Toward Reinforcement-Learning-Based Intelligent Network Control in 6G Networks

Junling Li, Huaqing Wu, Xi Huang, Qisheng Huang, Jianwei Huang, and Xuemin (Sherman) Shen

Abstract

Reinforcement learning (RL) is a critical enabler for optimizing performance, automating the deployment, and increasing the intelligence level of 6G networks. In this article, we first identify some advanced RL frameworks for diversified 6G service scenarios. We then envision RL-based intelligent network management for 6G from three different perspectives: cross-layer end-to-end network control for service-oriented software-defined networking (SOSDN), cross-network control for global coverage, and cross-service control for service customization. We also present the new challenges associated with RL-assisted network management in 6G networks and provide potential research directions. Finally, we use the smart grid as a typical 6G application scenario to demonstrate the critical role of RL-based methods in capacitating intelligent power system management.

Introduction

With the tremendous growth in mobile device connectivity and the continuous changes of the communications environment, the diversity of network service scenarios, and the complexity of network control in the 6G networks will far exceed those of the existing networks. More specifically, the 6G networks are foreseen to support:

- *Global coverage*, where the terrestrial communication network expands into an integrated space-air-ground-sea communication network.
- *Versatile application*, where communications, computing, storage, sensing, artificial intelligence (AI), and big data are deeply integrated and applied to support various vertical applications (e.g., autonomous driving, healthcare, energy networks, industrial IoT, internet of vehicles, etc.).
- *Utter digitalization*, which aims to realize the vision of "intelligent connection of everything (e.g., human, machine, things, and environment) and digital twins" [1].

However, there are several technical challenges to making the above 6G vision a reality.

Rapidly Changing Environmental Dynamics: In 6G networks, users manifest highly differentiated mobility patterns and spatial locations with dynamic traffic demands. In addition, the network conditions, such as network topologies, link connectivities, and network capacities, also change rapidly over time. These dynamics introduce high uncertainties and pose significant challenges in real-time network management and consistent service provisioning. It is difficult to achieve real-time network adjustments by only relying on traditional model-based optimization methods, and the network operators urgently need new and more adaptive network management schemes.

Highly Diverse and Strict QoS Requirements: The QoS requirements of 6G services will become stricter in terms of peak data rate (Tb/s), user-experienced data rate (1 Gb/s), end-to-end delay (<1ms), reliability (≥ 99.99999 percent), and cost efficiency (≥ 500 Gb/$)[2]. These highly diverse and strict QoS requirements make it challenging to achieve user-centric customized service provisioning. The network operators will need novel network architecture and flexible resource management schemes to satisfy the diversified QoS requirements simultaneously.

Extremely Complex Network Control: The 6G network will constitute multi-dimensional heterogeneous resources (e.g., communication, computing, caching) on the local server and cloud. New 6G services demanding higher data rates, lower latencies, and a greater number of connections will result in a tremendous increase in data volume and network control complexity. Therefore, it is desirable to have a lightweight and natively intelligent network architecture to achieve intelligent and simplified network control.

To deal with the aforementioned challenges, AI is expected to be deeply embedded into every aspect and component of the network stack from the physical layer to the application layer. The integration of AI will automate the deployment, optimize network performance, and improve the intelligence level of 6G networks. As a key area of AI, reinforcement learning (RL) is a powerful tool for achieving intelligent network control in complex dynamic network environments. In RL, network elements (e.g., central controllers, network edge nodes, and mobile devices) are treated as intelligent agents. These agents behave autonomously, and, by interacting with the network environment, continuously improve their behavior using reward and punishment mechanisms to maximize their accumulated rewards. Through effective problem modeling and reward mechanism design, the RL framework can enable agents to adapt to the dynamic changes of the environment quickly.

Junling Li is with Southeast University, China; Huaqing Wu is with the University of Calgary, Canada; Xi Huang is with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, China; Jianwei Huang (corresponding author) is with The Chinese University of Hong Kong, Shenzhen and Shenzhen Institute of Artificial Intelligence and Robotics for Society, China; Qisheng Huang is with Harbin Institute of Technology, Shenzhen, China; and Xuemin (Sherman) Shen is with the University of Waterloo, Canada.

Digital Object Identifier: 10.1109/MNET.003.2200641

Nevertheless, the highly diversified 6G service scenarios make a universal RL framework infeasible. Instead, we have to explore more advanced RL frameworks for different network domains, performance indicators, and application scenarios. For example, federated RL (FRL) can help achieve privacy-preserved service provisioning and hierarchical RL is expected to realize large-scale decision-making. By suitable RL framework designs in different 6G network domains, the RL-assisted 6G network architecture is expected to achieve intelligent resource management, automatic network adjustment, and intelligent service provisioning.

In this article, we first identify some advanced RL frameworks for diversified service scenarios. Then, we envision RL-based intelligent network management for 6G from three different perspectives, that is, cross-layer, cross-network, and cross-service. Finally, we take the smart grid as a typical 6G application scenario where RL plays an essential role, followed by the final section that concludes this article.

Advanced RL Frameworks for Diversified 6G Service Scenarios

A unified RL framework that fits all demands is almost unattainable due to the highly diversified 6G service scenarios and performance indicators. For 6G networks, different RL frameworks have to be designed and employed in different network domains. In the following, we identify several promising RL frameworks for 6G architecture design.

