Deep Reinforcement Learning for Economic Dispatch of Virtual Power Plant in Internet of Energy

Lin Lin, *Student Member, IEEE*, Xin Guan, *Member, IEEE*, Yu Peng, Ning Wang, Sabita Maharjan, *Senior Member, IEEE*, and Tomoaki Ohtsuki, *Senior Member, IEEE*

***Abstract*—With the high penetration of large-scale distributed renewable energy generation, the power system is facing enormous challenges in terms of the inherent uncertainty of power generation of renewable energy resources. In this regard, virtual power plants (VPPs) can play a crucial role in integrating a large number of distributed generation units (DGs) more effectively to improve the stability of the power systems. Due to the uncertainty and nonlinear characteristics of DGs, reliable economic dispatch in VPPs requires timely and reliable communication between DGs, and between the generation side and the load side. The online economic dispatch optimizes the cost of VPPs. In this article, we propose a deep reinforcement learning (DRL) algorithm for the optimal online economic dispatch strategy in VPPs. By utilizing DRL, our proposed algorithm reduced the computational complexity while also incorporating large and continuous state space due to the stochastic characteristics of distributed power generation. We further design an edge computing framework to handle the stochastic and large-state space characteristics of VPPs. The DRL-based real-time economic dispatch algorithm is executed online. We utilize real meteorological and load data to analyze and validate the performance of our proposed algorithm. The experimental results show that our proposed DRL-based algorithm can successfully learn the characteristics of DGs and industrial user demands. It can learn to choose actions to minimize the cost of VPPs. Compared with the deterministic policy gradient algorithm and DDPG, our proposed method has lower time complexity.**

***Index Terms*—Deep reinforcement learning (DRL), distributed generation, economic dispatch, edge computing, virtual power plants (VPPs).**

Manuscript received September 30, 2019; revised December 18, 2019; accepted December 27, 2019. Date of publication January 13, 2020; date of current version July 10, 2020. This work was supported by the Science and Technology Projects of State Grid Corporation of China under Grant SGHL0000DKJS1900883. *(Corresponding author: Xin Guan.)*

Lin Lin and Xin Guan are with the School of Data Science and Technology, Heilongjiang University, Harbin 150080, China (e-mail: ll.linlin@hotmail.com; guanxin.hlju@gmail.com).

Yu Peng and Ning Wang are with the Dispatching and Control Center, State Grid Heilongjiang Electric Power Company Ltd., Harbin 150080, China (e-mail: 142341@qq.com; wn007@126.com).

Sabita Maharjan is with the Simula Metropolitan Center for Digital Engineering, 0167 Oslo, Norway, and also with the University of Oslo, 0315 Oslo, Norway (e-mail: sabita@simula.no).

Tomoaki Ohtsuki is with the Department of Information and Computer Science, Keio University, Yokohama 223-8522, Japan (e-mail: ohtsuki@ics.keio.ac.jp).

Digital Object Identifier 10.1109/JIOT.2020.2966232

I. Introduction

DISTRIBUTED renewable energy resources will play a key role in accommodating the increasing power demand and in addressing environmental pollution [1]. Of particular interest in this context are large-scale distributed generation units (DGs). In order to realize coordinated control of DG, microgrid and virtual power plant (VPP) adopt reasonable economic dispatching methods for interconnected large-scale DGs. With "VPPs," various types of large-scale distributed power generation resources can be effectively integrated to supply power to industrial users in geographically dispersed areas. VPPs also play an increasingly important role in the operation of power market and ancillary service market [2].

In the previous studies on the optimal economic dispatch in VPPs, the demand response from end users was not fully incorporated into the objective function [3]. In this article, we introduce the compensation cost for controllable load from the user side while also explicitly including the inherent uncertainty associated with renewable power generation of the optimal economic dispatch problem formulation of VPPs. Stochastic optimization [4], such as particle swarm optimization (PSO), is used to solve the above-mentioned objective function. As we consider the real-time economic dispatching, the traditional heuristic method needs to rerun the optimization process for each decision and the computational complexity of PSO is high [5]. With the development of the VPP technology, the high-penetration rate of wind power, photovoltaic, and the active participation of industrial users make VPPs more uncertain. Traditional methods are difficult to solve the problem of uncertainty in more complex environments and fail to meet the future need of VPPs development, and it is difficult to find a solution for optimal solutions in large-scale data sets. Considering such limitations of the above solutions, in this article, we propose a deep reinforcement learning (DRL)-based algorithm to find the solution to the optimal economic dispatch problem. In a real VPP economic dispatch scenario, there are a large number of advanced metering infrastructures and wide-area monitoring systems on both the power generation side and the user side. They generate massive data, and these data provide a great opportunity for deep reinforcement training. Deep neural networks in DRL can process large amounts of data onto real scenarios.

The DRL-based asynchronous actor–critic algorithm [6] is an effective model-free method, and it can provide global long-term optimum rather than local short-term optimum. In addition, with the DRL-based algorithm, through the model trained offline, we can make a decision in milliseconds in the online dispatching phase. We make the computational complexity reasonably low despite the large-state space due to the stochastic characteristics of distributed renewable power generation.

There are many smart devices in the Internet of Energy (IoE) [7]. In the future development of IoE, the data generated by these smart devices are huge. The traditional centralized economic dispatch mechanism collects, processes, and transmits data through the central controller of the system to control and adjust the output power of each DG unit. For a centralized method, however, the system needs to collect a large amount of data for dispatch, which introduces further strain on communication infrastructure in terms of communication reliability and latency requirements. In addition, as large-scale solar and wind power access bring in a higher level of uncertainty into the VPPs, scalability, and adaptability become the critical issues in a centralized framework. On the other hand, for a fully decentralized dispatch system, optimal solutions can be obtained for each DG, but not for the whole VPP system.

Edge computing [8] offers considerable advantages in order to address the above challenges in terms of both communication performance and scalability. Due to the distributed nature of distributed power in VPPs, we set up edge nodes in cloud environment to handle the massive data generated by the distributed power and user-side smart devices of the VPP in different areas. As we consider the real-time economic dispatch scenario, that is, demand response and energy transfer are real time. Edge computing can accelerate the response time to users, and thus can contribute to making economic dispatch near real time. Storage and computation are usually moved near the terminal nodes of the network to reduce network overload and to process the collected information at edge to reduce latency. In addition, with the edge computing framework, privacy of the end users can be preserved. To this end, by considering the large size and various types of data generated by end users in VPPs, in this article, we introduce a three-layer architecture for economic dispatch in VPPs based on edge computing framework architecture. In the proposed framework, the first and second layers are edge computing layers, while the third layer is the cloud computing layer. At the first layer, the servers process and collect the information from both the generation side and the industrial side. At the second layer, in the online dispatching phase, the agent manages demand response and energy transfer in the local area. Compared to placing the dispatch of all areas in the cloud center, the proposed three-layer edge computing architecture reduces the computational complexity of processing the training task at the central node, and further reduces the communication load between the VPP operators and the DGs. We put the computing on edge nodes to enable applications on edge servers and use renewable energy to power the servers nearby, which can significantly reduce energy consumption.

Our contributions in this article are summarized as follows.

1) We formulate and optimize the economic dispatch problem for VPPs considering a more realistic scenario with a nonlinear and nonconvex objective function of VPPs and further improve the stability of IoE.
2) We propose a DRL-based algorithm to achieve the optimal solution of the formulated economic dispatch problem for VPPs with considerably low computational complexity despite the high-dimensional state space due to the randomness associated with distributed generation.
3) We design an edge-computing-based three-layer system architecture for VPPs for offloading the computation and communication load to the edge of the network in order to meet the near-real-time communication and computation requirements for economic dispatch in VPPs.

The remainder of this article is organized as follows. In Section II, we introduce related work. In Section III, we present the economic dispatch system architecture for VPPs based on edge computing. In Section IV, we formulate the economic dispatch mechanism as a nonlinear, nonconvex optimization problem. We propose a DRL-based algorithm that can be applied to different economic dispatching scenarios to minimize the cost of VPPs. In Section V, we provide numerical results to evaluate and validate the performance of our proposed framework and algorithm. Finally, Section VI concludes this article.

II. Related Works

In order to address and model the uncertainty associated with large-scale distributed renewable generation in VPPs, various methods have been proposed in the previous work. For instance, Liu combined interval optimization for optimal economic dispatch in VPPs in [9]. A multiobjective optimization model of VPPs dispatch was constructed through stochastic chance-constrained programming in [3]. A robust optimization method was proposed in [10] to optimize the dispatching scheme by adjusting the robustness coefficients. Liu proposed an optimal dispatch model to minimize the cost of VPPs, which takes into account the distributed renewable generation constraints, the supply demand balancing constraint, and also the security constraints of VPPs including network constraints [11]. Sousa introduced a multiobjective optimization model with three objective functions, maximizing operating revenue, and minimizing operating risk [3]. Chen presented a fully distributed method for economic dispatch of VPPs, using the alternating direction multiplier method and the consensus optimization algorithm [12]. In [13], a new consensus-based distributed control algorithm was proposed for economic dispatch when distributed generation was involved, through iterative coordination of local agents.

