing of computing tasks at the logical level and provides the identification information of target servers according to the characteristics of computing tasks and the load of computing domains. Moreover, the relevant information of the computing task (e.g., the description of the task, the data, and the identification of the target server) is packaged into packets and transmitted over the network. Secondly, D3QN-based deterministic transmission scheduling algorithm is combined with the current link condition to

determine the transmission path according to the address of the target server. Ensure that the packet reaches the destination server deterministically. Finally, the target server accepts and processes the packet, and the execution result is returned to the task initiator through the same network path.

Case Study

In this section, a deterministic transmission scheduling case study for computing tasks of large AI models is provided, aimed at improving the overall resource utilization ratio and reducing the total system cost.

Considered Scenario

We consider a communication and computation integrated network scenario (i.e., industrial IoT) that supports end-to-end deterministic transmission scheduling for computing tasks of large AI models (i.e., Transformer model), in which fault diagnosis of industrial machines is considered as computing tasks. In the scenario, 5 computing domains are considered to provide large AI model services, which are connected through TSN (TSN1, TSN2, and TSN3) switches and 5G core networks. Each computing domain is equipped with a number of CPUs and GPUs, which has 1000 cores and 1000 TFLOPS of computation resources. In addition, to support deterministic transmission, different bandwidth resources are allocated on each TSN switch, and three 5G spectrum resource blocks are considered, with subcarrier spacing (SCS) of 15KHZ, 30KHZ, and 60KHZ, respectively. To train a large AI model, we assume that 3000 computing tasks need to be transmitted between computing domain m_1 and m_2, and the date size of each task is set to 7 bits.

Considering the dynamic changes of link resources (e.g., time-varying wireless channels), a deterministic transmission scheduling optimization problem is formulated to minimize the total system cost when the number of successfully transmitted computing tasks is the same. The total system cost is defined as $P = \sum_{r=1}^{R} P_r$, where $P_r = \varepsilon_s P_s^r + \varepsilon_d P_d^r + \varepsilon_h P_h^r$. Here, P is the weighted sum of three cost components of all R computing tasks. Specifically, P_s^r is the link resource cost, which is the wireless spectrum resources and wired bandwidth resources consumed by a single transmission of the r-th computing task. The spectrum resources are allocated by a SCS in 15kHz, 30kHz, and 60kHz. P_d^r is the penalty for violating the latency constraints, which is to penalize the decisions when the service latency exceeds the maximum latency requirements. P_h^r is the retransmission cost for the lost packets, which is the extra spectrum and bandwidth resources consumed by the sender to re-transmit the lost packets. In addition, the weights of ε_s, ε_d, and ε_h are set to be 0.68, 8.43, and 0.45, respectively.

We propose a D3QN-based deterministic transmission scheduling algorithm to minimize the overall system cost. The D3QN-based algorithm adopts "online+target" framework, and both the online and target networks are three-layer fully-connected neural networks with 64 and 128 hidden neurons, respectively. The learning rate is set to 1×10^{-4}. For performance comparison, we adopt simulated annealing (SA), DQN-based, and DDQN-based transmission scheduling algorithms.

Simulation Results

As shown in Fig. 5a, the convergence performance of the proposed D3QN-based end-to-end transmission scheduling algorithm is evaluated. It can be seen that the D3QN-based transmission scheduling algorithm has converged after 1000 learning episodes, and the convergence performance is better than other benchmark algorithms.

As shown in Fig. 5b, the cumulative cost of a successful scheduling in terms of 3000 flows is presented. The result indicates that the D3QN-based transmission scheduling algorithms can reduce the overall cost of computing tasks by 18.9 percent, 9.1 percent, and 5.1 percent compared to the DQN-based, SA, and DDQN-based scheduling algorithms when the SCS is set to 60kHz. The reason is that the proposed D3QN-based algorithm can achieve more accurate scheduling performance than other benchmark algorithms via traversing more network states with the "dueling+double" structure, thus the cost of retransmission and latency violation is lower when successfully transmitting the same number of computing tasks. In addition, the simulation results show that the proposed D3QN-based deterministic transmission algorithm can reduce the packet loss rate while satisfying the latency and jitter requirements, and achieve a lower total system cost.

Future Research Directions

In this section, we discuss three future research directions for the construction and application of large AI models.

AIGC for Resource Optimization

AI-assisted resource optimization has received extensive attention, and many state-of-the-art schemes have been proposed to improve resource utilization efficiency. However, large AI models are more complicated, which requires collaborative use of geographically dispersed computation resources to support efficient training, deployment, and inference. With the emerging of generative AI techniques, AIGC is expected to be a new enabler for resource optimization in future networks, which has the potential to achieve dynamic resource allocation and scheduling while adapting to the changing requirements of computing tasks [1]. As such, the construction and application of large AI models in future networks would be more flexible and efficient.

Green Collaborative Computing

In the construction of large AI models, energy consumption has become an increasingly serious issue. In recent years, to realize energy-efficient training, deployment, and inference for large AI models, green collaborative computing has received wide attention [14]. Through distributed computing, model parallelism, and resource sharing, the unnecessary waste of computation resources can be greatly reduced. Hence, how to develop the corresponding solutions to support green collaborative computing is an important research issue. Moreover, multi-model joint training and cross-device collaborative inference are worth discussing to reduce resource consumption while ensuring model performance.