Federated RL for Privacy-Preserved Service Provisioning

Privacy-preserving is crucial in network management in 6G networks. Federated RL (FRL) [3] can help achieve privacy-preserved service provisioning and speed up the learning process of RL by training a shared global model. In FRL, data or knowledge obtained by one agent can be transferred to other agents by policy sharing and policy aggregation. This allows the agents to make more effective use of the accumulated data of all agents while protecting local data privacy. Moreover, unlike traditional federated learning (FL), where the training data is available at each agent at the very beginning, FRL agents obtain their raw data by exploring the unknown environment over time. Different exploration schemes will generate different data samples, resulting in different qualities of the global RL model. Upon the optimal trade-off between agents' exploration and exploitation, the FRL framework may produce a high-quality global RL model that can be shared by all the agents.

Game Theory-Based Multi-Agent RL for Fair Network Resource Management

Many real-world applications in 6G networks can be modeled as non-cooperative multi-agent RL (MARL) problems, where a group of agents with different objectives continually interact and learn in a shared environment. To guarantee fairness among the agents and address the issue of non-unique learning goals during the learning process, it is imperative to apply game theory to the MARL framework [4].

For different 6G service scenarios, multiple equilibrium points at each state may exist, corresponding to different collective behaviors. Therefore, an equilibrium selection algorithm can help choose the most proper equilibrium to achieve in each state, which results in a stable learning process. Moreover, since equilibrium computation may involve high computational costs, it is essential to design a low-complexity equilibrium computation method to achieve the best performance-complexity trade-off.

For different 6G service scenarios, multiple equilibrium points at each state may exist, corresponding to different collective behaviors. Therefore, an equilibrium selection algorithm can help choose the most proper equilibrium to achieve in each state, which results in a stable learning process.

Hierarchical MARL Framework for Scalable Large-Scale Decision Making

The 6G networks are foreseen to involve massive heterogeneous agents and frequent information interactions, leading to high signaling overheads. To resolve this issue, we need to design a suitable learning framework to reduce the frequency of information interactions. The hierarchical MARL framework [5] holds great potential to accelerate the learning process, improve system scalability, and realize large-scale decision-making. Under this framework, the agents' optimal policies can be learned in different time granularities. At the top level, network-wide management is conducted at a slow time scale, while at the bottom level, agents' real-time decisions are tuned at a fast time scale. Through hierarchical MARL, it is expected to significantly reduce the frequency of information interactions among the RL agents, which in turn reduces the signaling overhead and improves the scalability of large-scale 6G networks.

RL-Enabled Cross-Layer Network Control

As a technological foundation of 5G network control, software-defined networking (SDN) will continue to play a vital role in designing 6G networks. For example, recent research has illuminated a trend of introducing SDN-based design to network optimization, such as radio access resource management, network request scheduling, and proactive caching for content delivery.

Different from state-of-the-art *operator-oriented network-level* service delivery, the development of 6G will shift the focus toward *user-oriented application-level* service customization. This requires 6G networks to concentrate on application-level performances while supporting diverse scenarios, calling for service-oriented software-defined networking (SOSDN) with cross-layer network control, as illustrated in Fig. 1. Different from state-of-the-art SDN, SOSDN introduces a service plane above network's control and data planes, to tailor the underlying network administrations to distinct application scenarios. To this end, SOSDN requires an integrated control protocol across different planes vertically and cooperative resource allocations across distinct applications horizontally. However, SOSDN has two major concerns:

- *Horizontal heterogeneity*, in terms of the intrinsic heterogeneity in resource and networking across different scenarios.
- *Vertical discrepancy*, in terms of the tension between service-level user experiences and network-level performances.

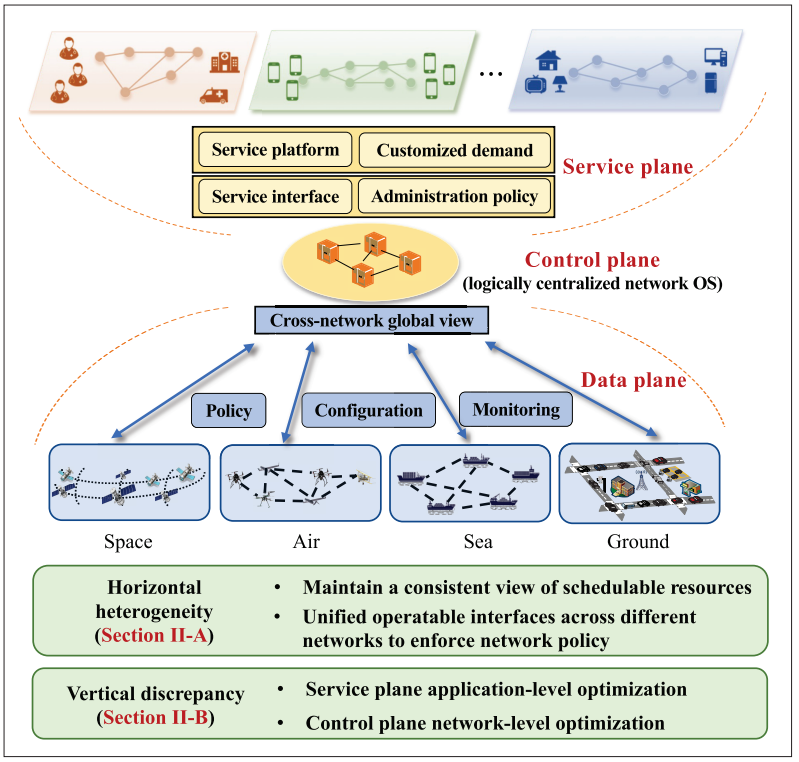

FIGURE 1. Illustration of SOSDN for 6G networks to realize cross-layer control through an integrated operation of the service, control, and data planes.

Facing such challenges, we envision RL as a promising enabler for utterly intelligent cross-layer network control under SOSDN, as explained below.