However, the above-mentioned methods cannot obtain a strategy suitable for large-scale data sets. In recent years, under the background of the IoE technology, the advanced machines learning methods [14]–[16], especially methods based on DRL. DRL integrates deep learning and reinforcement learning, which is emerging and has been widely adopted for

solving related problems in the IoE domain. For example, Sun *et al.* [17] mainly studied energy management in the IoE, and reinforcement learning methods were implemented to formulate the best operating strategies. Du *et al.* [18], studied the architecture design of the IoE in the context of large-scale renewable energy grid connection for the effective prediction and optimal use of energy. It takes the electric vehicle charging as a typical case to achieve optimal energy delivery based on reinforcement learning. For instance, DRL was used for designing emergency control strategy in [19]. Lu presented a demand response algorithm for the IoE system based on real-time execution. The algorithm combines reinforcement learning and deep neural network to predict unknown price and energy demand, balance energy fluctuation, overcome future uncertainty, and enhance grid reliability [20]. However, the utilization of the DRL-based algorithm to solve the economic dispatch problem of VPPs has not been explored. The objective function established in this article is a nonlinear cost function. Although this article does not add nonconvex characteristics, in real scenarios, power generation units are usually affected by the valve point effects, and the cost function is usually nonconvex. In order to solve these difficulties, the previous works mostly adopted heuristic methods. In [21], a distributed pattern search algorithm for nonconvex economic dispatching was proposed. Our DRL-based economic dispatch algorithm for VPPs can be adapted to this kind of nonlinear and nonconvex situations and loose the constraints of nonlinear and nonconvex characteristics, make it applicable to more realistic application scenarios regarding optimal dispatch in VPPs. The proposed algorithm contributes to further improve the stability of IoE. Besides, our DRL-based economic dispatch algorithm for VPPs to incorporate the stochastic characteristics reducing the high computational complexity.

The VPP architecture aggregates all of the DGs by utilizing advanced wireless communication technologies. Edge computing is the effective technology for economic dispatch in VPPs. Some previous works have explored the application of edge computing for the IoE. Liu *et al.* [22] proposed an energy management system based on edge computing infrastructure. Li *et al.* [23] proposed a unified energy management framework for enabling a sustainable edge computing paradigm with distributed renewable energy resources. These studies, however, do not focus on optimal economic dispatch issue of VPPs. Our proposed three-layer architecture based on edge computing is the first attempt for optimal economic dispatch in VPPs.

III. Edge-Computing-Based Architecture for the System

With the access of large-scale distributed generation in IoE, due to geographical constraints, the traditional microgrid has certain limitations that hinder effective utilization of large-scale distributed generation in multiregion, and the power curtailment happens. Due to the mismatch between the construction scale of renewable energy stations and the demand of local load, the accommodation capacity of renewable energy is limited, resulting in a certain number of power curtailments in the concentrated areas of wind power and photovoltaic power stations. Compared with the microgrid, VPP is a special type of power plant composed of different types of distributed power sources. It can coordinate the contradiction between the smart grid and distributed energy sources. It is also one of the important technologies to realize the intelligent distribution network. In order to achieve the coordinated control of distributed power generation and reduce power curtailment, VPPs effectively integrate various types of large-scale distributed power generation resources to adopt reasonable economic dispatch methods to supply power to industrial users in geographically dispersed areas.

Due to the complexity of economic dispatching scenarios, for instance, that includes distributed renewable power generation and industrial consumers, different types and a large amount of data need to be transmitted in real time. Due to the close relationship between industrial users and VPP operators, reasonable economic dispatch should take full account of user participation. Industrial users can participate in economic dispatch by signing contracts with VPP operators. The VPP operators need to receive data from the demand side industrial users and the DGs. Since data transmission between VPP operators and equipment require a certain level of performance to achieve optimal economic dispatch, the VPPs to adopt advanced control, sensing, and communication technology, to sense and collect data, and to transmit them to the economic dispatch control center of the VPPs to achieve optimal economic dispatch in the complex scenarios. Considering the wireless link between most devices and VPP operators, a large amount of data transmission can easily exceed the transmission capacity limit. Therefore, the batch equipment with limited resources cannot directly send the demand to the VPP operators, which makes a major challenge to effective economic dispatch.

Traditionally, VPP operators dispatch the geographically dispersed distributed power supply in a centralized manner. The information of users and real-time status data of DGs from multiple regions are sent to the cloud for storage and processing, which leads to a large amount of network communication load and computational resources consumption. This, however, leads to higher network latency. In a real scenario, the long distance data transmission from various DGs and industrial consumers to the cloud computing centers consumes considerable energy. Additionally, transmitted data raise privacy concerns for industrial users in different regions. Traditional cloud computing mode needs to upload local sensitive data to cloud computing center, which increases the risk of leakage of user privacy. In addition, the generation and transmission of a large amount of data make it difficult to accurately guarantee the reliability of data transmission in complex environments.

To address these challenges, edge computing is used to provide computing services on batch devices near the network edge of VPPs. Edge computing can greatly reduce data transmission from devices to VPP operators through preprocessing. Second, edge computing architecture can offload computational burden to the edge. Fig. 1 shows our proposed architecture for economic dispatch, which consists of four main components: 1) power-supply side server (PSS); 2) industrial

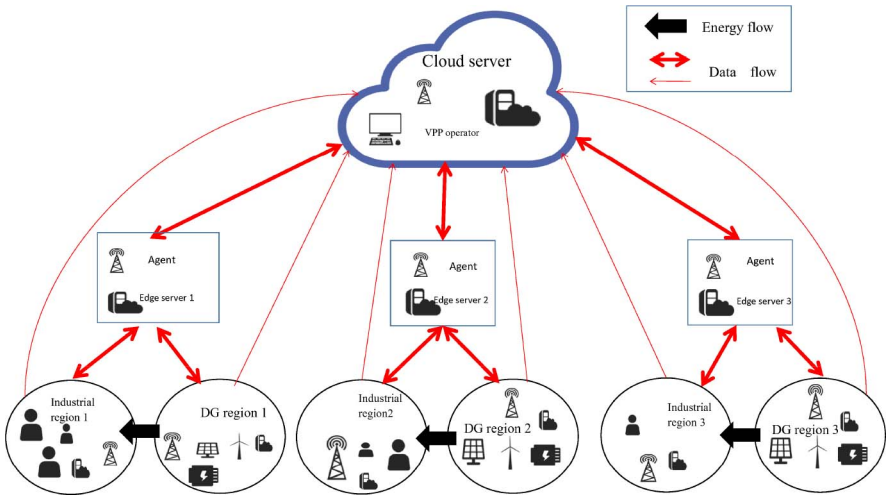


Fig. 1. Economic dispatch system architecture of VPP based on edge computing.

user side server; 3) agent edge server; and 4) VPP operator cloud server. The PSS connects the power equipment through different communication technologies (5G, WiFi, etc.). It collects and processes power generation data from the distributed power equipment and transmits data to agent edge server in real time. The PSS also receives the dispatch information of agent edge server and supplies power to industrial users. Industrial user side server connects the power equipment through different communication technologies (5G, WiFi, etc.). It collects and processes power consumption information of industrial users and transmits data to agent edge server in real time. According to the analysis results of industrial user side server and PSS, local economic dispatch decisions are made, and agent edge server interacts with the servers on both sides. VPP operator cloud server satisfies the computing requirements of agent edge server and manages each agent. It not only helps the agent server to provide real-time analysis and computation but also collects the dispatch information of the managed agent.

The proposed architecture consists of three layers, which are suitable for offline training and real-time online dispatch. First, in the offline training phase, it is necessary for industrial side servers and power servers to process and collect the information from generation side and user side in a specific area and transmit the collected information to the VPP operator cloud server. VPP operator cloud servers are trained based on the large-scale offline data, and the trained model is transferred to the agent edge side servers in specific areas. Due to the large number of industrial users and distributed power supply and considering the real-time demand response, the servers on both sides transfer the collected data to the agent edge server. The agent edge server will put it into the model trained before, to obtain a real-time economic dispatch strategy. Fig. 2 shows our proposed three-layer economic dispatch mode of VPP based on edge computing. First, VPP operators setup agents to manage distributed power generation and industrial users in different regions. On the demand side, the controllable load of user participates in demand response, which can reduce load demand during the peak period [24]. Compared with

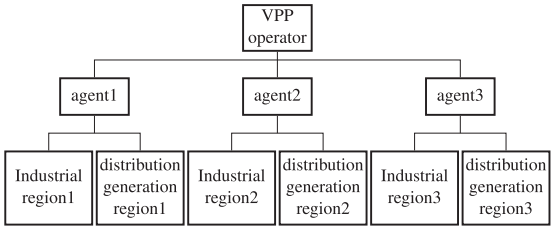


Fig. 2. Three-layer economic dispatch mode of VPP based on edge computing.

the VPP operators, each agent is an edge computing server. The industrial user side server and PSS collects the data from each distributed power generation unit and extracts and summarizes the data in real-time mode. These distributed power generations could be photovoltaic, wind power, and micro gas turbine. The agent server devices the optimal economic dispatch strategy for the region and, finally, transmits the decision information to the VPP operators. The three-layer economic dispatch model solves a large-scale and complex problem on the edge of VPP. To adapt to the distributed characteristics of power supply, the edge computing method is more flexible and suitable for the expansion of the dynamic network, thus making it a more scalable solution.

