Frequency Regulation Capacity Offering of District Cooling System: An Intrinsic-motivated Reinforcement Learning Method

Peipei Yu, *Student Member, IEEE,* Hongcai Zhang, *Member, IEEE,* Yonghua Song, *Fellow, IEEE,* Hongxun Hui, *Member, IEEE,* and Chao Huang, *Member, IEEE*

***Abstract*—District cooling system (DCS), a type of large-capacity air conditioning system that supplies cooling for multiple buildings, is an ideal resource to provide frequency regulation services for power systems. In order to provide high-quality services and maximize DCS's revenue from the electricity market, an accurate estimation of DCS's regulation capacity is indispensable. Inaccurate regulation capacity estimation may lead to unsatisfactory cooling supply for buildings and/or poor regulation service quality that may be penalized by the market. However, estimating a DCS's regulation capacity is quite challenging, because a DCS usually has complex thermal dynamics to model and its cooling demands and regulation signals are usually highly stochastic. To address the above challenges, this paper proposes a DCS regulation capacity offering strategy based on deep reinforcement learning. It is model-free and can effectively tackle various uncertainties. Furthermore, considering that the training process of DRL needs lots of "trial and errors," which may harm the actual physical system by making "bad" decisions. We propose a novel intrinsic-motivated method based on pseudo-count to improve the efficiency of the training. Numerical studies based on a realistic DCS system illustrate the effectiveness of the proposed method.**

***Index Terms*—Demand response, capacity offering, district cooling system, reinforcement learning, intrinsic-motivation.**

I. Introduction

Carbon neutrality facilitates the energy reform in the power system, which is leading to a higher penetration of renewable energies (RENs) in the future (e.g., over 80% in China by 2060) [1]. The intermittent and fluctuated power supply from RNEs requires more regulation capacity for maintaining power balance between supply and demand sides [2]. District cooling system (DCS) is a type of large-scale and high-efficient air conditioning system that provides cooling services for multiple buildings, which is being widely adopted worldwide [3]. Because a DCS with multiple buildings usually has large-capacity thermal inertia, it can modulate its power temporally with negligible impacts on buildings' indoor temperatures [4]. Therefore, DCS is an ideal demand-side resource to provide frequency regulation capacity for power systems.

In most power regulation markets, the resources (e.g., DCS) are required to offer their regulation capacities to the market ahead-of-time; and then commit their promised regulation services following the system operator's instructions during real time operation. If a resource's commitment is not satisfactory, i.e., fails to provide the services as promised, it will be penalized [5]. Therefore, to avoid the punishment caused by poor regulation performance and receive maximum revenue, a DCS must be able to evaluate and offer its future available regulation capacity properly. However, this is quite challenging because of the following aspects:

1) *Complexity*: A DCS is a networked system that serves cooling for multiple buildings within a radius up to 2 km [6]. As a result, its thermal dynamics are too complex to model in a real-world system, because accurate system parameters are influenced by external environment factors and impractical to measure. Thus, traditional model-based methods for calculating regulation capacity are probably infeasible for DCS.
2) *Time-coupling*: Because that a building's thermal inertia is limited, its regulation capacity is time-coupling and may get saturated (similar to a battery storage system) [7]. Specifically, a building may get overly heated or cooled in one hour because of providing regulation services, so that it may not be able to provide services continuously for the following hours. Thus, this is a sequential decision-making problem that requires the proposed strategy to consider both the current and future rewards so as to maximize the cumulative revenue.
3) *Uncertainty*: On the one hand, random human behaviors make buildings' cooling demands uncertain; on the other hand, real-time regulation signals are highly stochastic and unpredictable, which can affect DCS's capacity offering strategy. However, the prior knowledge of uncertainty distribution is unknown and hard to be expressed by explicit formulations. Hence, to ensure the regulation performance and building temperature comfort at the same time, these uncertainties shall be properly handled, which further complicates the problem.

At present, most literature focus on the operation or control of DCS to minimize the expected energy cost [8]. For example, Sam et al. [9] propose a control strategy to max-

This paper is funded in part by the Science and Technology Development Fund, Macau SAR (File no. SKL-IOTSC(UM)-2021-2023, File no. 0003/2020/AKP) and in part by the Zhuhai Science and Technology Innovation Bureau (File no. ZH22017002210022PWC). (*Corresponding author: Hongcai Zhang.*)

P. Yu, H. Zhang, Y. Song, and H. Hui are with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macao, 999078 China (email: hczhang@um.edu.mo).

C. Huang is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 10083, China (e-mail: chao.huang@my.cityu.edu.hk).

imize cost saving of DCS under the time-of-use electricity prices. Mohammad et al. [10] design a least-annualized-cost mathematical approach to determine the optimal control of the DCS's water flow for meeting required cooling demands. These papers commonly focus on demand-driven operation strategies, which only passively control DCS to satisfy cooling demands while cannot effectively utilize DCS's flexibility [11]. To the best of our knowledge, the regulation capacity offering problem of DCS has not yet been studied previously.

Similarly with the DCS, household air conditioners (ACs) have been a research hotspot to provide regulation services for years [12]. Many researchers have studied their strategic regulation capacity offering problem [13]. For example, Xie et al. [14] develops a probability density estimation method to offer the operating reserve capacity for large-scale aggregated ACs. Li et al. [15] formulate the regulation capacity assessment problem as quadratic programming to improve flexibility and minimize the cost. However, the above two methods rely heavily on accurate system models and steady model parameters, which may not work for the complex DCS in our problem [16]. Cai et al. [17] evaluate the maximum regulation capacity of heating, ventilation and air conditioning (HVAC) systems using a pseudo-optimization method, while the uncertainties from cooling demands and market signals are not taken into account. Besides, the traditional optimization technique requires the system model with detailed and accurate parameters, which may be infeasible to be applied in a real-world DCS. Anwar et al. [18] present an extensive multi-perspective method to assess the capacity of aggregated residential HVACs as regulation reserves, while the regulation performance is not considered. Compared with household ACs, a networked DCS's energy station and end-users are located far away, which involves transportation in outdoor pipelines and results in huge thermal inertia and transportation time delay in DCS [9]. Besides, a DCS's model parameters are influenced significantly by environmental changes, such as ambient temperature and humidity [19]. Furthermore, a DCS's operating power can only be regulated indirectly through controlling water mass flow or supply water temperature, where the relationship is nonlinear and hard to model [11]. Thus, most model-based methods used in household ACs [20]–[22] are not applicable for the DCS capacity offering problem.