Collaborative Control across Heterogeneous Networks

Compared to 5G, a distinct feature of 6G networks is the extensive and adaptive coverage based on Space-Air-Ground-Sea Networks (SAGSN). SAGSN offers two advantages over existing network architecture. The first is a joint utilization of spectrum resources across existing heterogeneous isolated networks. The second is an effective collaboration among existing network infrastructures, achieved by flexible service coverage of space-air networks and dense end-device deployment for ground-sea networking.

Despite the advantages, the heterogeneity of existing SAGSN also brings tremendous challenges to network control for 6G networks. In particular, to monitor network dynamics, the development of control strategy, and the enforcement of network policy. Specifically, to maintain a consistent global view, SOSDN in 6G networks needs to monitor the dynamics of heterogeneous network infrastructures at multiple time scales regarding network topology, network failure, and network protocol. The heterogeneity implies that the difficulty lies in not only the characterizations of uncertain dynamics of different networks but also the monitoring strategies for each infrastructure. Meanwhile, the heterogeneity in communication protocols and transmission technologies makes it impossible to come up with any one-size-fits-all solution for the development and enforcement of network policies. To this end, we contemplate leveraging model-based RL [6] to capture network dynamics adaptively and use federated RL [3] to develop network monitoring strategies and keep the updated global view collaboratively.

Coordinated Optimization across Service/Network Layers

Another feature of the 6G networks is the versatility in performance optimization of application-level service across various scenarios, leading to potential conflicts between operator-oriented and user-oriented optimizations. Network operators often aim at network-level resource efficiency and service quality improvement. However, application-level user demands are usually considered independent of network traffic. Consequently, traffic optimizations do not necessarily lead to user satisfaction. Such objective misalignment makes autonomous network administration challenging in SOSDN. An alternative to tackling the problem is hierarchical multi-agent RL [5]. Specifically, by considering network administration as a task performed by agents, the RL-enabled master agent orchestrates the interplay and network policies proposed by two types of agents: one aiming at network-level performance optimization and the other optimizing application-level service quality for users.

RL-Enabled Cross-Network Control for Global Coverage

Global coverage is essential to ubiquitous service provisioning in 6G networks. However, the existing terrestrial networks alone cannot always offer reliable, flexible, and cost-effective services because of the limitations of scarce radio spectrum resources, high operational and maintenance costs, and geographically-constrained infrastructure deployment [7]. The SAGSN, as shown in Fig. 2, integrates terrestrial networks with space networks consisting of satellites, aerial networks formed by unmanned aerial vehicles (UAVs), and maritime networks with maritime unmanned surface vehicles (USVs). The SAGSN is cost-effective in supporting global network access for widely dispersed users and can flexibly adapt to dynamic service demands. However, network control in SAGSN encounters challenges caused by unparalleled service and device heterogeneity, unprecedented data traffic growth, and increasingly rigorous service requirements. Therefore, RL-based approaches have attracted increasing attention in addressing the high dynamics and complexity to empower intelligent and efficient network control.

RL-Based UAV/USV Trajectory Planning

In SAGSN, one crucial research direction is to leverage the agility and full controllability of UAVs and USVs to adapt to service demands and improve service quality. Considering that the trajectory planning of UAVs and USVs is a sequential decision-making process in uncertain network environments, RL-based approaches are a natural fit for UAV/USV control and trajectory planning problems. In [8], we proposed a hierarchical multi-agent deep RL (DRL)-based multi-UAV trajectory planning and resource allocation scheme for high-mobility users to maximize the accumulative network throughput while ensuring user fairness. In [9], Nguyen *et al.* proposed a DRL-based UAV trajectory planning algorithm to balance the trajectory flight time and total throughput while satisfying the QoS constraints. In [10], Su *et al.* presented a USV-aided marine data collection network and designed a target-oriented double

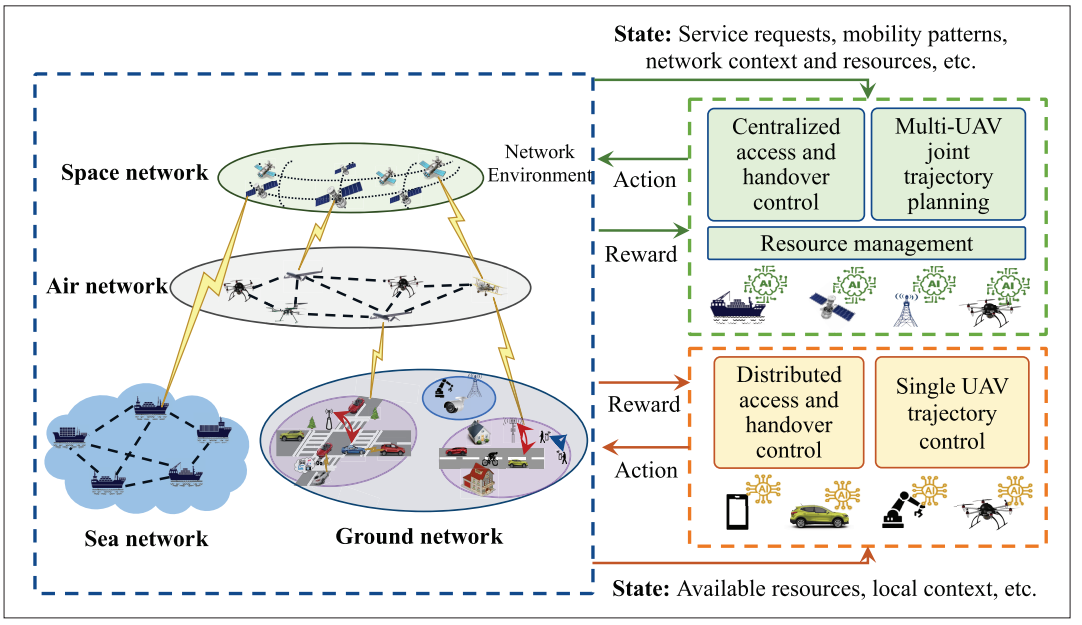

FIGURE 2. Illustration of the RL-enabled SAGSN.

deep Q-learning-based USV trajectory planning algorithm. By applying RL-based approaches, each position in the trajectory can be adaptively determined to accommodate the spatially and temporally-varying service demands.