AI-assisted resource optimization has received extensive attention, and many state-of-the-art schemes have been proposed to improve resource utilization efficiency. However, large AI models are more complicated, which requires collaborative use of geographically dispersed computation resources to support efficient training,

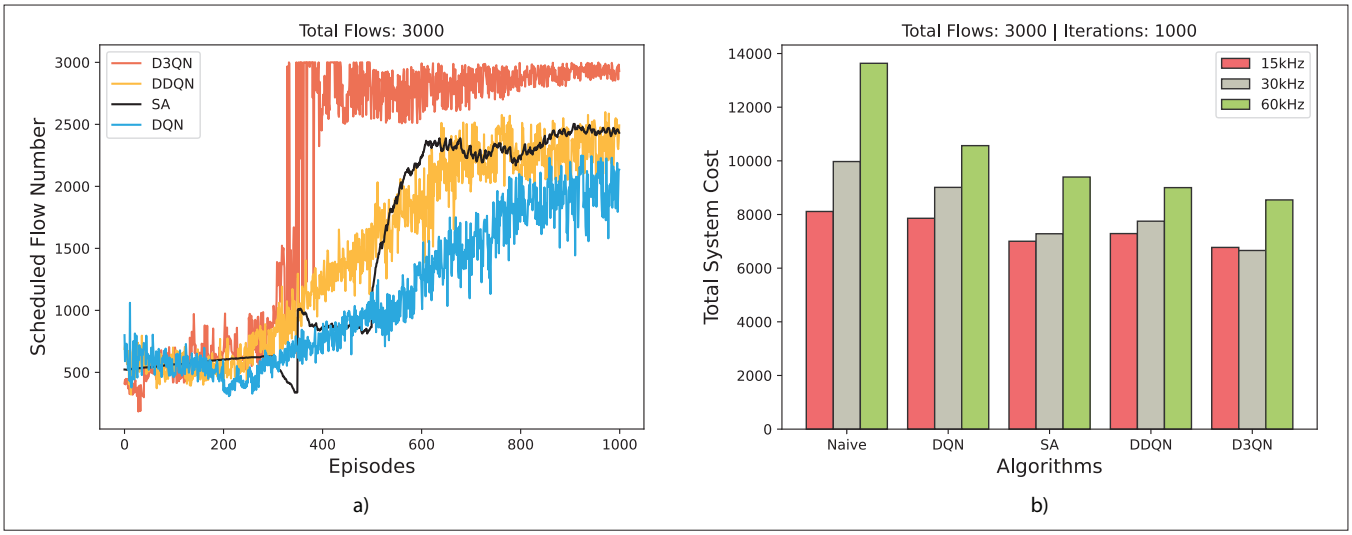

FIGURE 5. Performance evaluation of the proposed D3QN-based deterministic transmission scheduling algorithm: a) Convergence performance; b) Total system cost.

Secure Transmission for Large Model Interaction

To train a large AI model, huge amounts of model parameters need to be interacted among multiple computing domains. How to ensure the data privacy and model integrity during the interaction process is a key issue, especially when transmitting large AI models across dispersed servers or computing domains [15]. In particular, large AI models may involve sensitive information such as safety-related industrial data and private trade secrets. Hence, developing new secure transmission techniques, such as encryption algorithms, security protocols, and security hardware, deserve to be investigated to effectively guarantee the data confidentiality, model integrity, and transmission security during the interaction process.

Conclusion

In this article, we have proposed the integrated network architecture for deterministic communication and computation to support large AI model services in future networks. To achieve cross-domain computation orchestration and deterministic transmission, learning-based solutions have been developed, by which the diverse computing tasks of large AI models can be dynamically scheduled among multiple computing domains with the guaranteed transmission deterministic requirements of bounded latency, jitter, and packet loss. To facilitate the development pace of deterministic communication and computation integration, several research issues illustrated in the future research directions deserve to be explored.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62201029 and 62394321, in part by the China Postdoctoral Science Foundation under Grants 2022M710007 and BX20220029, and in part by the National Key Research and Development Program of China under Grant 2022YFB2901302.