Recently, deep reinforcement learning (DRL) has been paid more attention as a model-free method and has successfully addressed many decision-making problems in power systems [23]. For example, Chen et al. [24] use the DRL method to seek the optimal demand response strategy, which can adapt to the changing environment information automatically. Liang et al. [25] propose an HVAC control algorithm in commercial buildings based on DRL to cope with the unknown thermal dynamic models and parameter uncertainties. To the best of our knowledge, there is no published paper that has adopted DRL in DCS for frequency regulation capacity offering problems, which is the focus of this paper. Although DRL does not require model knowledge of DCS, successfully applying it to our problem is nontrivial. First, the complex capacity offering problem is required to be formulated into a Markov decision process (MDP) mathematically, based on the operating characteristic of DCS. Second, the model-free algorithm is required to improve the low training efficiency caused by the absence of model knowledge, where the agent updates the policy through random explorations [26]. Last, because of poor policy and aggressive exploration at the early learning stage, the agent in traditional online DRL methods needs lots of "trial and errors" resulting in "bad" decisions [27].[1] A "bad" decision may lead to uncomfortable indoor temperatures in buildings and low regulation performance. In summary, when adopting DRL method in a DCS system, we need to formulate the problem in MDP mathematically, improve the training efficiency, and decrease the times of bad decisions.

To address the above issues, an intrinsic-motivated DRL method is proposed in this paper for a large DCS to determine its hour-ahead regulation capacity. The main contributions can be summarized as follows:

1) A model-free DRL-based capacity offering strategy is developed for DCS to provide regulation services, whose objective is to maximize revenue while considering both the buildings' temperature comfort and DCS's regulation performance. It can effectively address the challenges brought by the model's complexity, time-coupling and uncertainties.
2) A novel intrinsic-motivated method is proposed to improve the RL exploration efficiency. Compared with the traditional RL method, our method can converge to a similar result through fewer training episodes and enhance the convergence performance. Thus, the times of operation constraint violations during training can be effectively decreased.

This paper is organized as follows. Section II develops the preliminary model of a DCS. Section III formulates the sequence decision-making problem as an MDP and proposes the intrinsic-motivated DRL method to solve the control strategy. Section IV illustrates the effectiveness of the proposed method by numerical studies. Finally, Section V concludes this paper.

II. Preliminary Models

In this section, we first provide the DCS thermal dynamic model that considers the thermal inertia, ramp rate and temperature transmission time delay of mass flows. Second, we introduce the PJM's regulation market rules to calculate hourly performance score and revenue [5],[2] which further formulates the corresponding regulation capacity offering problem. It should be noted that the DCS model is established as a simulation environment for interacting with the DRL agent. While the agent does not know any knowledge about the system model, and its training process is still model-free.

[1]An RL model can be trained either offline based on historical data or online by interacting with the environment [28]. In this paper, considering that it is hard to collect adequate historical data to cover all the scenarios, especially those extreme scenarios that offer a too large or too small capacity to the market. Therefore, we adopt the online RL algorithm to train our model by interacting with the environment.

[2]The proposed method is also adaptable to other regulation markets.

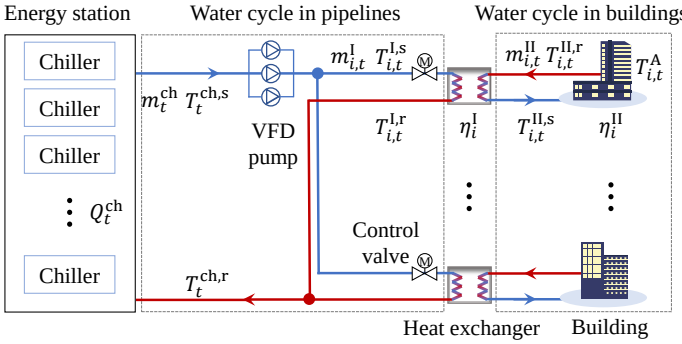


Fig. 1: Schematic diagram of a DCS.

A. Modelling of the DCS

1) Thermal dynamic process: The schematic diagram of a DCS is shown in Fig. 1, where the DCS is composed of one energy station, multiple pipelines and buildings. The blue and red lines represent the chilled and warm water, respectively. Chillers in the energy station, as the system's main power consumer, produce chilled supply water. The chilled supply water is pumped into pipelines to exchange heat with buildings and becomes warm return water. The following thermal dynamic model describes the thermal transmission from chillers to buildings.

According to the energy balance, the chillers' cooling power can be calculated as follows:

$$P_t^{\mathrm{ch}} = \frac{Q_t^{\mathrm{ch}}}{\mathrm{COP}}, \quad \forall t, \tag{1}$$

$$Q_t^{\mathrm{ch}} = c^{\mathrm{w}} m_t^{\mathrm{ch}} (T_t^{\mathrm{ch,r}} - T_t^{\mathrm{ch,s}}), \quad \forall t, \tag{2}$$

$$m_t^{\mathrm{ch}} = \sum\nolimits_{i\in\mathcal{I}} m_{i,t}^{\mathrm{I}}, \quad \forall t, \tag{3}$$

where P_t^{ch} and Q_t^{ch} are chillers' electricity power and cooling power at time t, respectively, in kW. The parameter COP is the coefficient of performance, and c^{w} is the specific heat capacity of water, in kJ/(kg·°C). The total mass flow rate in pipelines, m_t^{ch} (in kg/s), is the sum of buildings' mass flow rates, $m_{i,t}^{\mathrm{I}}$ (in kg/s), where set $\mathcal{I}$ denotes the set of all buildings. It can be seen from Eq. (2) that the difference between the supply and return water temperatures $T_t^{\mathrm{ch,r}} - T_t^{\mathrm{ch,s}}$ (in °C) can reflect the actual cooling demands and influence the cooling power. Further, the return water temperature $T_t^{\mathrm{ch,r}}$ in pipelines is determined by the warm return water from buildings, which can be calculated as follows:

$$c^{\mathrm{w}} m_t^{\mathrm{ch}} T_t^{\mathrm{ch,r}} = \sum\nolimits_{i\in\mathcal{I}} c^{\mathrm{w}} m_{i,t}^{\mathrm{I}} T_{i,t}^{\mathrm{I,r}}, \quad \forall t, \tag{4}$$

where $T_{i,t}^{\mathrm{I,r}}$ is the i-th building's return water temperature.

In the heat exchanger, the water in pipelines and buildings are isolated. The water in the two cycles only exchanges heat to achieve thermal transmission. The total exchanging heat $Q_{i,t}$ (in kW) can be calculated as follows:

$$\begin{aligned} Q_{i,t} &= c^{\mathrm{w}} m_{i,t}^{\mathrm{I}} (T_{i,t}^{\mathrm{I,r}} - T_{i,t}^{\mathrm{I,s}}) \cdot \eta_i^{\mathrm{I}} \\ &= c^{\mathrm{w}} m_{i,t}^{\mathrm{II}} (T_{i,t}^{\mathrm{II,r}} - T_{i,t}^{\mathrm{II,s}}), \quad \forall i \in \mathcal{I}, \forall t, \end{aligned} \tag{5}$$

where η_i^{I} is the heat transfer efficiency of the i-th building's heat exchanger. Symbols $m_{i,t}^{\mathrm{II}}$ (in kg/s), $T_{i,t}^{\mathrm{II,r}}$ and $T_{i,t}^{\mathrm{II,s}}$ are parameters of the ith building's water cycle, which indicate the water mass flow rate, return water temperature and supply water temperature, respectively.