RL-Assisted Mobility Management

In SAGSN, satellites, UAVs, USVs, and users have different mobility patterns, resulting in highly dynamic network topologies, intermittent network connections, and frequent handovers. Thus, efficient mobility management to cope with the constantly changing network node positions and channel conditions is imperative. RL-based approaches have great potential in supporting mobility-aware access control and efficient handover in SAGSN, guaranteeing seamless connectivity and uninterrupted service provisioning.

User access and handover control (AHC) problems are generally modeled as integer or mixed integer programming problems, which are NP-hard to solve. Traditional optimization approaches are often too dependent on the system modeling assumptions and are inadequate to solve these problems due to the excessive problem complexity. RL methods, without requiring prior knowledge or complete information of the network environment, are suitable to address the AHC problems for massive users. RL can be adopted for decentralized AHC by treating each individual user as an agent, which can interact with the environment to learn the characteristics of different network segments (space, aerial, terrestrial, and maritime networks) and the optimal AHC policy. In this case, each agent independently makes decisions to improve its own utility (e.g., to maximize throughput or minimize latency) based on local network information. Another potential solution is centralized RL-based AHC by regarding each access point (e.g., terrestrial BS, UAV, or USV) or edge controller as an agent, which jointly makes decisions for multiple users. In this centralized case, the agent needs to collect information from all related network components to learn a globally optimal strategy to improve network performance.

RL-Enabled Resource Management

RL-enabled resource management algorithms are promising in coping with the multi-dimensional resources (for communication, computing, caching, and sensing) from the space, air, terrestrial, and maritime networks in SAGSN. With RL-based resource management, the mobility pattern and traffic demand characteristics of different network nodes can be learned to facilitate precise and adaptive resource management. Different from supervised learning methods that output the predicted network environment, RL-based resource management methods learn the environment and provide the optimal resource management policy in a more intuitive fashion.

Resource management in SAGSN needs to be conducted with different granularities and timescales. For example, flying UAVs have fast-changing coverage and communication channel conditions, thus requiring real-time resource allocation decision-making. On the other hand, satellite resources allocated to different beams should be determined based on the service demand within the beam coverage, which needs to be adjusted with a relatively long period and coarse granularity. RL-based methods enable different granularities for resource management and network operation to support multifarious and customizable service provisioning in SAGSN. As shown in Fig. 3, RL-based resource management approaches can adapt to the time-varying number of users and effectively guarantee diversified service requirements, that is, reducing the delay of delay-tolerant services and satisfying delay-sensitive services' delay requirement of 10 ms.

New Challenges and Potential Research Directions for RL-Assisted SAGSN

Despite the great potential of applying RL-based approaches in SAGSN network control, various technical challenges remain. *First*, data collected from different network segments and operators generally have diversified data types, formats, defi-

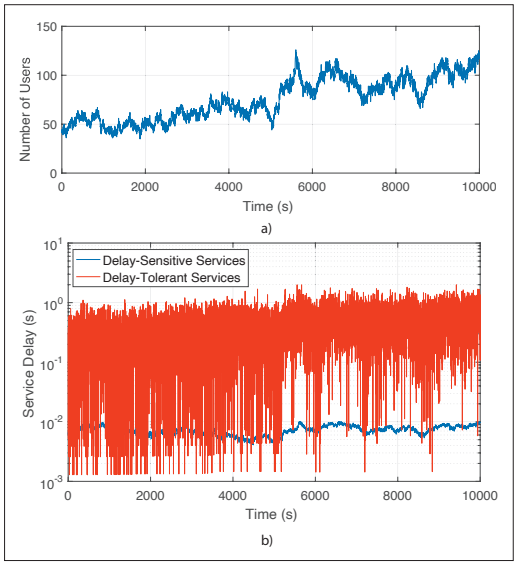

FIGURE 3. RL-based resource management for delay-sensitive services (with a delay requirement of 10 ms) and delay-tolerant services: a) time-varying number of users; b) service delay.

nitions, and granularities. Therefore, the difficulty and cost of effective data acquisition (including data collection, processing, and analysis) should be considered for RL-enabled network control. *Second*, the transmission of the tremendous amount of data required for RL model training consumes enormous bandwidth resources. Techniques such as edge computing and FL can be adopted for data pre-processing and de-redundancy to downsize the information exchange, which, however, may lead to important information loss. Therefore, the trade-off between RL-enabled network control performance and bandwidth resource consumption should be investigated. *Third*, in the large-scale and dynamic SAGSN, the RL-enabled approaches may suffer from the long training time and heavy calculation, especially for centralized RL with large action and state space. One potential solution is adopting MARL and hierarchical MARL frameworks to decompose the network system into multiple regional networks to downscale the action and state space. However, the lack of awareness of the global network status will lead to contradictory and inefficient decision-making. Therefore, the balance of overall performance and network operation efficiency and flexibility requires further study. *Finally*, users, UAVs, USVs, and satellites in SAGSN may struggle with the RL-based model training and inference due to the limited available energy and computing resources, which calls for the development of lightweight RL algorithms.