IV. Problem Formulation and Deep Reinforcement Learning Method

A. Problem Formulation for Economic Dispatch of VPP

The goal of economic dispatch for VPP operators is to minimize the operating cost for managing industrial users and the DGs, including photovoltaic, wind turbines, and micro gas turbines. Based on the minimizing cost of VPP operators, the proposed optimal economic dispatch algorithm fully takes into account the operation and maintenance cost of DG units defined as C^{pom}, C^{wom}, and C^{dom}. Specifically, we also consider the environmental protection cost C^{de} and fuel cost C^{d} of micro gas turbine. Overall, the initial depreciation cost of DG units are taken into account and, respectively, defined as C^{pdp}, C^{wdp}, and C^{ddp}. The variable notations of this paper are

TABLE I
VARIABLE NOTATIONS

Symbol	Definition
C	VPP operation cost for managing DGs and industrial users
C_i	Operation cost for managing DGs and industrial users in region i
C_i^g	Operation cost of DGs in region i
C_i^{dr}	Compensation cost to the load side in region i
$p_i^{p'}(k)$	Photovoltaic actual consumption in time slot k
$p_i^{w'}(k)$	Wind turbine actual consumption in time slot k
$p_i^{d'}(k)$	Micro gas turbine actual consumption in time slot k
$C_i^{pdp}\left(p_i^p(k)\right)$	Initial investment depreciation cost of photovoltaic in time slot k
r	Annual interest rate
$C^{p_{n^2}}$	Unit capacity installation cost of photovoltaic cells
K_p	Photovoltaic capacity factor
n^p	Operation lifetime of photovoltaic
$C_i^{pom}\left(p_i^p(k)\right)$	Maintenance operation cost for photovoltaic in time slot k
K_{pom}	Maintenance operation cost coefficient for photovoltaic
$C_i^{wdp}\left(p_i^w(k)\right)$	Initial depreciation cost of wind turbines in time slot k
$C^{w_{a^2}}$	Installation cost per unit capacity of wind turbines
K_w	Capacity factor of wind turbines
n^w	Operation life of wind turbines
$C_i^{de}\left(p_i^d(k)\right)$	Environmental protection cost of micro-gas turbines in time slot k
M	Total number of pollutant types
β_m	Treatment cost of the m unit emission of pollutants
α_{dm}	Emission amount of pollutants when the micro gas turbine produces unit electric energy
η_d	Generation efficiency of micro gas turbine
$p_i^d(k)$	Output power of micro gas turbine in time slot k
$C_i^d\left(p_i^d(k)\right)$	Fuel cost in time slot k
c^d	Natural gas price
L	Lowest energy released by natural gas
$p_i^{con}(k)$	Controllable load in time slot k
$p_i^{base}(k)$	Uncontrollable load in time slot k
λ	Compensation coefficient for contract signing
$x_i(k)$	Interruptible load percentage in time slot k
$p_i^{v'}(k)$	Actual power generated by photovoltaic in time slot k
$p_i^{w'}(k)$	Actual power generated by wind turbine in time slot k
$p_i^{p'}(k)$	Actual power generated by micro-gas turbine in time slot k
n^w	Operation lifetime of wind turbines

listed in Table I. We consider the requirements of industrial users and also include the compensation cost to industrial users who participate in demand response, represented as C^{dr}. The proposed algorithm reduces the economic loss of VPPs during the peak periods of power consumption by reducing the controllable load. Increasing the flexibility of users may lead to load peak-valley offset and reduce the uncertainty of renewable energy generation. Meanwhile, we treat the industrial users as a dispatchable resource that participates in the system operation in VPPs. In such case, the industrial users are equivalent to a virtual generation resource [25]. Therefore, in the objective function of the proposed model, the compensation cost to the demand side is C^{dr}. The cost compensates for users who choose to cut the controllable load. The objective function consists of two parts, the first part is the operation cost of DGs, and the second part is the compensation cost for controllable load when the demand side participates in the operation of the system [26]

$$\begin{cases} \min C \\ C = \sum_i^I C_i \\ C_i = C_i^g + C_i^{\mathrm{dr}} \end{cases} \tag{1}$$

where C is the VPP operation cost for managing DG units and industrial users in VPPs. The number of regions is denoted as I. C_i is the operation cost for managing DG units and industrial users in region i. C_i^g is the operation cost of DGs in region i and C_i^{dr} is the compensation cost to the demand side of industrial users in region i.

In real-time scenarios, the edge of VPP operators is denoted as agent i. In our proposed optimal economic dispatch model, three types of DGs are considered, which are photovoltaic, wind power, and micro gas turbines. The temporary power data from photovoltaic, wind power, and micro gas turbines are extracted from the data set in [27]. The operation cost of DG units includes the initial depreciation, operation, and maintenance cost of VPPs. Specifically, the environmental protection and fuel cost of micro gas turbines are also considered. k denotes the time slot intervals, $p_i^{p'}(k), p_i^{w'}(k), p_i^{d'}(k)$ represent the actual consumption of photovoltaic, wind turbine, and micro gas turbine, respectively, in time slot k.

1) *Photovoltaic:* The initial depreciation cost of photovoltaic investment can be expressed as [28]

$$C_i^{\mathrm{pdp}}\left(p_i^p(k)\right) = \frac{C^{p_{n^2}}}{8760K_p} \times \frac{r(1+r)^{n^p}}{(1+r)^{n^p}-1} \times p_i^p(k) \tag{2}$$

where r is the annual interest rate, $C^{p_{n^2}}$ is the unit capacity installation cost of photovoltaic cells, K_p is the photovoltaic capacity factor, and n^p is the operation lifetime of photovoltaic. The operational and maintenance costs of photovoltaic [28]

$$C_i^{\mathrm{pom}}\left(p_i^p(k)\right) = K_{\mathrm{pom}} \times p_i^p(k) \tag{3}$$

where $C_i^{\mathrm{pom}}(p_i^p(k))$ is the maintenance operation cost for photovoltaic and K_{pom} is the maintenance operation cost coefficient for photovoltaic.

2) *Wind Turbines:* The initial investment cost of wind turbines is converted to the output power per unit time. As the depreciation cost of wind turbines, it is included in the operation cost of wind turbines [28]

$$C_i^{\mathrm{wdp}}\left(p_i^w(k)\right) = \frac{C^{w_{a^2}}}{8760K_w} \times \frac{r(1+r)^{n^w}}{(1+r)^{n^w}-1} \times p_i^w(k) \tag{4}$$

where $C_i^{\mathrm{wdp}}(p_i^w(k))$ is the initial depreciation cost of wind turbines, $C^{w_{a^2}}$ is the installation cost per unit capacity of wind turbines, K_w is the capacity factor of wind turbines, r is the annual interest rate, and n^w is the operation lifetime of wind turbines. The operation and maintenance costs of wind turbines during operation can be expressed as [28]

$$C_i^{\mathrm{wom}}\left(p_i^w(k)\right) = K_{\mathrm{wom}} \times p_i^w(k) \tag{5}$$

where K_{wom} is the operation cost coefficient of wind turbines.

3) *Micro Gas Turbines:* The initial depreciation cost of micro gas turbine is modeled as [29]

$$C_i^{\mathrm{ddp}}\left(p_i^d(k)\right) = \frac{C^{d_{a2}}}{8760K_d} \times \frac{r(1+r)^{n^d}}{(1+r)^{n^d}-1} \times p_i^d(k) \quad (6)$$

where $C^{d_{a2}}$ is the installation cost per unit capacity of micro gas turbine, K_d is the capacity factor of micro gas turbine, and n^d is the operation life of micro gas turbine. The operation and maintenance costs of micro gas turbines can be expressed as [29]

$$C_i^{\mathrm{dom}}\left(p_i^d(k)\right) = K_{\mathrm{dom}} \times p_i^d(k) \quad (7)$$

where K_{dom} is the operation and maintenance cost coefficient of micro gas turbines.

The environmental protection cost of micro gas turbine is [29]

$$C_i^{\mathrm{de}}\left(p_i^d(k)\right) = \sum_{m=1}^{M} 10^{-3}\beta_m \alpha_{dm} p_i^d(k) \quad (8)$$

where m is the pollutant discharged, M is the total number of pollutant types, β_m is the treatment cost of the m unit emission of pollutants, and α_{dm} is the emission amount of pollutants when the micro gas turbine produces unit electric energy.

The relationship function between generation efficiency and output power of micro gas turbine is [29]

$$\eta_d = 0.0753\left(\frac{p_i^d(k)}{65}\right)^3 - 0.3095\left(\frac{p_i^d(k)}{65}\right)^2 + 0.4174\left(\frac{p_i^d(k)}{65}\right) + 0.1068 \quad (9)$$

where η_d is the generation efficiency of micro gas turbine and $p_i^d(k)$ is the output power of micro gas turbine.

The consumption characteristics of micro gas turbines can be expressed as follows [29]:

$$C_i^d\left(p_i^d(k)\right) = \frac{c^d}{L\eta_d} p_i^d(k) \quad (10)$$

where $C_i^d(p_i^d(k))$ is the fuel cost, c^d is the natural gas price, and L is the lowest energy released by natural gas.

Based on the above description, we have the operating cost for DGs as follows:

$$C_i^g = \sum_k^K \Big[C_i^{\mathrm{pdp}}\left(p_i^p(k)\right) + C_i^{\mathrm{pom}}\left(p_i^p(k)\right) + C_i^{\mathrm{wdp}}\left(p_i^w(k)\right) + C_i^{\mathrm{wom}}\left(p_i^w(k)\right) + C_i^{\mathrm{ddp}}\left(p_i^d(k)\right) + C_i^{\mathrm{dom}}\left(p_i^d(k)\right) + C_i^{\mathrm{de}}\left(p_i^d(k)\right) + C_i^d\left(p_i^d(k)\right) \Big]. \quad (11)$$

Demand response can effectively integrate the potential of user side response to enhance the security, stability, and economy of power grid operation [30], [31]. In this article, we consider the demand response of industrial users during the process of model construction. In order to achieve the optimal economic dispatch strategy, each agent chooses the size of controllable load for cutting. Since the controllable load is reduced, it leads to inconvenience for industrial users, therefore, they need to be compensated. VPP operator should provide power compensation to users who choose to cut the controllable load. The variable of controllable load models $X_i(k)$ requires the compensation coefficient λ. $X_i(k)$ is a variable obtained from the electricity information of all industrial users within a region, which is defined as the percentage of the maximum interruptible controllable load in each time slot of the industrial area, considering that agent i's compensation cost on the load side is C_i^{dr}. This method can reduce or move part of the power consumption to avoid peak load for industrial users. The mentioned control method is beneficial to the economic cost of VPP operators and industrial users. The industrial user's load is obtained from [32] and is divided into controllable load $p_i^{\mathrm{con}}(k)$ and uncontrollable load $p_i^{\mathrm{base}}(k)$. Since controllable load can directly respond to the economic dispatch in VPPs, this article mainly focuses on the reduction of controllable load when participating in the dispatch process in VPPs.