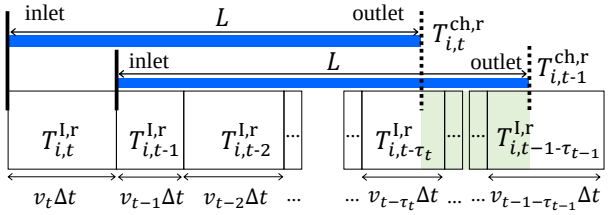


Fig. 2: Water transportation in a pipe.

In the building, the thermal energy brought by the building's cooling water provides cooling services and maintains its indoor temperature $T_{i,t}^{\mathrm{A}}$. The temperature dynamic is as follows:

$$D_i \frac{\partial T_{i,t}^{\mathrm{A}}}{\partial t} = \frac{T_t^{\mathrm{O}} - T_{i,t}^{\mathrm{A}}}{R_i} + \zeta_{i,t} - \eta_i^{\mathrm{II}} Q_{i,t}, \quad \forall i \in \mathcal{I}, \forall t, \tag{6}$$

where D_i and R_i are the i-th building's thermal capacity and thermal resistance, in kJ/°C and °C/kW, respectively. Symbol T_t^{O} (in °C) is the ambient temperature and η_i^{II} is the heat transfer efficiency. The uncertain heat loads (because of human activities) in buildings are denoted by $\zeta_{i,t}$ (in kW).

2) Control process: As shown in Fig. 1, control valves in each building can regulate the water mass flow $m_{i,t}^{\mathrm{I}}$ to adjust the cooling supply for buildings. However, the control process is not instantaneous, which exists the ramp rate, inertia and time delay. We assume that the aim value of the mass flow's regulation is represented by $\Delta m_{i,t}^{\mathrm{aim}}$ at time t, then the ramp rate limitation $\Delta m_i^{\mathrm{ramp}}$ can be expressed as follows:

$$\Delta m_{i,t}^{\mathrm{aim}} \le \Delta m_i^{\mathrm{ramp}}, \quad \forall i \in \mathcal{I}, \forall t. \tag{7}$$

For the inertia process, after a period of time t_1, the mass flow changes to a new value as follows:

$$m_{i,t+t_1}^{\mathrm{I}} = m_{i,t}^{\mathrm{I}} + (1 - e^{\frac{-t_1}{G_i}}) \Delta m_{i,t}^{\mathrm{aim}}, \quad \forall i \in \mathcal{I}, \forall t, \tag{8}$$

where G_i is the i-th building's inertia time constant.

The change of the mass flow influences the building's return water temperature, which further changes the cooling power. However, the temperature change exists time delay. As shown in Fig. 2, in a pipe with length L, its inlet water is from the building with temperature $T_{i,t}^{\mathrm{I,r}}$, and its outlet water goes to the chillers with temperature $T_t^{\mathrm{ch,r}}$. At each time step t, the water velocity v_t in the pipe equals to:

$$v_t = \sum\nolimits_{i\in\mathcal{I}} m_{i,t}^{\mathrm{I}} / A\rho^{\mathrm{w}}, \quad \forall t, \tag{9}$$

where A is the pipe's sectional area and ρ^{w} is the density of the water. Symbol τ_t represents the time delay index, which can be calculated as follows:

$$\tau_t = \min \tau_t, \quad \text{s.t.} \ \sum\nolimits_{j=0}^{\tau_t} v_{t-j} \Delta t \ge L, \tag{10}$$

where Δt is the interval between two adjacent time slots. The outlet water at time t is composed by the water with different temperatures, as the green area in Fig. 2. The average mixing temperature of the water can be calculated as follows:

$$v_t T_{i,t}^{\mathrm{ch,r}} = (\sum\nolimits_{j=0}^{\tau_t} v_{t-j} - \frac{L}{\Delta t}) T_{i,t-\tau_t}^{\mathrm{I,r}}$$

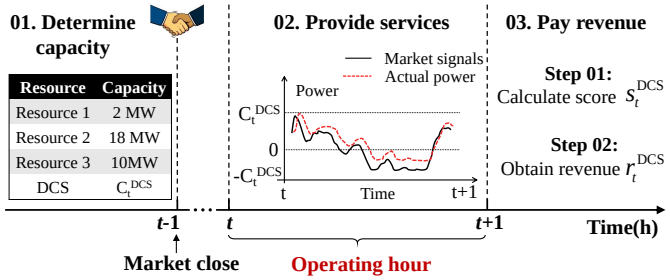


Fig. 3: Market rule of regulation services in PJM.

$$+\left(\frac{L}{\Delta t}-\sum\nolimits_{j=0}^{\tau_{t-1}-1} v_{t-1-j}\right)T^{\mathrm{I,r}}_{i,t-1-\tau_{t-1}} + \sum\nolimits_{j=\tau_t}^{\tau_{t-1}-1} v_{t-j}T^{\mathrm{I,r}}_{i,t-j}, \quad \forall i\in\mathcal{I},\forall t. \tag{11}$$

B. Regulation Capacity Offering in PJM

The market rule for resources to provide frequency regulation services is shown in Fig. 3. In PJM, the day-ahead regulation market closes at 14:15 the day before the operating day. However, in order to accurately reflect the resource's available capacity during the operating day, the regulation capacity offering is allowed to be changed and resubmitted until 60 minutes prior to the operating hour, at which time the hour-ahead market closes [29]. The market determines all resources' capacities one time slot before the operating time [t, t+1] [5]. When DCS finishes providing services at time t+1, it can receive its hourly revenue r_t^{DCS} as follows:

$$r_t^{\mathrm{DCS}} = C_t^{\mathrm{DCS}} s_t^{\mathrm{DCS}} p_t^{\mathrm{PJM}}, \quad \forall t, \tag{12}$$

where C_t^{DCS} (in MW) and $s_t^{\mathrm{DCS}} \in [0,1]$ are DCS's hourly regulation capacity and performance score, respectively. Parameter p_t^{PJM} is the market clearing price, which is uncertain and determined by all market participators. The DCS is assumed as a market price-taker because its regulation capacity is not large enough to influence the market price.

The performance score s_t^{DCS} reflects the resource's regulation quality, which assesses the difference between the actual regulation and the received signal from three aspects: 1) precision; 2) correlation; and 3) time delay, as follows:

$$s_t^{\mathrm{DCS}} = (s_t^{\mathrm{pre}} + s_t^{\mathrm{cor}} + s_t^{\mathrm{delay}})/3, \quad \forall t, \tag{13}$$

where s_t^{pre}, s_t^{cor} and $s_t^{\mathrm{delay}} \in [0,1]$ are the corresponding scores of precision, correlation and time, respectively. Because of space limitations, we cannot introduce this score in detail. The score calculation rules can refer to the PJM manual [29].