RL-Enabled Cross-Service Control for Service Customization

The 6G network is expected to be an intelligent network for providing user-centric and personalized service customization. RL-enabled cross-service control is an important solution to achieve agile and flexible 6G service customization. However, directly deploying RL on network elements (e.g., mobile terminals) to realize service customization may face several significant challenges.

First, the computing resources on the network elements are typically limited, which poses severe limitations on the scale of RL models and algorithms running on them. Second, the "trial-and-error" learning mechanism of RL requires the agents to interact with the unknown environment frequently to acquire sufficient data for optimal decision-making, resulting in high communications and computational costs. The cost of exploring the real network environment increases dramatically as the size of the network increases. Third, due to the highly diversified network scenarios, obtaining data that covers all possible scenarios is extremely difficult. Consequently, the RL model may not be able to provide optimal policies due to insufficient environmental explorations.

To address the aforementioned challenges, RL-enabled cross-service control is envisioned to be performed under the network function virtualization (NFV) and digital twin (DT) paradigms to achieve cost-effective 6G service customization.

RL-Enabled Cross-Service Control Under NFV Paradigm

Under the NFV network paradigm, the service customization is realized by splitting the network resources (e.g., communication, computing, and storage) into many virtual logical slices, each representing a customized end-to-end service. To better support customized services, heterogeneous resources must be allocated to different network slices through proper management and orchestration.

However, in a real-world 6G network scenario, user service requests arrive at the system dynamically with diverse QoS requirements, and the state of the underlying physical network changes rapidly over time. Existing static resource allocation strategies cannot meet the dynamic properties of NFV-enabled network services and may result in low resource utilization. RL algorithms running in the NFV management and orchestration module can help achieve efficient resource sharing among heterogeneous services and improve resource utilization. The role of RL is to gather the service requests' statistics and network status in real-time and learn the optimal resource allocation policy through interacting with the environment.

RL-based resource allocation is typically performed in three stages. First, given the service requests' properties and network conditions, the RL-based network orchestrator determines the resource demands for each virtual slice to form a virtual network topology. Second, the RL-based network orchestrator determines how to embed the virtual nodes and links in the slice onto the physical network cost-effectively. Finally, an RL-enabled virtual network function (VNF) scheduler determines the processing sequence of the VNFs embedded in the same physical node to minimize the overall completion time for the slices. By integrating with RL, NFV realizes cost-effective network element deployment for seamless device access, flexible VNF placement for service customization, and enhanced resource utilization to accommodate high traffic volume.

Specifically, in [11], the VNF scheduling problem is reformulated as a Markov decision process (MDP) and the VNF scheduler is regarded as an RL agent. Based on the specific representations of the MDP's components under the VNF scheduling circumstances, the agent can learn the optimal VNF scheduling policy with the objective of minimizing

the service completion time. Figure 4a illustrates the overall framework designed for RL-enabled VNF scheduling. The agent (i.e., VNF scheduler) makes the scheduling decision at the beginning of each time slot. The agent first determines a set of feasible actions based on the current system state. Next, it chooses a feasible action via the e-greedy policy based on the agent's current learned policy. The agent's reward, which reflects the service completion time and service end-to-end delay, is fed back to the agent from the environment, which is subsequently used to improve the VNF scheduling policy. Figure 4b shows an example VNF pattern determined by the learned optimal scheduling policy for a given time instant, where the interval between two vertical lines represents a time slot, the bricks in the same color represent the VNFs in a service request, and the vertical axis indicates the NFV node index. These results demonstrate the great potential of the RL-based approach in facilitating QoS-guaranteed service provisioning in future NFV-enabled networks.

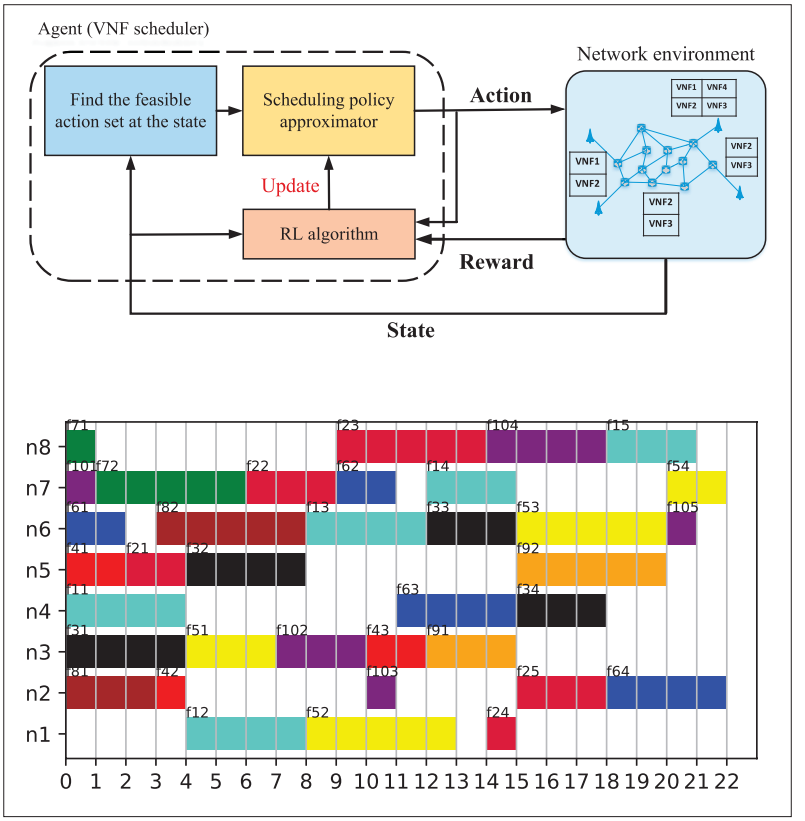

FIGURE 4. a) The overall learning framework for RL-enabled VNF scheduling; b) the VNF pattern determined by the learned optimal scheduling policy for a given time instant [11].