References

[1] H. Du *et al.*, "AI-Generated Incentive Mechanism and Full-Duplex Semantic Communications for Information Sharing," *IEEE JSAC*, vol. 41, no. 9, June 2023, pp. 2981–97.
[2] P. Ding *et al.*, "Distributed Q-Learning-Enabled Multi-Dimensional Spectrum Sharing Security Scheme for 6G Wireless Communication," *IEEE Wireless Commun.*, vol. 29, no. 2, Apr. 2022, pp. 44–50.
[3] Y. Cai *et al.*, "Ultra-Reliable Distributed Cloud Network Control With End-to-End Latency Constraints," *IEEE/ACM Trans. Networking*, vol. 30, no. 6, June 2022, pp. 2505–20.
[4] C. Yi and J. Cai, "A Truthful Mechanism for Scheduling Delay-Constrained Wireless Transmissions in IoT-Based Healthcare Networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, Dec. 2019, pp. 912–25.
[5] D. Yang *et al.*, "DetFed: Dynamic Resource Scheduling for Deterministic Federated Learning Over Time-Sensitive Networks," *IEEE Trans. Mobile Computing*, vol. 23, no. 5, May 2024, pp. 5162–78, 2023, DOI: 10.1109/TMC.2023.3303017.
[6] X. Tang *et al.*, "Computing Power Network: The Architecture of Convergence of Computing and Networking Towards 6G Requirement," *China Communications*, vol. 18, no. 2, Feb. 2021, pp. 175–85.
[7] H. Chang *et al.*, "Decentralized Deep Reinforcement Learning Meets Mobility Load Balancing," *IEEE/ACM Trans. Networking*, vol. 31, no. 2, Aug. 2023, pp. 473–84.
[8] P. Dong *et al.*, "Multipath TCP Meets Reinforcement Learning: A Novel Energy-Efficient Scheduling Approach in Heterogeneous Wireless Networks," *IEEE Wireless Commun.*, vol. 30, no. 2, July 2023, pp. 138–46.
[9] W. Wu *et al.*, "Dynamic RAN Slicing for Service-Oriented Vehicular Networks via Constrained Learning," *IEEE JSAC*, vol. 39, no. 7, Dec. 2021, pp. 2076–89.
[10] W. Zhang *et al.*, "Optimizing Federated Learning in distributed Industrial IoT: A Multiagent Approach," *IEEE JSAC*, vol. 39, no. 12, Dec. 2021, pp. 3688–3703.
[11] D. Guo *et al.*, "Joint Optimization of Handover Control and Power Allocation based on Multi-Agent Deep Reinforcement Learning," *IEEE Trans. Vehicular Technology*, vol. 69, no. 11, Sept. 2020, pp. 13,124–38.
[12] N. Huynh *et al.*, "Optimal and Fast Real-Time Resource Slicing With Deep Dueling Neural Networks," *IEEE JSAC*, vol. 37, no. 6, Mar. 2019, pp. 1455–70.
[13] H. Hu *et al.*, "Intelligent Resource Allocation for Edge-Cloud Collaborative Networks: A Hybrid DDPG-D3QN Approach," *IEEE Trans. Vehicular Technology*, vol. 72, no. 8, Mar. 2023, pp. 10,696–709.
[14] W. Zhang *et al.*, "(Com)2Net: A Novel Communication and Computation Integrated Network Architecture," *IEEE Network*, Early Access, 2024, DOI: 10.1109/MNET.2024.3355922.
[15] B. Rong, "Security in Wireless Communication Networks," *IEEE Wireless Commun.*, vol. 30, no. 1, Feb. 2023, pp. 10–11.

Biographies

Weiting Zhang [S'20, M'21] (wtzhang@bjtu.edu.cn) earned the Ph.D. degree in Communication and Information Systems with the Beijing Jiaotong University, Beijing, China, in 2021. From Nov. 2019 to Nov. 2020, he was a visiting Ph.D student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Starting from Dec. 2021, he works as an associate professor with the School of Electronic and Information

Engineering, Beijing Jiaotong University. His research interests include industrial Internet of Things, deterministic networks, edge intelligence, and machine learning for network optimization.

NIAN TANG (22120127@bjtu.edu.cn) is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Beijing Jiaotong University, Beijing, China. Her research interests include industrial Internet of Things, deterministic networks, and machine learning for network optimization.

DONG YANG [M'11] (dyang@bjtu.edu.cn) received his B.S. degree from Central South University, Hunan, China, in 2003 and Ph.D. degree in Communications and Information Science from Beijing Jiaotong University, Beijing, China, 2009. From March 2009 to June 2010, he was a Post-Doctoral Research Associate with Jonkoping University, Jonkoping, Sweden. In August 2010, he joined the School of Electronic and Information Engineering, Beijing Jiaotong University. Since 2017, he is a Full Professor of Communication Engineering in Beijing Jiaotong University. His research interests include network architecture, wireless sensor networks, industrial network and Internet of Things.

RUIBIN GUO [S'23] (20111022@bjtu.edu.cn) received the B.E. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include deterministic networks, machine learning for resource allocation, and industrial Internet of Things.

HONGKE ZHANG [M'13, SM'16, F'20] (hkzhang@bjtu.edu.cn) received the M.S. and Ph.D. degrees in Electrical and Communication Systems from the University of Electronic Science and Technology of China, Chengdu, China, in 1988 and 1992, respectively. From 1992 to 1994, he was a Postdoctoral Researcher with Beijing Jiaotong University, Beijing, China, where he is currently a Professor with the School of Electronic and Information Engineering and the Director of the National Engineering Research Center on Advanced Network Technologies. His research has resulted in many papers, books, patents, systems, and equipment in the areas of communications and computer networks.

XUEMIN (SHERMAN) SHEN [M'97, SM'02, F'09] (sshen@uwaterloo.ca) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular ad hoc and sensor networks. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.