For a DCS, in order to maximize the revenue r_t^{DCS}, it should strategically offer its regulation capacity and maintain a high-enough performance score. However, if a DCS offers a too large regulation capacity while cannot commit, it may receive a low-performance score and be kicked out of the market. In PJM, there are two score requirements for the participants, which are the disqualification threshold score 0.4 and the test qualification score 0.75 [30].[3] To let the agent's strategy meet both two score requirements, we adopt the higher one as the agent's performance criterion $\underline{s}$=0.75.[4] This is because we should ensure that the trained agent can work well both in the qualification test stage before entering the market and the daily operation stage after entering the market. Considering that a DCS's regulation capacity is time coupled, the objective of DCS is to maximize its cumulative revenue during a day as follows:

$$\max_{C_t^{\mathrm{DCS}}} \quad r^{\mathrm{DCS}} = \sum\nolimits_t C_t^{\mathrm{DCS}} s_t^{\mathrm{DCS}} p_t^{\mathrm{PJM}}, \tag{14}$$

$$\text{s.t.: } s_t^{\mathrm{DCS}} \geq \underline{s}, \quad |T^{\mathrm{A}}_{i,t} - T^{\mathrm{set}}_{i,t}| \leq \Delta\overline{T_i}, \quad \forall i\in\mathcal{I},\forall t, \tag{15}$$

where $T^{\mathrm{set}}_{i,t}$ and $\Delta\overline{T_i}$ are the building's set temperature and its required comfortable range, respectively. Eq. (14) shows the objective of the cumulative revenue. In our objective, the additional power cost and valve life influence caused by regulation services do not be taken into account, because we assume that they are insignificant compared with the regulation revenue from markets. Eq. (15) shows two constraints when DCS provides the regulation capacity to the market, which are the performance score requirement from the market and the temperature comfort requirement from buildings.

Remark 1. *The capacity offering is expected to be large to increase the revenue according to Eq.(12). However, the requirements from the performance score and buildings' indoor temperature comforts limit the offered regulation capacity. Considering the model complexity and the environment uncertainty, it is infeasible to adopt traditional model-based methods. Meanwhile, the offer strategy should consider the time-coupled decision, which presents another challenge to our problem. To address these challenges, a model-free DRL method is proposed in Section III.*

III. Regulation Capacity Offering based on DRL

This section proposes an intrinsic-motivated DRL framework to determine the regulation capacity offering strategy. First, the decision-making problem is formulated as an MDP mathematically, whose objective is to maximize DCS's accumulative revenue with qualified services. In the proposed framework, both the uncertainties of real-time prices and regulation signals are considered. Then, considering that the DCS's model is usually unavailable, the strategical policy of the MDP is solved by a model-free method based on the policy gradient. Furthermore, intrinsic motivation is designed to improve training efficiency, which can decrease the times of score violations during the training process.

A. Regulation Capacity Offering based on MDP

In order to use a model-free algorithm, we first need to design a proper MDP to describe the DCS capacity offering problem in a mathematical way. A well-designed MDP requires a state space that captures necessary system information and guarantees Markov property. Besides, the reward function

[3]The participant must satisfy the test qualification score before it is permitted to enter into the regulation market. During daily operation, if the participant violates the disqualification threshold score, it will be kicked out of the regulation market. Training two different controllers for the two different stages will be considered as our future research direction.

[4]The effectiveness of the proposed method does not rely on the given constrained performance criterion. Our method can be readily extended to consider other criteria based on the different market rules.

is required to achieve the trade-off between the revenue reward and constraint violation penalties. In our problem, the observed state variables mainly consist of the DCS operating state (e.g., power consumption and indoor temperature) and historical market information (e.g., prices and signals). At each time step, the state can be fully observed because the historical market information is already known and the DCS operating state can be measured. Besides, the next state only depends on the current state and uncertain future environment, so the state space satisfies the Markov property. Meanwhile, the state transition can be considered as a memory-less stochastic process, where the next state is influenced by uncertainties from both the real-time market and cooling demands. Therefore, our problem is qualified to be modeled as an MDP.

The regulation capacity offering problem can be described as an MDP with a 5-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. Symbol $\mathcal{S}$ is a state space; $\mathcal{A}$ is an action space; $\mathcal{P}$ is a transition probability function; $\mathcal{R}$ is a reward function; $\gamma \in [0, 1]$ is a discount factor to balance the immediate and future rewards. Considering that the accurate DCS model can not be obtained, the transition probability function is not necessary to be defined in the proposed model-free method. The other elements are introduced in detail as follows:

1) Action: The decision variable at each time step t is the agent's action, which is defined as the DCS's hourly capacity offering to the electricity market, $\boldsymbol{a}_t = [C_t^{\text{DCS}}]^\intercal, \forall t$. The action is a continuous variable that satisfies $\boldsymbol{a}_t \in [C^{\min}, C^{\max}]$, where $C^{\min}$ and $C^{\max}$ represents the DCS's minimum and maximum regulation capacity, respectively. Because the market signal σ_t^{s} is a value between $[-1, 1]$, the power regulation range of DCS during services is within $[-\boldsymbol{a}_t, \boldsymbol{a}_t]$. In other words, the power shall be able to be regulated both up and down according to the signals. If the signal is negative, the power is regulated down; otherwise, it is regulated up.

2) State: State space is the observation of the DCS's real-time operation, which is also the input of the agent. Here, the observed state space at each time step t is defined as follows:

$$\boldsymbol{s}_t = [t, p_{t-1}^{\text{PJM}}, P_{t-1}^{\text{ch}}, \mathbb{E}[\Delta T_{i,t-1}], \mathbb{E}[\sigma_{t-1}^{\text{s}}], f(\sigma_{t-1}^{\text{s}})]^\intercal, \quad \forall t,$$

where the scale of the state space is $|\mathcal{S}| = 6$. Note that the key idea for the state design is to select those state variables that have strong relevance to the decision-making. Because the state that includes too much irrelevant information may take counterproductive effect on the convergence of DRL [31]. Hence, the designed MDP does not include all the DCS state information, and the lost information is assumed as the environment uncertainty to simplify the problem [32].[5] The observed parameter t is the operating time that reflects the ambient temperature and cooling demands; p_{t-1}^{PJM} is the last hour's market prices. Parameter P_{t-1}^{ch} is DCS's power consumption to indicate the maximum regulation capacity, where a higher power means a larger capacity. Symbol $\Delta T_{i,t-1}$ is the building's temperature deviation from its set value, i.e., $\Delta T_{i,t-1} = T_{i,t-1}^{\text{A}} - T_{i,t-1}^{\text{set}}$. The expectation of all the buildings' deviations $\mathbb{E}[\Delta T_{i,t-1}]$ reflects buildings' temperature comfort and regulation thermal potential for the next time step. The parameters $\mathbb{E}[\sigma_{t-1}^{\text{s}}]$ and function $f(\sigma_{t-1}^{\text{s}})$ describe the statistical characteristics of the market signal σ_{t-1}^{s}, which can help the agent predict the signal trend. The former parameter $\mathbb{E}[\sigma_{t-1}^{\text{s}}]$ is the expectation of regulation signals, which denote the average regulation degree. The later $f(\sigma_{t-1}^{\text{s}})$ is a self-defined function to represent the maximum cumulative regulation power brought by regulation signals during the operating hour, whose definition is as follows:

$$f(\sigma_{t-1}^{\text{s}}) = \max_{x \in [t-1, t]} \left| \int_{t-1}^{x} \sigma_x^{\text{s}} \mathrm{d}x \right| \cdot C_t^{\text{DCS}}, \quad \forall t. \tag{16}$$

Specifically, as shown in Fig. 4, the DCS's power is regulated up when the signal is positive and down when it is negative. When the signal keeps in the same regulation direction in $[t-1, t_1]$, the cumulative regulation power keep increasing with time, as the shadow area shows. When the signal changes its direction at time t_1, the regulation power also changes the direction so that the cumulative regulation decreases.