RL-Enabled Cross-Service Control Under DT Paradigm

The DT builds a mapping between the physical world and the virtual world. Every physical network element has a corresponding digital twin network element in a digital twin network (DTN). DTNs are usually deployed on edge servers or cloud servers equipped with powerful computing capabilities to avoid the limitation of computing resource scarcity on the network elements.

One of the major benefits of deploying RL in the DTN is the significant "trial-and-error" cost reduction due to the existence of a high-fidelity model of the real physical world. As shown in Fig. 5, the physical network acquires the data from the environment at different levels. It provides DTN with model parameters and training data to construct high-fidelity models. With the trained models for network elements, channels, and network slices, the DTN can simulate the behavior of different network parties and predict future network states. The RL agents can thus directly interact with the DTN, where DT simulates their rewards and decides on the final action to assist network performance optimization. The learned final policies will be fed back to the physical networks to guide the decision-making of network elements in real network environments. The network performance will be verified in physical networks and fed to the DTN to calibrate the DT models. By this iterative process, the simulation accuracy of the DT models will become sufficiently high to reflect the dynamic user behaviors and the network performance will converge to the maximized reward. By using RL to predict future network states and make real-time decisions, the DTNs are expected to help 6G networks realize the vision of self-optimization, self-management, and self-detection.

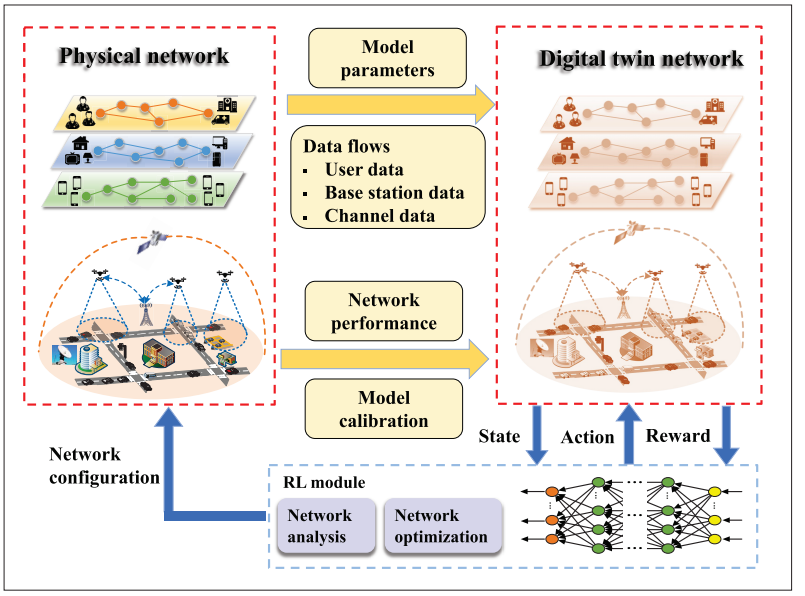

FIGURE 5. Exploring RL in DTN-enabled service customization.

RL-Enabled Intelligent Power System Management

6G is foreseen to support various vertical applications, including autonomous driving, healthcare, supply chain optimization, and personalized recommendations. One typical application of 6G is integrating smart grids with advanced wireless communication technologies. The seamless integration of smart grids with 6G networks will revolutionize the way we generate, transmit, and consume energy, creating more efficient, reliable, and sustainable energy systems. However, the fast-growing penetration of different kinds of resources creates severe challenges in the control, operation, and management of the smart grid. RL is considered a promising method to overcome these challenges through real-time data-driven control and management. Furthermore, devel-

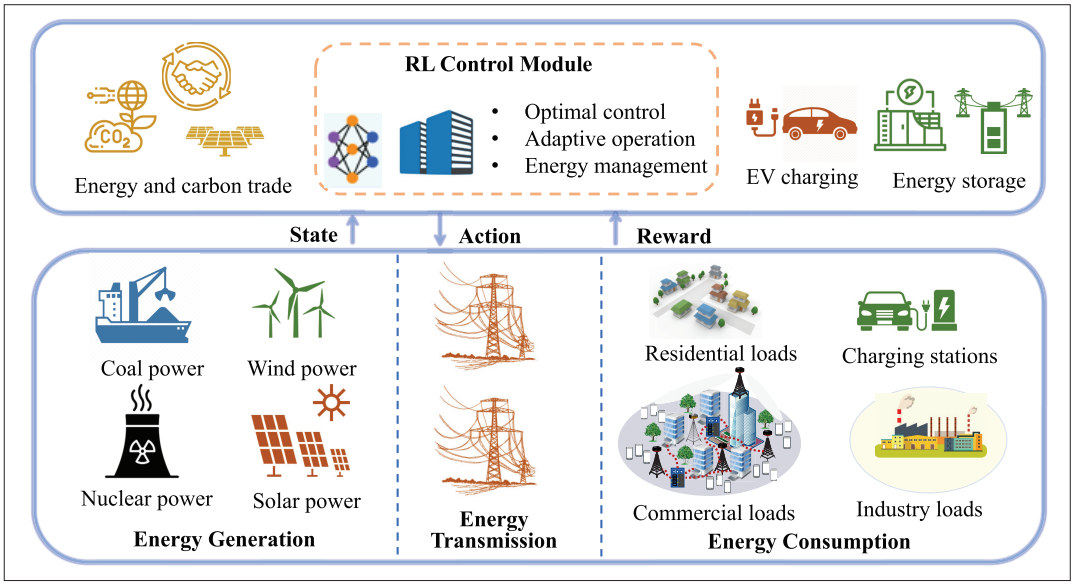

FIGURE 6. RL-enabled smart grids for the 6G vision of versatile applications.

oping more advanced communication and computing technologies, such as 6G and advanced metering infrastructures (AMIs), enables the versatile deployment of RL in smart grids, as shown in Fig. 6. In this section, we present some typical RL-based application scenarios in the smart grid.