The compensation cost on the load side of agent i can be expressed as

$$C_i^{\mathrm{dr}} = \sum_k^K \lambda p_i^{\mathrm{con}}(k) x_i(k) \quad (12)$$

where λ is the compensation coefficient, and $x_i(k)$ is expressed as vectors for selecting interruptible load percentage, the range of value is [0, 1].

The objective function of economic dispatch for each agent i can be expressed as

$$\min C_i = \sum_k^K \Big[C_i^{\mathrm{pdp}}\left(p_i^p(k)\right) + C_i^{\mathrm{pom}}\left(p_i^p(k)\right) + C_i^{\mathrm{wdp}}\left(p_i^w(k)\right) + C_i^{\mathrm{wom}}\left(p_i^w(k)\right) + C_i^{\mathrm{ddp}}\left(p_i^d(k)\right) + C_i^{\mathrm{dom}}\left(p_i^d(k)\right) + C_i^{\mathrm{de}}\left(p_i^d(k)\right) + C_i^d\left(p_i^d(k)\right) + \lambda p_i^{\mathrm{con}}(k) x_i(k) \Big]. \quad (13)$$

For the entire VPP system, power balance constraints are the fundamental issue and should be fully considered during the process of model construction. In each management area of agent i, the total power consumption by various DG units should be equal to the total power demand of industrial users. The total power demand of industrial users takes the reduction of controllable load of agent i to industrial users into account, as shown in the following equation:

$$p_i^d(k) + p_i^w(k) + p_i^p(k) = p_i^{\mathrm{base}}(k) + p_i^{\mathrm{con}}(k)(1 - x_i(k)). \quad (14)$$

The actual power consumption by DG units in each agent management region is limited by the actual power generation in this area. The actual power generated by the DG units are $p_i^{v'}(k)$, $p_i^{w'}(k)$, and $p_i^{p'}(k)$, which are photovoltaic, wind power, and micro gas turbine, respectively, as follows:

$$0 \le p_i^d(k) \le p_i^{d'}(k) \quad (15)$$

$$0 \le p_i^w(k) \le p_i^{w'}(k) \quad (16)$$

$$0 \le p_i^p(k) \le p_i^{p'}(k). \quad (17)$$

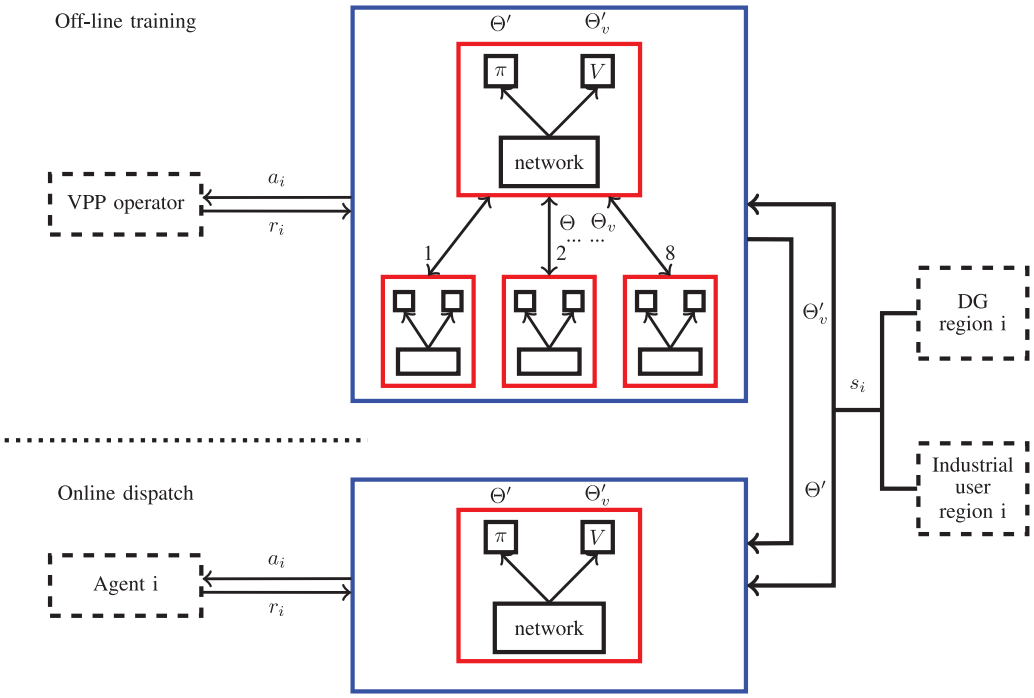


Fig. 3. Information acquisition and DRL mechanism for VPP economic dispatch.

The percentage of interruptible load in the industrial area managed by agent i should not exceed the percentage of maximum interruptible controllable load in each time slot, i.e.,

$$0 \le x_i(k) \le X_i(k). \tag{18}$$

VPP operators manage all regions and aggregate the dispatch information of each region. Based on the above description, we define the objective function for the optimal economic dispatch strategy as follows:

$$\begin{aligned} \min \ C = & \sum_{i}^{I} \sum_{k}^{K} \\ & \times \Big[C_i^{\mathrm{pdp}}\left(p_i^p(k)\right) + C_i^{\mathrm{pom}}\left(p_i^p(k)\right) \\ & + C_i^{\mathrm{wdp}}\left(p_i^w(k)\right) + C_i^{\mathrm{wom}}\left(p_i^w(k)\right) + C_i^{\mathrm{ddp}}\left(p_i^d(k)\right) \\ & + C_i^{\mathrm{dom}}\left(p_i^d(k)\right) + C_i^{\mathrm{de}}\left(p_i^d(k)\right) + C_i^{d}\left(p_i^d(k)\right) \\ & + \lambda p_i^{\mathrm{con}}(k) x_i(k) \Big] \\ \text{s.t.} \ & (14) - (18). \end{aligned} \tag{19}$$

In this article, our proposed optimal economic dispatch strategy minimizes the generation cost of DGs while also satisfying the restrictions on power balance and generation capacity for VPP.

B. Deep Reinforcement Learning Algorithm

In order to make the scenario more realistic, we have included various cost components into the objective function. In the real economic dispatch scenario, the dispatch process should normally be completed within a short time period. Since the stochastic characteristics due to photovoltaic and wind power generation and also from the flexibility in load, the state of the systems from previous time slot to the next time slot constitutes a larger state space, and the state information needs to be updated quickly.

DRL, as an effective artificial intelligence algorithm, has achieved great success in solving many problems of the areas such as the Internet of Things [33]–[35], and can find the optimization strategies in different scenarios, within a reasonable time frame. We adopt such advantages of DRL to the complex VPPs-economic dispatch scenario in this article. When we apply the offline trained model to online economic dispatch, we can continuously optimize the model. The proposed DRL-based algorithm looses the constraints of nonlinear and nonconvex characteristics. Problem (19) constitutes a high-dimensional and continuous state space. The proposed algorithm fits the value function to precisely control the continuous state space by using deep learning algorithm and, furthermore, improves the accuracy of solution. In this article, the economic dispatch problem is nonlinear with unknown transfer probability and large continuous state space. The efficiency of the computation is prominently enhanced by our proposed DRL-based optimal economic dispatch algorithm. The DRL requires no environment information to compute the transfer probability distribution. Thus, the optimal economic dispatch algorithm is run in the real-time mode.

The information acquisition and the DRL mechanism for VPP economic dispatch are shown in Fig. 3. The proposed algorithm adopts the offline data training mode, the PSS, and the user side server to collect the historical temporary data and transfers the information to the VPP cloud server. The VPP cloud server uses DRL to train the network separately according to the data transferred from different regions, and the economic dispatch strategies of different regions are obtained.

In the online economic dispatch phase, each agent edge server obtains the corresponding network weight value from the VPP cloud server. The PSS and the industrial user side server aggregate the real-time transmission information and power demand and then transmit all the aggregated information to the corresponding agent edge server. If there are slight changes to the online environment, our trained model can learn these changes by default and dynamically adjust the actions to achieve optimal dispatching. During the online dispatching phase, distributed power generation data and industrial user demand data can be directly transferred to edge computing nodes without being transferred to the cloud center, which is more suitable for real-time economic dispatch scenarios.