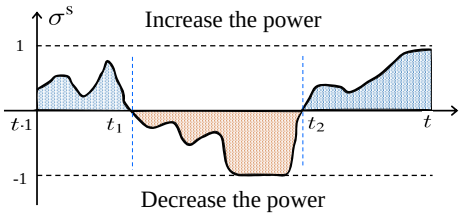


Fig. 4: The accumulated regulation brought by signals.

3) Reward: The reward function is designed to achieve the maximum revenue from the electricity market. Considering the constraint of the performance score, we use symbol κ_t to indicate the satisfaction of the score requirement (i.e., when $s_t^{\text{DCS}} \geq \underline{s}$, $\kappa_t = 1$; otherwise, $\kappa_t = 0$). Then, the reward is composed of two parts as follows:

$$\mathcal{R}_t = \beta_1 \kappa_t C_t^{\text{DCS}} s_t^{\text{DCS}} p_t^{\text{M}} - \beta_2 (1 - \kappa_t) \frac{C_t^{\text{DCS}}}{s_t^{\text{DCS}}}, \quad \forall t, \tag{17}$$

$$\kappa_t = \left(1 + \frac{s_t^{\text{DCS}} - \underline{s}}{|s_t^{\text{DCS}} - \underline{s}|} \right) / 2, \quad \forall t. \tag{18}$$

In Eq. (17), the former part is the obtained revenue and the later part is the penalty for score violations. It means the reward is negative for punishing the agent's decision, when the performance score is lower than the required value. Parameters β_1 and β_2 are weight factors of the corresponding items.

B. Regulation Capacity Offering based on Deep Deterministic Policy Gradient

In this subsection, we introduce the deep deterministic policy gradient (DDPG) algorithm, an effectively RL algorithm, to solve the proposed MDP. In the proposed MDP, the mapping rule from the state space to the action space is defined as a policy: $\pi : \mathcal{S} \to \mathcal{A}$. A deep neural network with parameter θ^π is adopted to express the policy π, called *actor network*. The agent's objective is to seek the optimal parameter $\theta^{\pi *}$ (i.e., π^* is the optimal policy) that can maximize the expected

[5]This simplification still satisfies the Markov property, because the next DCS operating state has no direct relationship with the historical state.

This article has been accepted for publication in IEEE Transactions on Smart Grid. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSG.2022.3220732

long-term reward $J(\pi)$, as follows:

$$\max_{\pi} J(\pi) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t \mathcal{R}_t\right], \tag{19}$$

where τ=$\{\boldsymbol{s}_0, \boldsymbol{a}_0, \boldsymbol{s}_1, ..., \boldsymbol{a}_{T-1}, \boldsymbol{s}_T\}$ is the sequence of continuous states and actions; $\mathcal{R}_t$ is the immediate reward for a single time step t. This objective guides the trained agent to select an action that can maximize the long-term revenue, instead of maximizing the current hour's revenue.

In order to update the policy and converge to the optimal one $\theta^{\pi} \rightarrow \theta^{\pi^*}$, an *action-value* function is defined to critique the probability distribution of the next state as follows:

$$Q^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t \mathcal{R}_t | \boldsymbol{s}_0 = \boldsymbol{s}, \boldsymbol{a}_0 = \boldsymbol{a}\right], \tag{20}$$

where Q value denotes the expected reward taking action $\boldsymbol{a}$ at state $\boldsymbol{s}$ following the policy π. Note that at the same state $\boldsymbol{s}$, a larger Q value means a better action $\boldsymbol{a}$. Here, because of the absence of the model knowledge, function Q is approximated by a neural network with parameter θ^Q, called *critic network*. After training, the critic network can achieve the estimation of the state probability distribution, which addresses the uncertainties.

With the interaction between the agent and the environment, the transitions $(\boldsymbol{s}_t, \boldsymbol{a}_t, \mathcal{R}_t, \boldsymbol{s}_{t+1})$ are stored into an *experience reply buffer* as the training data. After each episode, the agent randomly samples L training data from the reply buffer to update its policy parameter θ in the policy gradient direction [33], as follows:

$$\theta \leftarrow \theta + \lambda \bigtriangledown_{\theta} J^{\pi}, \tag{21}$$

$$\bigtriangledown_{\theta} J^{\pi} = \frac{1}{L}\sum_{l=1}^{L} \bigtriangledown_{\boldsymbol{a}} Q(\boldsymbol{s}_l, \pi(\boldsymbol{s}_l)) \bigtriangledown_{\theta} \pi(\boldsymbol{s}_l), \tag{22}$$

where λ is the learning rate, $\bigtriangledown_{\theta} J^{\pi}$ is the gradient of the agent's expected reward with respect to parameters θ. The key idea of the gradient equation (22) is to increase the action's possibility that can gain a higher reward.

In Eq. (22), in order to calculate the gradient of Q function (i.e., critic network) and policy π (i.e., actor-network), their corresponding loss functions are required [33], as follows:

$$L(\theta^Q) = \frac{1}{L}\sum_{l=1}^{L}[\mathcal{R}_l + Q'(\boldsymbol{s}_{l+1}, \pi'(\boldsymbol{s}_{l+1})) - Q(\boldsymbol{s}_l, \boldsymbol{a}_l)]^2, \tag{23}$$

$$L(\theta^{\pi}) = -\frac{1}{L}\sum_{l=1}^{L} Q(\boldsymbol{s}_l, \boldsymbol{a}_l), \tag{24}$$

where Q' and π' are the target networks that are the copies of networks Q and π, respectively. The loss function of critic network Q adopts the mean squared error to approximate its target value, which considers the next step's Q value by $\mathcal{R}_l + Q'(\boldsymbol{s}_{l+1}, \pi'(\boldsymbol{s}_{l+1}))$. The target value is designed according to the temporal-difference learning method, which dynamically approximates the Q value using the next step's Q [34]. In this way, the next action's Q value can be propagated to the current action to consider all the future steps. The loss function of actor-network π adopts $-Q$ value. Based on the Eqs. (21)-(24), the actor and critic networks can be updated