RL-Based Energy Storage Operation

Energy storage can help balance the energy demand and supply in real-time. Merchant energy storage owners seek to maximize their profit by making energy storage operation decisions. There are a number of studies that have investigated the optimal energy storage operation decisions in electricity markets [12]. Wang *et al.* in [13] first proposed a temporal arbitrage policy for energy storage in electricity markets via reinforcement learning. The authors designed a reward function that incorporates the history information as well as the instant profit of charging/discharging decisions. However, RL-based optimal energy storage operation problems in joint energy and frequency regulation markets remain unexplored.

RL-Based Electric Vehicle Charging

Electric vehicles (EVs) play an essential role in power systems and transportation systems due to their low carbon and flexibility features. Adopting RL algorithms can solve the optimal EV charging and operation problems. Jin et. al in [14] proposed an RL-based method to study the charging routing problem, where a smart EV wants to fulfill its battery charging demand by seeking an EV charging station. A real-world transportation network has been adopted to evaluate the performance of their proposed approach, showing superior performance in comparison with the classical actor–critic (A2C) method. However, human beings' behaviors have a significant impact on the optimal scheduling and charging decisions of charging stations. Hence, it is important to design RL-based methods that capture the dynamic features of the consumers' charging behaviors.

RL-Based Energy and Carbon Trading

RL can help market players learn their optimal bidding strategies in electricity markets. In such RL-based trading frameworks, the market players repeatedly interact with the market-clearing process (i.e., environment) and utilize the experiences acquired from the interactions to improve their bidding strategies. For example, in [15], Ye et. al presented a deep reinforcement learning (DRL) based methodology, which combines a prioritized experience replay (PER) strategy with the deep deterministic policy gradient (DDPG) method. They numerically demonstrated that their proposed methodology could increase the profit significantly compared to the state-of-the-art methods. An interesting direction is to consider RL in the joint energy and carbon emission trading market, which helps power generation companies dynamically manage their energy production and trade their carbon emission quotas.

New Challenges and Potential Future Directions for RL-Enabled Smart Grid Applications

Despite the versatile applications of applying RL-based approaches in smart grid control and management, there exist several critical challenges. First, a smart power network is a vital part of modern society. We must ensure that any controllers in the smart grid do not lead to violations of network physical constraints or system reliability issues. Therefore, safety is a major issue of applying RL in a smart grid. Second, most of the existing literature conducts tests on a small-scale test bed with a few decision-making agents and a few buses. It is crucial to deal with the scalability issues of RL in large-scale multi-agent power systems. To deal with the above challenges, there are several potential future directions. First, we can integrate model-free and model-based methods together and obtain the advantages of both to help deal with the safety and scalability issues. Second, we can properly use different variants of RL in the literature, such as multi-agent RL, robust RL, transfer RL, inverse RL, etc., to tackle the safety and scalability issues.

Conclusion

In this article, we investigated how to empower 6G with reinforcement learning (RL) toward

utter intelligent network control from three aspects: RL-enabled cross-layer network control under the SDN paradigm; RL-enabled cross-network control in the SAGSN for global coverage; and RL-enabled cross-service control under the NFV and digital twin paradigms to achieve agile 6G service customization. A related case study has also been presented. The visions presented in this article shed light on the development of RL-enabled intelligent network management in highly dynamic, diversified, and complex 6G environments. Hopefully, the challenges discussed in this article will inspire continuous discussions and more research efforts on RL-enabled 6G to promote pervasive network intelligence. For future research, some open issues deserve further investigation, including proper economic mechanism designs under the RL framework to ensure fairness of user experience and resource sharing; and adaptive orchestration schemes between RL model performance and resource consumption to realize cost-effective network management.

Acknowledgment

This work is supported by the Guangdong Basic and Applied Basic Research Foundation (2021A1515110949, 2021B1515120008), the Start-up Research Fund of Southeast University under Grant RF1028623062, Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2023-03759, University of Calgary Start-up 10040064, the National Natural Science Foundation of China (Project 62271434), Shenzhen Science and Technology Program (Project JCYJ20210324120011032), Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (No. ZDSYS20220606100601002), and the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