For the offline training of VPP operators, we consider a 24-h time frame denoted by $k \in (0, 1, \ldots, 23)$. The goal of economic dispatch is to find an optimal economic dispatch solution to minimize the operation cost of VPPs. For region i, the state set S_i, for $s_i \in S_i$, $s_i = \{p_i^{p'}(k), p_i^{w'}(k), p_i^{d'}(k), p_i^{\text{con}}(k), p_i^{\text{base}}(k)\}$ is aggregated by PSS and the industrial user side server, which, respectively, represents the actual power generation in slot k, photovoltaic, wind power, micro gas turbine, industrial user controlled load, and uncontrollable load demand. The action set A_i, $a_i \in A_i$, $a_i = \{p_i^p(k), p_i^w(k), p_i^d(k), x_i(k)\}$, represents the actual power consumption of photovoltaic, wind power, and micro gas turbine, respectively, and the control coefficient of the controllable load in slot k. A is the continuous action space that satisfies the power and capacity balance constraints, a_i is the selected action which satisfies action constraints. In any time slot, in order to find the mapping relationship from state to action, we introduce strategy π. The strategy represents the conditional probability distribution of each action under the condition that the current state is known. The next state is denoted as s'_i and the initial state is denoted as $s^0{}_i$. i.e., $\pi = P(s^0{}_i)\prod_{k=1}^{K} P(a_i|s_i)P(s'_i|s_i, a_i)$. In a real scenario, the state transition probability is unknown, and the state space and the behavior space are continuous. When s_i, a_i are known that the reward $r_i(s_i, a_i)$ related to the objective function can be obtained

$$\begin{aligned} r_i = &-K_1\left(C_i^{\text{pdp}}\left(P_i^p\right) + C_i^{\text{pom}}\left(P_i^p\right)\right) \\ &- K_2\left(C_i^{\text{wdp}}\left(P_i^w\right) + C_i^{\text{wom}}\left(P_i^w\right)\right) \\ &- K_3\left(C_i^{\text{ddp}}\left(P_i^d\right) + C_i^{\text{dom}}\left(P_i^d\right) + C_i^{\text{de}}\left(P_i^d\right) + C_i^d\left(P_i^d\right)\right) \\ &- K_4\lambda P_i^{com} x_i \end{aligned} \tag{20}$$

where $K_1, K_2, K_3,$ and K_4 are the weight values we set. Because we want to minimize the cost of the VPP, the reward value is negative.

We can get a total reward for K hours as

$$\begin{aligned} \bar{r_i}(s_i, a_i) &= \sum_k^K r_i(s_i, a_i)\pi \\ &= \underset{\pi}{E}\left[\sum_k^K r_i(s_i, a_i)\right]. \end{aligned} \tag{21}$$

To maximize the reward, we utilize the gradient ascend method to update the strategy in our proposed algorithm, i.e.,

$$\begin{aligned} \nabla\bar{r}_i(s_i, a_i) &= \sum_k^K r_i(s_i, a_i)\nabla\pi \\ &= \sum_k^K r_i(s_i, a_i)\pi\frac{\nabla\pi}{\pi} \\ &= \underset{k\sim\pi}{E}\left[r_i(s_i, a_i)\nabla\log\pi\right]. \end{aligned} \tag{22}$$

We can obtain the state value function $V^\pi(s_i)$ and state action value function $Q^\pi(s_i, a_i)$, γ is the discount factor, representing the discount rate of reward over time

$$V^\pi(s_i) = E\left(\sum_{k=0}^{K}\gamma^k r_{i,k}|s_i\right) \tag{23}$$

$$\begin{aligned} Q^\pi(s_i, a_i) &= E\left(\sum_{k=0}^{K}\gamma^k r_{i,k}|s_i, a_i\right) \\ &= E\left(r_i + \gamma V^\pi\left(s'_i\right)|s_i, a_i\right). \end{aligned} \tag{24}$$

The goal is to select the optimal strategy and maximize the state-action value function, which is the optimal state-action value function

$$Q^{\pi^*}(s_i, a_i) = \max_\pi E\left(\sum_k^K \gamma^k r_i(s_i, a_i)\right). \tag{25}$$

In order to find the optimal economic dispatch strategy, we usually consider utilizing the data table representation function. However, this method limits the scale of the reinforcement learning algorithm. When the scale of the problem is too large and the memory space for storing a table is huge, it takes a long time to accurately compute each value in the table. In case the learning experience is obtained from the small-scale data sets of training, the generalization ability of the training mode is insufficient. In order to solve the above problems, considering the large-scale state action space, the deep neural network is used to parameterize the state value function and state-action value function. In our proposed algorithm, we use deep neural networks to extract the features of large-scale input state data for training the economic dispatch model to make the trained model more adaptable. From the first layer of neurons, through the nonlinear activation function into the next layer of neurons, continue to downstream, so that the cycle until the output layer. Since the nonlinear functions are indispensable for deep neural networks, the deep neural network has sufficient capacity to extract data characteristics. θ_v is utilized to approximate the state value function $V(s_i)$ and the state-action value function $Q(s_i, a_i)$

$$Q(s_i, a_i) \approx Q(s_i, a_i, \theta_v) \tag{26}$$

$$V(s_i) \approx V(s_i, \theta_v). \tag{27}$$

Deep neural networks are based on the function approximator and its parameter θ is the strategy parameter. π obeys the Gaussian distribution, i.e.,

$$\pi(a_i|s_i, \theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\frac{-(a_i - s_i^T\theta)^2}{2\sigma^2}. \tag{28}$$

The reward value for each region i is shown in (22)

$$R_i = E\left(\sum_{k}^{K} \gamma^k r_i(s_i, a_i)\right). \tag{29}$$

In our scenario, for increasing the probability of the strategy with high reward, the policy gradient is employed to complete the goal. The calculation of gradient is updated as [6]

$$\nabla \log \pi(a_i|s_i, \theta)(R_i - b(s_i)) \tag{30}$$

where R_i is the total reward in the region i and is estimated by $Q(s_i, a_i)$, i.e., $R_i \approx Q(s_i, a_i)$. $b(s_i)$ is a baseline that is utilized to reduce the error of estimation. $V(s_i)$ is employed to estimate the baseline, i.e., $b(s_i) \approx V(s_i)$

$$A^{\pi}(s_i, a_i; \theta, \theta_v) = Q^{\pi}(s_i, a_i, \theta_v) - V^{\pi}(s_i, \theta_v). \tag{31}$$

We use (31) to substitute $(R_i - b(s_i))$. Equation (31) is the advantage function that means the advantage of the action value function over the value function. In case the action value function is larger than the value function, the advantage function is positive, and if the action value function is smaller, the advantage function is negative. When the advantage function is positive, the parameters are updated along the direction of increasing the probability of the strategy, and when the advantage function is negative, the parameters are updated along the direction of decreasing the probability of the strategy. Therefore, when the advantage function is adopted, the convergence of the algorithm is faster. The advantage function is presented by (31). Therefore, the calculation of gradient is transformed into

$$\nabla \log \pi(a_i|s_i, \theta)\big(A^{\pi}(s_i, a_i; \theta, \theta_v)\big). \tag{32}$$

Then, policy gradient is updated as

$$\partial (A^{\pi}(s_i, a_i; \theta, \theta_v))^2 / \partial \theta_v. \tag{33}$$

Therefore, the parameters of θ and θ_v are updated as

$$d\theta \leftarrow d\theta + \nabla \log \pi(a_i|s_i, \theta) A^{\pi}(s_i, a_i; \theta, \theta_v) \tag{34}$$

$$d\theta_v \leftarrow d\theta_v + \partial \big(A^{\pi}(s_i, a_i; \theta, \theta_v)\big)^2 / \partial \theta_v. \tag{35}$$

To make the training strategy more adaptable and to prevent the premature convergence of the suboptimal deterministic policy, we include the entropy regularization term in the policy gradient, i.e.,

$$H(\pi(s_i; \theta)) = -\sum_{a_i} \pi(s_i, a_i) \log \pi(s_i, a_i, \theta) \tag{36}$$

$$d\theta \leftarrow d\theta + \nabla \log \pi(a_i|s_i, \theta) A^{\pi}(s_i, a_i; \theta, \theta_v) + \beta \nabla H(\pi(s_i; \theta)) \tag{37}$$

where β is the hyperparameter, which controls the strength of the entropy. The first term updates the value function in a larger direction, and the second term updates the regular penalty parameter, which works in the convergence of the value function.

When training a neural network, the required data is independent and identically distributed. In order to break the correlation between different data sets, we utilize the asynchronous mechanism for updating the value function. The proposed algorithm utilizes the multiple threads method instead of a single thread. During the training process, multiple threads maintain a global actor–critic network. Each thread keeps a local network-weight value copy of the global network. The local network cumulated gradients, and the gradients are sent to the global network for parameter updating. After that, the local network synchronizes the parameters from the global network. The local network not only updates its own independent network by learning the state of environment but also interacts with the global actor–critic network. We define global shared parameter vectors θ'_v and θ' as

$$d\theta'_v \leftarrow d\theta'_v + \partial \big(A^{\pi}(s_i, a_i; \theta, \theta_v)\big)^2 / \partial \theta_v \tag{38}$$

$$d\theta' \leftarrow d\theta' + \nabla \log \pi(a_i|s_i, \theta) A^{\pi}(s_i, a_i; \theta, \theta_v) + \beta \nabla H(\pi(s_i; \theta)). \tag{39}$$

In this sense, each region reaches the optimal economic dispatch strategy. The offline training procedure is summarized in Algorithm 1. In the numerical section, we implement eight threads, the VPP operators communicate with each region and calculate C. Based on the algorithm, the economic dispatch model for region i can be obtained. In the online dispatch stage, each agent edge server first obtains the corresponding network weight value from the VPP cloud server, i.e., the agent i. The economic dispatch model based on DRL is shown in Fig. 3.