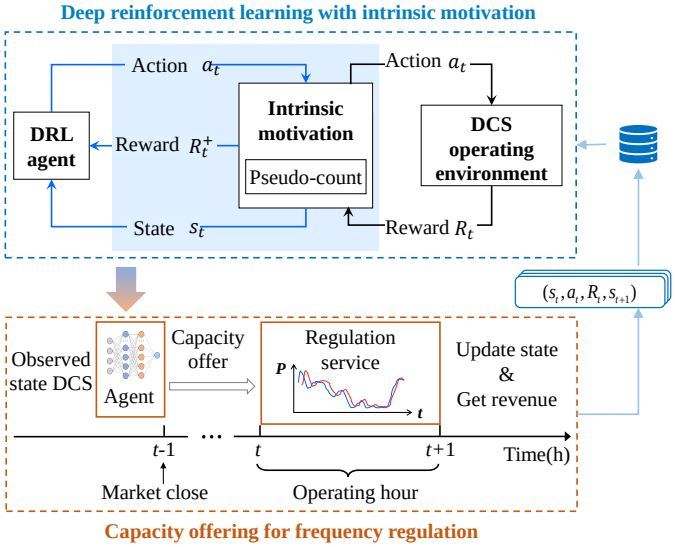


Fig. 5: Framework of the proposed strategy.

to improve the probability of high-reward actions at each time step. Further, the environment uncertainty can be addressed without any knowledge of the DCS model [35].

Remark 2. *The hourly capacity offering problem is formulated as an MDP and solved by DDPG to seek an effective strategy. According to the real-time operating state of DCS, the well-trained agent can make continuous decisions automatically, which adapt to the uncertain market signals and stochastic cooling demands.*

C. Count-based Exploration and Intrinsic Motivation

As shown in Fig. 5, the well-trained agent is applied online to offer regulation capacity, which further collects the experience data for the offline training process. *In the capacity offering process:* According to the DCS's current state, the agent's policy determines the capacity offering action before the market closes and sends it to the DCS. Then the DCS executes the decision following market signals during the operating hour. After the operating hour, the DCS receives the revenue from the market and observes its updated operating state. Finally, the DCS sends the revenue and new state to the agent for the next decision. *In the training process:* To update the agent's policy, the agent uses all the historical experiences $(s_t, a_t, R_t.s_{t+1})$ generated by the interactions with the environment. Considering that the designed MDP does not have a system model to know the deterministic state transition, the model-free algorithm's training efficiency will become relatively low compared with a model-based one. Meanwhile, the agent's exploration during the training process based on lots of "trial and errors" in the traditional DRL framework may lead to many "bad" decisions. As a result, the DCS may often violate the constraints in Eq. (15) and receive a low regulation performance score.

To address this issue, we propose the *intrinsic motivation method* over the environment's reward to improve the training efficiency and ensure a "safer" training process. The key idea of the intrinsic motivation method is to add an exploration

bonus in the reward function to encourage the exploration of new states. The bonus is designed based on the state visit count $N(\boldsymbol{s})$ that describes the total historically visited times of states, where a larger visit count means a smaller information gain for the agent [36]. According to the state visit count, the agent tends to exploit historical experiences in its familiar states with large $N(\boldsymbol{s})$, and explore more unknown states with small $N(\boldsymbol{s})$ [37]. As a result, the agent will explore the state space with fewer episodes, which can significantly accelerate the training process and reduce the number of "bad decisions."

However, in our problem, the state space is continuous, which causes the *empirical-count number* $N(\boldsymbol{s})$ to be invalid because the same state is rarely revisited. To overcome this challenge, we propose the *pseudo-count number* $\widehat{N}(\boldsymbol{s})$ to approximate the empirical-count number, which is derived by a density model of experienced states introduced as follows.

1) Pseudo-count number: To derive the calculation formulation of pseudo-count, we first give the definition of the *probability* and *re-coding probability*. For the state space $\mathcal{S}$, $\boldsymbol{s}_{1:n}$ represents a sequence states $(\boldsymbol{s}_1, ..., \boldsymbol{s}_n) \in \mathcal{S}^n$ with length n. If a state's number $\boldsymbol{s}_t$ in the sequence $\boldsymbol{s}_{1:n}$ is $N_n(\boldsymbol{s}_t)$, then for each $\boldsymbol{s}_{1:n}$, the probability distribution over $\mathcal{S}$ is given as:

$$\rho_n(\boldsymbol{s}_t) = \rho(\boldsymbol{s}_t|\boldsymbol{s}_{1:n}) = \frac{N_n(\boldsymbol{s}_t)}{n}, \quad \forall t, \tag{25}$$

where $N_n(\boldsymbol{s}_t)$ is the empirical-count number in all the past n experiences. Distribution $\rho_n(\boldsymbol{s}_t)$ is a density model that assumes all states are distributed independently. Inspired from the statistical compression, when the density model $\rho_n(\boldsymbol{s}_t)$ observes a new state $\boldsymbol{s}_t$, the probability of state $\boldsymbol{s}_t$ is called *re-coding probability*, formulated as follows:

$$\begin{aligned}\rho'_n(\boldsymbol{s}_t) &= Pr_\rho(S_{n+2} = \boldsymbol{s}_t|S_1...S_n = \boldsymbol{s}_{1:n}, S_{n+1} = \boldsymbol{s}_t) \\ &= \rho(\boldsymbol{s}_t|\boldsymbol{s}_{1:n}\boldsymbol{s}_t), \quad \forall t,\end{aligned} \tag{26}$$

where $\rho'_n(\boldsymbol{s}_t)$ is a conditional probability distribution; $\boldsymbol{s}_{1:n}\boldsymbol{s}_t$ means the connection of n samples $\boldsymbol{s}_{1:n}$ and a state $\boldsymbol{s}_t$. Then, the relationship between the state $\boldsymbol{s}_t$'s pseudo-count number $\widehat{N}_n(\boldsymbol{s}_t)$ and the pseudo-count total number (of all states) $\widehat{n}$ is formulated as the following two constraints:

$$\rho_n(\boldsymbol{s}_t) = \frac{\widehat{N}_n(\boldsymbol{s}_t)}{\widehat{n}}, \qquad \rho'_n(\boldsymbol{s}_t) = \frac{\widehat{N}_n(\boldsymbol{s}_t) + 1}{\widehat{n} + 1}, \quad \forall t. \tag{27}$$

It means after observing one sample of state $\boldsymbol{s}_t$, the prediction probability of the same state $\boldsymbol{s}_t$ in the density model would increase correspondingly. Therefore, the pseudo-count number can be solved by the linear equation as:

$$\widehat{N}_n(\boldsymbol{s}_t) = \frac{\rho_n(\boldsymbol{s}_t)(1 - \rho'_n(\boldsymbol{s}_t))}{\rho'_n(\boldsymbol{s}_t) - \rho_n(\boldsymbol{s}_t)}, \quad \forall t. \tag{28}$$

Based on the above equation, the pseudo-count number $\widehat{N}$ can be derived through the density model $\rho_n(\boldsymbol{s}_t)$, which can approximate the empirical-count number N effectively in a continuous space.