References

[1] X. You *et al.*, "Towards 6G Wireless Communication Networks: Vision, Enabling Technologies, and New Paradigm Shifts," *Science China Information Sciences*, vol. 64, no. 1, Jan. 2021.
[2] C.-X. Wang *et al.*, "On the Road to 6g: Visions, Requirements, Key Technologies and Testbeds," *IEEE Commun. Surveys and Tutorials*, early access, 2023, doi: 10.1109/COMST.2023.3249835.
[3] R. Ali *et al.*, "A Federated Reinforcement Learning Framework for Incumbent Technologies in Beyond 5G Networks," *IEEE Network*, vol. 35, no. 4, 2021, pp. 152–59.
[4] K. Zhang, Z. Yang, and T. Bas¸ar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," *Handbook of Reinforcement Learning and Control*, 2021, pp. 321–84.
[5] S. Pateria *et al.*, "Hierarchical Reinforcement Learning: A Comprehensive Survey," *ACM Computing Surveys*, vol. 54, no. 5, 2021, pp. 1–35.
[6] M. Okada and T. Taniguchi, "Dreaming: Model-Based Reinforcement Learning by Latent Imagination Without Reconstruction," *Proc. 2021 IEEE Int'l. Conf. Robotics and Automation*, 2021, pp. 4209–15.
[7] H. Wu *et al.*, "Resource Management in Space-Air-Ground Integrated Vehicular Networks: SDN Control and AI Algorithm Design," *IEEE Wireless Commun.*, vol. 27, no. 6, 2020, pp. 52–60.
[8] W. Shi *et al.*, "Drone-Cell Trajectory Planning and Resource Allocation for Highly Mobile Networks: A Hierarchical DRL Approach," *IEEE Internet of Things J.*, vol. 8, no. 12, 2020, pp. 9800–13.
[9] K. K. Nguyen *et al.*, "3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning," *IEEE Trans. Commun.*, vol. 70, no. 4, 2022, pp. 2358–71.
[10] N. Su *et al.*, "Unmanned-Surface-Vehicle-Aided Maritime Data Collection Using Deep Reinforcement Learning," *IEEE Internet of Things J.*, vol. 9, no. 20, 2022, pp. 19,773–86.
[11] J. Li *et al.*, "Delay-Aware Vnf Scheduling: A Reinforcement Learning Approach With Variable Action Set," *IEEE Trans. Cognitive Commun. Networking*, vol. 7, no. 1, 2020, pp. 304–18.
[12] Q. Huang *et al.*, "Market Mechanisms for Cooperative Operation of Price-Maker Energy Storage in a Power Network," *IEEE Trans. Power Systems*, vol. 33, no. 3, 2017, pp. 3013–28.
[13] H. Wang and B. Zhang, "Energy Storage Arbitrage in Real-Time Markets via Reinforcement Learning," *Proc. 2018 IEEE Power & Energy Society General Meeting*, IEEE, 2018, pp. 1–5.
[14] J. Jin and Y. Xu, "Shortest Path Based Deep Reinforcement Learning for Ev Charging Routing Under Stochastic Traffic Condition and Electricity Prices," *IEEE Internet of Things J.*, 2022.
[15] Y. Ye *et al.*, "Deep Reinforcement Learning for Strategic Bidding in Electricity Markets," *IEEE Trans. Smart Grid*, vol. 11, no. 2, 2019, pp. 1343–55.

Despite the versatile applications of applying RL-based approaches in smart grid control and management, there exist several critical challenges.

Biographies

Junling Li [S'18, M'21] received her B.S. degree from Tianjin University, Tianjin, China, and her M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013 and 2016, respectively. In 2020, she received her Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada. She is currently an Associate Professor in the National Mobile Communications Research Laboratory at Southeast University, Nanjing, China. Her research interests include game theory, machine learning, software-defined networking, network function virtualization, and vehicular networks.

Huaqing Wu [S'15, M'21] received her Ph.D. degree from the University of Waterloo, Ontario, Canada, in 2021, and B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively. She worked as a postdoctoral fellow with the Department of Electrical and Computer Engineering, MacMaster University, from 2021 to 2022. She is currently an Assistant Professor with the Department of Electrical and Software Engineering, University of Calgary, Alberta, Canada. Her current research interests include B5G/6G, space-air-ground integrated networks, Internet of vehicles, mobile/edge computing/caching, and artificial intelligence (AI) for future networking.

Xi Huang [S'19, M'20] received his B.Eng. degree from Nanjing University, China, in 2014 and his Ph.D. degree from ShanghaiTech University, China, in 2021. Currently, he is a Research Scientist at the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), China, and a Distinguished Researcher of the Shenzhen Pengcheng Peacock project. His research is focused on multi-agent collaborations for edge intelligence and data marketplaces. He was a visiting researcher at the Department of Electrical Engineering and Computer Sciences, UC Berkeley, in 2017.

Qisheng Huang [S'19, M'20] received his B.S. degree in Electrical Engineering from Xi'an Jiaotong University, Xi'an, China, in 2014, and his Ph.D. degree from Singapore University of Technology and Design (SUTD), Singapore, in 2019. He was a postdoc research fellow at SUTD, Singapore, in 2019–2020, and a postdoctoral researcher at the Chinese University of Hong Kong, Shenzhen, Shenzhen, China, in 2020–2022. He is currently an Assistant Professor at Harbin Institute of Technology, Shenzhen. His research interests include power system economics, electricity markets, and carbon market.

Jianwei Huang [F'16] is a Presidential Chair Professor and Associate Vice President of the Chinese University of Hong Kong, Shenzhen. His research interests are in the area of network optimization, network economics, and network science, with applications in communication networks, energy networks, data markets, and crowd intelligence. He is the Editor-in-Chief of *IEEE Trans. Network Science and Engineering (TNSE)*, and was an Associate Editor-in-Chief of *IEEE Open J. the Communications Society*.

Xuemin (Sherman) Shen [M'97, SM'02, F'09] is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo. His research focuses on network resource management, wireless network security, social networks, 5G and beyond, and vehicular ad hoc networks. He is a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, an Engineering Institute of Canada Fellow, and a Chinese Academy of Engineering Foreign Fellow. He received President's Excellence in Research from the University of Waterloo in 2022, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R.A. Fessenden Award in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019.