Algorithm 1 Asynchronous Advantage Actor-Critic for VPP in Region i

Initialize θ', θ, θ_v, θ'_v, K'
while $k < K'$ **do**
  Reset gradients:$d\theta' \leftarrow 0$ and $d\theta'_v \leftarrow 0$
  Set thread parameters to global θ=θ', θ'_v=θ_v
  $k_{start} = k$
  Observe state $s_i(k)$
  repeat
    Perform $a_i(k)$ according to policy $\pi(a_i(k)|s_i(k); \theta)$
    Receive reward $r_i(k)$ and new state $s_i(k+1)$
    $k \leftarrow k + 1$
  until $k - k_{start} == 23$
  $R_i(k) = V(s_i(k), \theta_v)$
  for $t \in \{k-1, k-2, ..., k_{start}\}$ **do**
    $R_i(t) \leftarrow r_i(t) + \gamma R_i(t)$
    $A^{\pi}(s_i, a_i; \theta, \theta_v) \leftarrow R_i(t) - V(s_i(t); \theta_v)$
    $H(\pi(s_i(t); \theta)) \leftarrow$
    $-\sum_{a_i(t)} \pi(s_i(t), a_i(t)) log \pi(s_i(t), a_i(t), \theta)$
    Accumulate gradients θ:
    $d\theta' \leftarrow d\theta' + \nabla_{\theta} log \pi(a_i|s_i; \theta) A^{\pi}(s_i, a_i; \theta, \theta_v) + \beta \nabla H(\pi(s_i; \theta))$
    Accumulate gradients θ_v:
    $d\theta'_v \leftarrow d\theta'_v + \partial (A^{\pi}(s_i, a_i; \theta, \theta_v))^2 / \partial \theta_v$
  end for
  Perform asynchronous update of θ' using $d\theta$ and of θ'_v using $d\theta_v$
end while

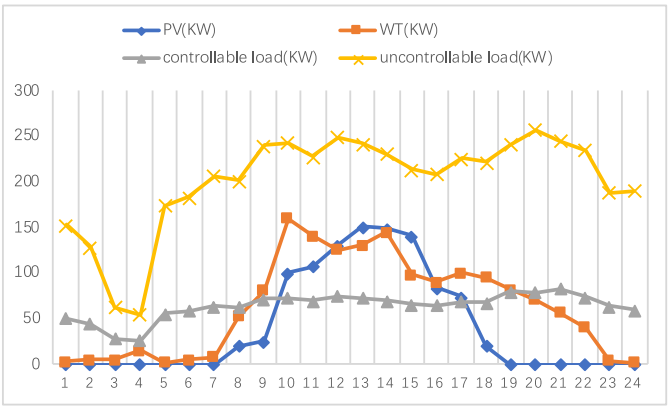


Fig. 4. Power of photovoltaic and wind power generation, and the demand of controllable loads and uncontrollable loads.

TABLE II
POLLUTION EMISSION DATA

Power generation technology (g/kwh)	SO_2	NO_x	CO_2	CO
Photovoltaic	0	0	0	0
Wind power	0	0	0	0
Micro gas turbine	0.000928	0.6188	184.0829	0.1702

TABLE III
DISTRIBUTION GENERATION UNIT COST

Power generation technology ($/kwh)	Power generation cost	Environmental cost
Photovoltaic	0.1736	0
Wind power	0.636	0
Micro-gas turbine	0.0868	0.00231

V. NUMERICAL RESULTS

A. System Description

In order to train the economic dispatch model based on DRL, we utilize offline data set for training with the load data of photovoltaic, wind power, micro gas turbine, and industrial users from [27] and [32]. Fig. 4 shows the power of photovoltaic and wind power generation on a random day, and the demand of controllable loads and uncontrollable loads. Wind power generation is denoted as WT and photovoltaic is denoted as PV. The maximum power of the micro gas turbine is set to 200 kW, since the industrial load mainly varies accounting to the industrial production, and there are no particularly obvious peak-to-valley costs. The period with higher load demand are 9.00–10.00, 12.00–14.00, and 19.00–21.00, and the period with lower load demand is 1.00–5.00. It can be seen that there is a large peak-to-valley difference between photovoltaic and wind power generation. The peak period of photovoltaic is 10.00–16.00, and the peak period of wind power generation is 10.00–18.00. The emission cost of pollution and the operation and maintenance cost of photovoltaic, wind power, and micro gas turbines are presented in Tables II and III.

The reward value is a key component of assessing the quality of movements and guiding the effectiveness of the learning process. In order to better set the reward value, we set the reward value as a function related to cost through repeated experiments. Because we want to minimize the cost of the VPP, the reward value is negative. The greater the weight, the greater the penalty. K_1, K_2, K_3, and K_4 are the weight values we set. The photovoltaic and wind power are cheaper and more environmentally friendly than micro gas turbines power generation in a real scenario. We set K_3 and K_4 to 10, and set K_1 and K_2 to 1. Therefore, the load is mainly powered by wind power and photovoltaics, and the remaining part is supplemented by micro gas turbines or the controllable load is reduced through demand response.

In this article, we run the numerical experiment on a PC with an 8-core i5 CPU and 16 GB. The number of threads is 8, that is, each local actor and critic network is equivalent to a subthread. The environment is learned asynchronously through subthreads, and the learning results are updated to the global network at intervals. Parallel simulation of state solves the problem of nonconvergence caused by continuous state updating. The discount factor is 0.90 and the entropy weight value is 0.01.

Usually, actor updating is guided by critic to produce a high reward, where critic updating is faster than actor. When the learning rate increases, the convergence rate is faster. However, the higher learning rate may lead to local optimum rather than global optimum. We, therefore, set the learning rate moderate.

In this section, we would preserve the structure of the neural network model in the DRL-based algorithm in detail below. The action is obtained by random sampling with the Gaussian distribution according to the state. In our algorithm, the state is represented as a 5-D vector, and the resulting action has four dimensions. We use the neural network model to calculate the Mu and Sigma parameters required for the Gaussian distribution. We input the state into the Mu network and the Sigma network, respectively, producing 4-D Mu and Sigma parameters. Among them, the Mu network is composed of two MLP layers. The input dimension of the first layer is 5, and Tanh is used for activation. The output is activated by Softplus. The Sigma network is also composed of two MLP layers, the first layer input dimension is 5, activated by Tanh and input two-layer neural network, and output dimension is 4, activated by Softplus. After that, we randomly sampled the 4-D actions through the Gaussian distribution. We calculate Q value based on state and action using the critic network. Actions are encoded using another MLP, input dimension 5, activated using Tanh. We then concatenate the two encoded outputs using a linear layer, with the final output dimension of 1. There are many random choices at the starting point of learning. However, through many iterations, the economic dispatch model converges and learns to select actions for the optimization goal. We use offline data set for training the optimal economic dispatch strategy. The main advantage of DRL is that the model can be applied online in real environment after training completely on such offline data. This online environment changes slightly, the DRL model can learn these

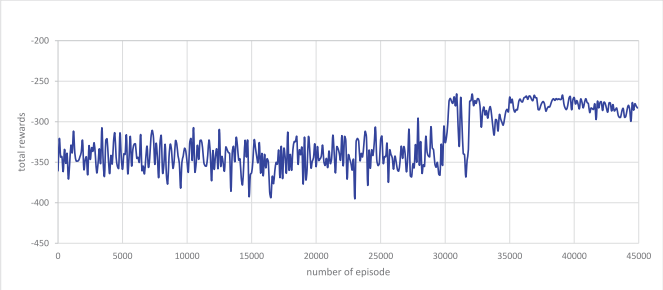


Fig. 5. Convergence process of our proposed algorithm.

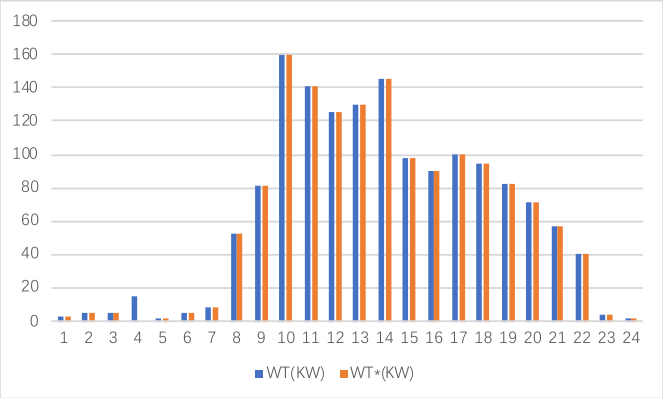


Fig. 6. Optimized results for wind power.

changes and dynamically adjust actions to achieve optimal dispatch.

B. Result and Discussion

In order to verify the convergence of the proposed algorithm, we sampled 100 days of data as the training data. Each episode is run on each day of 100 days. After running 45 000 episodes, the model can choose the best action. The algorithm is converged to take a total of 132 000 s. There are 24 steps in each episode, and each step is 1 h. The iterative process is shown in Fig. 5. The action is randomly sampled with the Gaussian distribution based on the state. We can see that the algorithm has large fluctuations in the first 30 000 episodes, mainly due to the randomness of strategy selection. As the exploring procedures, there is a significant fluctuation for the reward value. Due to the constraint of the action interval and the constraint of the equation, the fluctuation interval is about −300 to −400 kW. After training 32 000 episodes, the training makes a good breakthrough, because the model learned how to choose the best action. After 35 000 episodes, the model started to converge. The training results show that the proposed model can minimize the cost of a fully trained VPP operator. Although there are many random choices and many iterations at the beginning of learning, our proposed DRL-based model can converge to choose actions close to the best target value.