2) Density model: For estimating the density $\rho_n(\boldsymbol{s}_t)$ based on historical samples $\boldsymbol{s}_{1:n}$, we use the Gaussian Mixture Model (GMM) as the density estimation model [38], which is a probabilistic model that is composed of K Gaussian distributions as follows:

$$\mu(\boldsymbol{s}_t) = \sum\nolimits_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{s}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \forall t, \tag{29}$$

where each Gaussian density $\mathcal{N}(\boldsymbol{s}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ has its own mean $\boldsymbol{\mu}_k$ and co-variance $\boldsymbol{\Sigma}_k$. Parameters c_k are the weight factors. The parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ of GMM are estimated by the Expectation Maximization algorithm [39]. Then, we normalize the state densities between zero and one (i.e., $0 \leq \rho \leq 1$) as:

$$\rho_n(\boldsymbol{s}_t) = \frac{\mu(\boldsymbol{s}_t)}{\sum_{j=1}^{n} \mu(\boldsymbol{s}_j)}, \quad \forall t. \tag{30}$$

Similarly, the re-coding probability $\rho'_n(\boldsymbol{s}_t)$ can be calculated by the above two equations based on state samples $\boldsymbol{s}_{1:n}\boldsymbol{s}_t$. After that, taking $\rho_n(\boldsymbol{s}_t)$ and $\rho'_n(\boldsymbol{s}_t)$ into Eq. (28), the pseudo-count number $\widehat{N}_n(\boldsymbol{s}_t)$ can be calculated.

3) Exploration bonus as intrinsic motivation: After calculating the state's pseudo-count number $\widehat{N}_n(\boldsymbol{s}_t)$, the agent adds an exploration bonus as the intrinsic motivation to its empirical reward function $\mathcal{R}_t$. The bonus form is as follows:

$$\mathcal{R}^+{}_t(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathcal{R}_t + \beta_3(\widehat{N}_n(\boldsymbol{s}_t) + 0.01)^{-1/2}, \quad \forall t, \tag{31}$$

where β_3 is the mixing coefficient. The new reward $\mathcal{R}^+{}_t$ replaces the traditional reward function $\mathcal{R}_t$ to calculate the gradient in Eq. (22), and updates the policy neural network π. Note that a smaller pseudo-count $\widehat{N}$ in Eq. (31) leads to a larger step reward to encourage the agent's exploration.

Remark 3. *The proposed method uses the state's pseudo-count number to approximate the empirical count, which is then used to incentivize the agent to explore unknown states. This method can decrease the total times of constraint violations by improving training efficiency, so that it is "safer" compared with the traditional RL.*

IV. Case Study

A. System environment

Based on the system schematic diagram of the realistic DCS in Hengqin, China, our simulation environment is established for both the training and testing process.[6] The aim of the simulation environment is to validate the effectiveness of our proposed method. It means, if we can deploy the agent in a real-world system, the agent can be trained directly in the real-world system. All the case studies are implemented in the simulated DCS that is composed of one energy station and 12 connected buildings. The total cooling capacity of DCS is 144 MW. The system parameters are illustrated in Table I, which are designed according to both the DCS technical guidelines [40] and Chinese National Standards (GB 31349-2014, GB 50019-2003, DBJ 15-51-2020 and NB/T 47004.1-2017). The market regulation instruction (σ_t^{s}) and hourly price (p_t^{PJM}) are collected from the realistic data in PJM [5] (one typical day is shown in Fig. 6). Meanwhile, the designed performance criterion ($\underline{s}$) for our agent is 0.75, which is the test qualification score in PJM. -The minimum capacity

[6]The DRL agent is trained and tested in the same environment, which is the characteristic of online DRL and there are lots of environmental randomnesses that need to be handled by the agent.

This article has been accepted for publication in IEEE Transactions on Smart Grid. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSG.2022.3220732

TABLE I: Parameter setting of DCS.

Symbols	Definitions	Values
COP	Coefficient of performance	5
$\eta_i^{\mathrm{I}}, \eta_i^{\mathrm{II}}$	Heat transfer efficiency	90%
c^{w}	Heat capacity	4.2 kJ/(kg·°C)
D_i	Buildings' thermal capacity	$10^6 \sim 10^7$ kJ/°C
R_i	Buildings' thermal resistance	0.001∼0.003 °C/kW
$T_{i,t}^{\mathrm{set}}$	Indoor set temperature	20∼23°C
$\Delta T_{i,t}$	Comfortable temperature range	-1∼1°C
$T_t^{\mathrm{ch,s}}$	Supply water temperature	3 °C
C^{min}	Minimum regulation capacity	0 MW
C^{max}	Maximum regulation capacity	144 MW

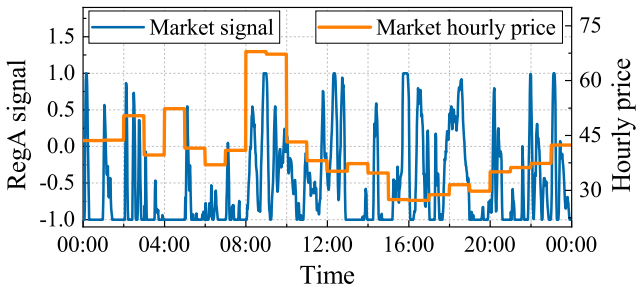


Fig. 6: RegA signal and hourly price in PJM.

requirement to access the market is set as 1MW, which means the DCS can participate in the regulation market only when its capacity offer is larger than 1MW.

B. Benchmarks

We implement three benchmarks to validate the benefits of the proposed method: 1) the proposed count-based DDPG strategy that uses the proposed reward function in Eq. (31) is denoted by B1; 2) the DDPG strategy that uses the original reward function in Eq. (17) is denoted by B2; 3) the heuristic method based on particle swarm optimization (PSO) is denoted by B3; and 4) the equal-proportional strategy is denoted by B4. The capacity offering in the benchmark B4 is directly proportional to the DCS's operating power. Thus, the B4 strategy participates in the regulation market every hour as long as its capacity offering satisfies the required value. Considering the actual operating power is unknown when determining the regulation capacity, we adopt DCS's average power of the same hour in the past five days as its base power. According to the base power P_t^{base} and a constant proportion $g\%$, the DCS calculates its hour-ahead capacity offering by:

$$C_t^{\mathrm{DCS}} = P_t^{\mathrm{base}} \cdot g\%, \quad \forall t. \tag{32}$$

To show the results based on different proportions, we set $g\%$ as 10%, 20% and 30%, and name the corresponding scenarios by B4-10%, B4-20% and B4-30%, respectively.