In order to verify the effectiveness of our proposed algorithm, we select the data from different agents in the last stage of training. Photovoltaic and wind power are cheaper and more environmentally friendly than micro gas turbines power generation in the VPP. Therefore, the load is wished to

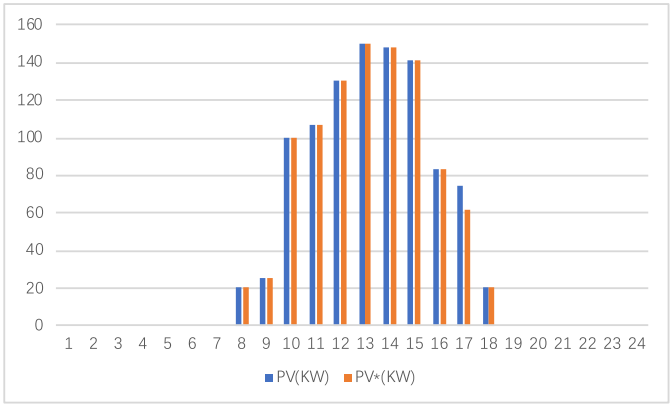


Fig. 7. Optimized results for photovoltaic.

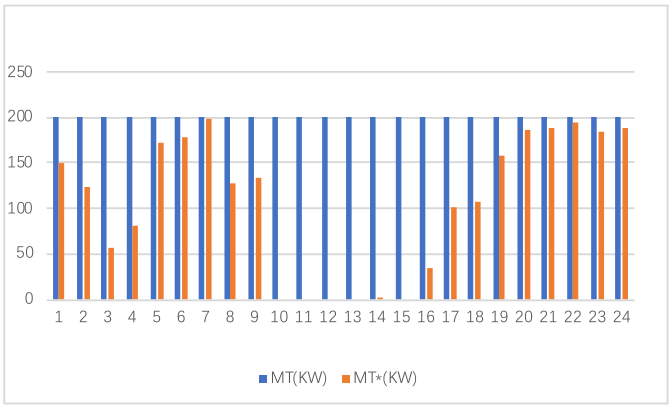


Fig. 8. Optimized results for micro gas turbine.

be powered by photovoltaic and wind power, and the remaining part of the load is supplemented by micro gas turbines or the controllable load through demand response. Figs. 6–8 show the comparison between the power generated by wind power, photovoltaic, and micro gas turbines and the actual power consumption. The yellow line is the power consumption of wind power, photovoltaic, and micro gas turbines. The blue line is the power generation of wind power, photovoltaic, and micro gas turbines. The horizontal axis is time (h), and the vertical axis is power (KWs). From Figs. 6 and 7, it can be seen that the difference between the actual power generation and the final power consumption of wind and photovoltaic power is approximately 0 per hour. At 1.00–7.00 and 23.00–24.00, the actual power of both photovoltaic and wind power generation is small, the load needs to be powered by a micro gas turbine. As can be seen from Fig. 8, 1.00–7.00 and 23.00–24.00, micro gas turbines are the main power supply units. In Fig. 9, it can be seen that from 20.00 to 24.00, due to the large electricity demand of industrial users and the high cost of micro gas turbines, the proportion of controllable load reduced greatly in this period. Therefore, it can be concluded that the cost of the VPP is minimized using our proposed algorithm. Under the preset reward value, the early stages of learning are relatively random. During the training process, over time, the model learns the correct strategy selection, how to minimize the cost of VPPs with stable control of distributed power generation and demand response.

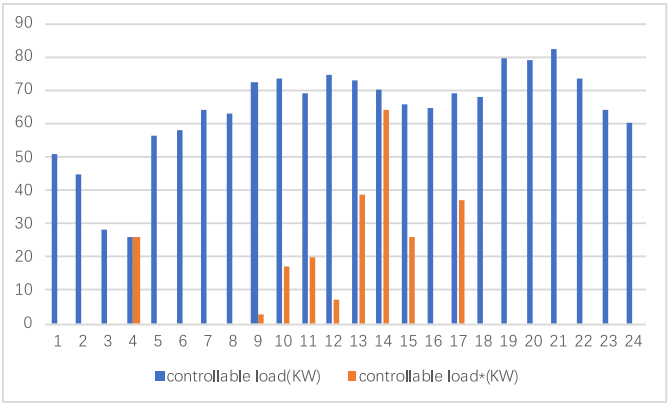


Fig. 9. Optimized results for controllable loads.

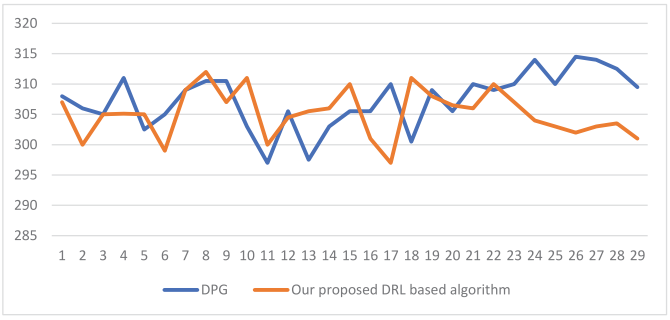


Fig. 10. Comparison of DPG and our proposed DRL-based algorithm averaged cumulative cost.

In order to verify the effectiveness of the proposed method, we further compare our proposed algorithm to other reinforcement learning algorithms. The deterministic policy gradient algorithm (DPG) [36] is used as a comparison algorithm to solve this kind of continuous action space problem. We compare our proposed algorithm with the DPG as shown in Fig. 10. The yellow curve is the DPG and the blue curve is the proposed algorithm. We use the 29-day data to compare the financial costs between the proposed method and the DPG method. We can see that from the 22nd day, the cost of our proposed method is significantly lower. Compared with our proposed method, because the DPG method uses the current moment's reward value as the unbiased estimate of the action state value function under the current strategy, the obtained strategy has higher variance and less generalization. However, the proposed method gets a smaller variance by subtracting the baseline. In order to break the correlation between data, our proposed method uses the asynchronous update mechanism to create multiple parallel environments, so that four workers with subthread update the parameters of the main network in parallel environment at the same time, since the workers in parallel do not interfere with each other.

In order to verify the lowest time complexity of the proposed method, we compare it with the DDPG and DPG methods. We set 45 000 episodes to analyze the running time of different methods. The results are shown in Table IV. Compared with different DRL methods suitable for solving VPP economic dispatch, the proposed method has the lowest time cost. Since each episode time is several milliseconds, in the real-time VPP economic dispatching scenario, it can make decisions in milliseconds based on the state's input. The traditional heuristic method needs to rerun the optimization process for each state, and its time cost is higher.

TABLE IV
TIME COMPLEXITY OF DIFFERENT METHODS

Methods	Total time(s)	The time of every episode(s)
Our proposed DRL based algorithm	13200	0.29
DPG	32010	0.71
DDPG	60300	1.34

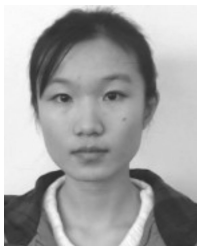
VI. CONCLUSION

In recent years, different types of distributed renewable power generation have been incorporated into the IoE. How to coordinate for integrating DGs effectively into the power grid, however, has been a challenging issue. VPPs can play a crucial role in addressing this issue. As for the core technique, economic dispatch plays an important role in minimizing cost. In this article, we proposed the DRL-based algorithm for the optimal economic dispatch in VPPs explicitly incorporating the stochastic characteristics of distributed renewable power generation. We further utilize an edge computing-based framework such that the optimal dispatch solution can be achieved with a reasonably low computation complexity. We evaluated the performance of our proposed algorithm with real meteorological and load data. The experimental results show that our proposed DRL-based algorithm can successfully learn the characteristics of DGs and industrial user demands. It can learn to choose actions to minimize the cost of VPPs. Compared with DPG and DDPG, our proposed method has a lower time cost.

REFERENCES

[1] N. L. Panwar, S. C. Kaushik, and S. Kothari, "Role of renewable energy sources in environmental protection: A review," *Renew. Sustain. Energy Rev.*, vol. 15, no. 3, pp. 1513–1524, 2011.
[2] D. Koraki and K. Strunz, "Wind and solar power integration in electricity markets and distribution networks through service-centric virtual power plants," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 473–485, Jan. 2018.
[3] T. Sousa, H. Morais, Z. Vale, and R. Castro, "A multi-objective optimization of the active and reactive resource scheduling at a distribution level in a smart grid context," *Energy*, vol. 85, pp. 236–250, Jun. 2015.
[4] L. Ju, H. Li, J. Zhao, K. Chen, Q. Tan, and Z. Tan, "Multi-objective stochastic scheduling optimization model for connecting a virtual power plant to wind-photovoltaic-electric vehicles considering uncertainties and demand response," *Energy Convers. Manag.*, vol. 128, pp. 160–177, Nov. 2016.
[5] P. Faria, J. Soares, Z. Vale, H. Morais, and T. Sousa, "Modified particle swarm optimization applied to integrated demand response and DG resources scheduling," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 606–616, Mar. 2013.
[6] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.
[7] L. Cheng, N. Qi, F. Zhang, H. Kong, and X. Huang, "Energy Internet: Concept and practice exploration," in *Proc. IEEE Conf. Energy Internet Energy Syst. Integr. (EI2)*, 2017, pp. 1–5.
[8] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[9] Y. Liu, M. Li, H. Lian, X. Tang, C. Liu, and C. Jiang, "Optimal dispatch of virtual power plant using interval and deterministic combined optimization," *Int. J. Elect. Power Energy Syst.*, vol. 102, pp. 235–244, Nov. 2018.
[10] Y. Zhang, X. Ai, J. Wen, J. Fang, and H. He, "Data-adaptive robust optimization method for the economic dispatch of active distribution networks," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3791–3800, Jul. 2019.
[11] E. Mashhour and S. M. Moghaddas-Tafreshi, "Bidding strategy of virtual power plant for participating in energy and spinning reserve markets—Part I: Problem formulation," *IEEE Trans. Power Syst.*, vol. 26, no. 2, pp. 949–956, May 2011.
[12] G. Chen and J. Li, "A fully distributed ADMM-based dispatch approach for virtual power plant problems," *Appl. Math. Model.*, vol. 58, pp. 300–312, Jun. 2018.
[13] S. Yang, S. Tan, and J.-X. Xu, "Consensus based approach for economic dispatch problem in a smart grid," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4416–4426, Nov. 2013.
[14] P. Pławiak, M. Abdar, and U. R. Acharya, "Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring," *Appl. Soft Comput.*, vol. 84, Nov. 2019, Art. no. 105740.
[15] M. Abdar, W. Książek, U. R. Acharya, R.-S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 179, Oct. 2019, Art. no. 104992.
[16] W. Książek, M. Abdar, U. R. Acharya, and P. Pławiak, "A novel machine learning approach for early detection of hepatocellular carcinoma patients," *Cognitive Syst. Res.*, vol. 54, pp. 116–127, May 2019.
[17] Q. Sun, D. Wang, D. Ma, and B. Huang, "Multi-objective energy management for we-energy in energy Internet using reinforcement learning," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, 2017, pp. 1–6.
[18] L. Du, L. Zhang, X. Tian, and J. Lei, "Efficient forecasting scheme and optimal delivery approach of energy for the energy Internet," *IEEE Access*, vol. 6, pp. 15026–15038, 2018.
[19] W. Liu, D. Zhang, X. Wang, J. Hou, and L. Liu, "A decision making strategy for generating unit tripping under emergency circumstances based on deep reinforcement learning," *Proc. CSEE*, vol. 38, no. 1, pp. 109–119, 2018.
[20] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Appl. Energy*, vol. 236, pp. 937–949, Feb. 2019.
[21] F. Li, J. Qin, and Y. Kang, "Multi-agent system based distributed pattern search algorithm for non-convex economic load dispatch in smart grid," *IEEE Trans. Power Syst.*, vol. 34, no. 3, pp. 2093–2102, May 2019.
[22] Y. Liu, C. Yang, L. Jiang, S. Xie, and Y. Zhang, "Intelligent edge computing for IoT-based energy management in smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 111–117, Mar./Apr. 2019.
[23] W. Li *et al.*, "On enabling sustainable edge computing with renewable energy resources," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 94–101, May 2018.
[24] X. Zhang, T. Bao, T. Yu, B. Yang, and C. Han, "Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid," *Energy*, vol. 133, pp. 348–365, Aug. 2017.
[25] T.-C. Chiu, Y.-Y. Shih, A.-C. Pang, and C.-W. Pai, "Optimized day-ahead pricing with renewable energy demand-side management for smart grids," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 374–383, Apr. 2017.
[26] B. Chai, J. Chen, Z. Yang, and Y. Zhang, "Demand response management with multiple utility companies: A two-level game approach," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 722–731, Mar. 2014.
[27] NREL. *Measurement and Instrumentation Data Center (MIDC) of NREL*. Accessed: Feb. 14, 2017. [Online]. Available: http://www.nrel.gov/midc
[28] J. Liu, Y. Miura, H. Bevrani, and T. Ise, "Enhanced virtual synchronous generator control for parallel inverters in microgrids," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2268–2277, Sep. 2016.
[29] M. Carrión and J. M. Arroyo, "A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem," *IEEE Trans. Power Syst.*, vol. 21, no. 3, pp. 1371–1378, Aug. 2006.
[30] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Başar, "Demand response management in the smart grid in a large population regime," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 189–199, Jan. 2016.
[31] S. Li, J. Yang, W. Song, and A. Chen, "A real-time electricity scheduling for residential home energy management," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2602–2611, Apr. 2019.
[32] *Open Energy Information and Data (Openei)*. Accessed: Feb. 12, 2017. [Online]. Available: http://en.openei.org/wiki/Main_Page
[33] Y. Liu, H. Yu, S. Xie, and Y. Zhang, "Deep reinforcement learning for offloading and resource allocation in vehicle edge computing and networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11158–11168, Nov. 2019.
[34] C. Wu, T. Yoshinaga, Y. Ji, T. Murase, and Y. Zhang, "A reinforcement learning-based data storage scheme for vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6336–6348, Jul. 2017.
[35] Y. Dai, D. Xu, S. Maharjan, G. Qiao, and Y. Zhang, "Artificial intelligence empowered edge computing and caching for Internet of vehicles," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 12–18, Jun. 2019.
[36] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. ICML*, 2014, pp. 387–395.



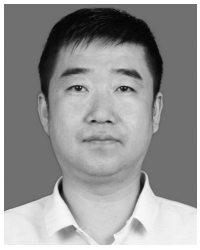
Lin Lin (Student Member, IEEE) received the B.S. degree in computer science and technology from Qufu Normal University, Rizhao, China, in 2017. She is currently pursuing the master's degree with the School of Data Science and Technology, Heilongjiang University, Harbin, China.

Her current research interests include the Internet of Energy and machine learning.

Xin Guan (Member, IEEE) received the bachelor's degree from the School of Computer Science and Technology, Heilongjiang University, Harbin, China, in 2001, the master's degree from the Harbin Institute of Technology, Harbin, in 2007, and the Ph.D. degree from the Graduate School of Science and Technology, Keio University, Tokyo, Japan, in 2012.

He is currently an Associate Professor with Heilongjiang University. His research interests include Internet of Things, energy Internet, and machine learning.

Dr. Guan was a recipient of the Ninth International Conference on Communications and Networking in China 2014 Best Paper Award. He served as the Technical Committee Member for IEEE-GLOBECOM, PIMRC, HPCC, and IWCMC. He served as a Reviewer for IEEE NETWORK, the *KSII Transactions on Internet and Information Systems*, the *EURASIP Journal on Wireless Communications and Networking*, *Security and Communication Networks* (Wiley), the *International Journal of Ad Hoc and Ubiquitous Computing*, and *ACM/Springer Mobile Networks and Applications*.



Yu Peng received the B.E. degree from Northeast Electric Power University, Jilin City, China, in 2000, and the M.E. degree from the Harbin University of Science and Technology, Harbin, China, in 2009.

He currently works with the Office of State Grid Heilongjiang Electric Power Company, Ltd., China. His research interests include power management, power grid operation technology, and relay protection technology.

Ning Wang received the B.E. degree from the Harbin Institute of Electrical Technology, Harbin, China, in 1995, and the M.E. degree from the Harbin University of Science and Technology, Harbin, in 2004.

He currently works as the Chief Engineer with the 93 Electric Power Bureau of State Grid Heilongjiang Electric Power Company, Ltd., China. His research interests include electrical operation of power plant, power dispatching management, power grid operation technology, and renewable energy operation management.

Sabita Maharjan (Senior Member, IEEE) received the Ph.D. degree in networks and distributed systems from the University of Oslo, Oslo, Norway, and Simula Research Laboratory, Fornebu, Norway, in 2013.

She is currently a Senior Research Scientist with Simula Metropolitan Center for Digital Engineering, Norway, and an Associate Professor (adjunct position) with the University of Oslo. She worked as a Research Engineer with the Institute for Infocomm Research, Singapore, in 2010. She was a Visiting Scholar with Zhejiang Univeristy, Hangzhou, China, in 2011, and a Visiting Research Collaborator with the University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 2012. She was a Postdoctoral Fellow with Simula Research Laboratory from 2014 to 2016. She publishes regularly in prestigious journals in her field, such as the IEEE TRANSACTIONS ON SMART GRID, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *IEEE Communications Magazine*, *IEEE Network Magazine*, *IEEE Wireless Communications Magazine*, and the IEEE INTERNET OF THINGS JOURNAL. Her current research interests include vehicular networks and 5G, network security and resilience, smart grid communications, Internet of Things, machine-to-machine communication, software defined wireless networking, and advanced vehicle safety.

Dr. Maharjan serves/has served in the Technical Program Committee of conferences, including top conferences like IEEE INFOCOM and IEEE IWQoS.

Tomoaki Ohtsuki (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1990, 1992, and 1994, respectively.

From 1994 to 1995, he was a Postdoctoral Fellow and a Visiting Researcher of electrical engineering with Keio University. From 1993 to 1995, he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005, he was with the Science University of Tokyo, Tokyo, Japan. In 2005, he joined Keio University, where he is currently a Professor. From 1998 to 1999, he was with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, USA. He has published more than 180 journal papers and 400 international conference papers. He is engaged in research on wireless communications, optical communications, signal processing, and information theory.

Prof. Ohtsuki was a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, the Ericsson Young Scientist Award 2000, the 2002 Funai Information and Science Award for Young Scientist, the IEEE First Asia–Pacific Young Researcher Award 2001, the Fifth International Communication Foundation Research Award, the 2011 IEEE SPCE Outstanding Service Award, the 27th TELECOM System Technology Award, the ETRI Journal's 2012 Best Reviewer Award, and the Ninth International Conference on Communications and Networking in China 2014 Best Paper Award. He served the Chair of IEEE Communications Society, Signal Processing for Communications, and Electronics Technical Committee. He served a Technical Editor for *IEEE Wireless Communications Magazine* and an Editor for *Physical Communications* (Elsevier). He is currently serving an Area Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Editor for IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He has served the General-Co Chair, the Symposium Co-Chair, and a TPC Co-Chair of many conferences, including IEEE GLOBECOM 2008, SPC, IEEE ICC2011, CTS, IEEE GCOM2012, SPC, IEEE APWCS, IEEE SPAWC, and IEEE VTC. He gave tutorials and keynote speech at many international conferences, including IEEE VTC and IEEE PIMRC. He was a Vice President of Communications Society of the IEICE and the President of Communications Society of the IEICE. He is a fellow of IEICE.