For the DRL methods B1 and B2, all parameter settings are the same except that B1 adopts the intrinsic motivation in Eq. (31). The weight factors are set as β_1=0.005, β_2=0.003 and β_3=0.05 [7]. The discount factor is set as γ=0.9. The number of mini-batch samples is set as L=200. The actor and critic networks both have two hidden layers with a scale of 128×128, whose learning rates are set as 0.001 and 0.0001, respectively. The smooth factors of the two target networks are all set as 0.005. The noise added to the action at each time step subjects to Gaussian distribution $\xi \sim N(0, \sigma^2)$, where σ is set as 0.3. The simulation is implemented in a Windows system, using the PyTorch framework in Python with an Intel core i7 CPU @3.0 GHz and 16GB memory.

C. Training Efficiency of RL Methods

Fig. 7 depicts the results of training episode rewards based on methods B1 and B2. In order to compare the same reward of the two methods, the reward curve of method B1 in Fig. 7 does not include the added exploration bonus item in Eq. (31). Two agents in B1 and B2 are initialized with the same random policy, then their policies are trained through interactions with the same environment for 7500 episodes. In the early training stage, two agents' policies are not good and unstable due to the absence of prior knowledge. After nearly 2000 episodes, the quality of the two policies are improved and become stable. The episode rewards of B1 are higher than those of B2 during the training process. Moreover, the final converged reward of B1 is also significantly higher than that of B2. It is noted that the converged policy in the two DRL algorithms are probably not the global optimums, but they are both acceptable sub-optimal solutions for our complex capacity offering problem. The optimality of the DRL policy for a specific work is difficult to prove mathematically [42].

Fig. 8 shows the violation penalties during the training process based on methods B1 and B2. If the performance score satisfies the requirement, the penalty is zero; otherwise, it will be negative. It can be seen that the agent in method B2 frequently violates the score constraint until it converges. However, the proposed method B1 can avoid significantly violating the score constraint after about 150 episodes. It is noted that, in Figs. 7 and 8, the effectiveness of the strategy at the beginning cannot be guaranteed, where the strategy's reward is quite low and the score violation is significant. This is a common limitation of online DRL when it is applied in a practical system. To avoid the unstable policy at the beginning, many researchers adopt a high-quality pre-trained policy to initialize the agent, such as combining transfer learning with DRL [43]. This is out of this paper's scope and will be an important future work.

D. Regulation Revenue and Performance Score

To validate the superiority of the proposed method B1, two months' regulation capacity offerings are simulated adopting all the benchmarks. For the model-free methods B1 and B2,

[7]Weight factors β_1, β_2 and β_3 are determined one by one through the following steps: Firstly, we adjust the reward scale and find a proper order of magnitude for β_1 through testing, to ensure that each step's reward lies near to 1 [41]; Secondly, based on the fixed β_1, we tune β_2 to penalize the score violations until the agent achieves the balance between the reward and penalty; Finally, we select β_3 from a coarse parameter sweep.

This article has been accepted for publication in IEEE Transactions on Smart Grid. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSG.2022.3220732

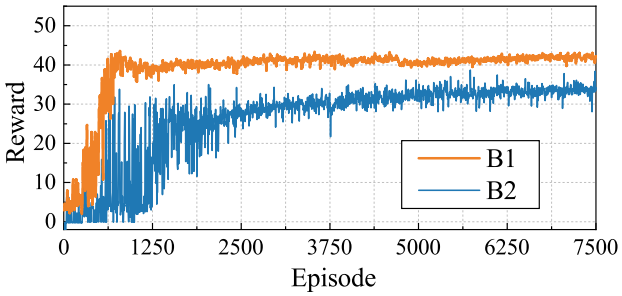


Fig. 7: Episode reward with (B1) and without (B2) the intrinsic motivation during the training process.

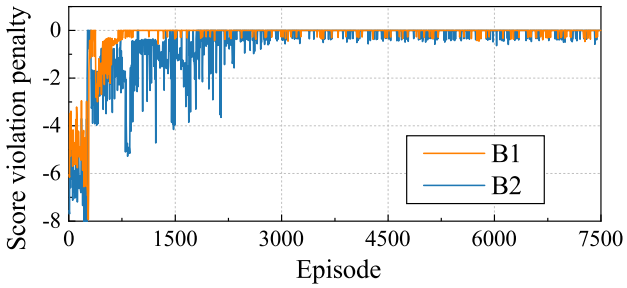


Fig. 8: Violation penalty with (B1) and without (B2) the intrinsic motivation during the training process.

the well-trained agents (after convergence) are adopted in the simulation. The distributions of all the daily revenues and regulation score violations in the test days based on different strategies are shown in Figs. 9(a) and 9(b), respectively. The profiles of hourly regulation scores in a typical operation day is shown in Fig. 10. If the hourly performance score satisfies the market requirement, it falls above the dotted line.

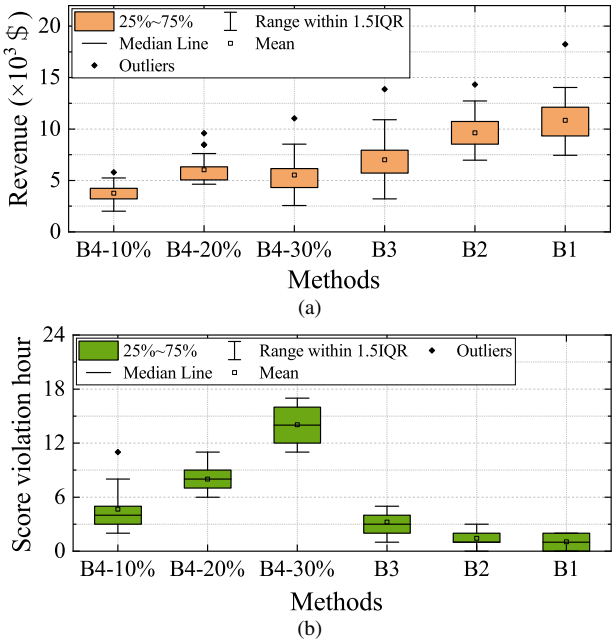


Fig. 9: The distributions of (a) the daily revenue, and (b) the daily score violation hours.

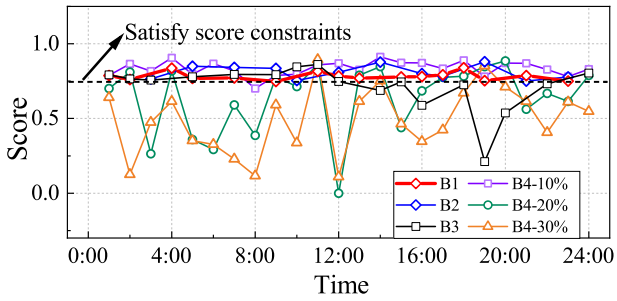


Fig. 10: Regulation performance score with five different capacity offering strategies.

For the three equal-proportion strategies, the score violations increase significantly with the regulation capacity offering. When the regulation capacity offering is too large, many buildings may become uncomfortable and quit the regulation services to recover their indoor temperatures during operation hours, which will affect the regulation performance scores. In B4-10%, the DCS's hourly capacity offerings are too small. Though it has high performance scores, it still cannot gain high revenue. By contrast, in B4-30%, the DCS offers two large regulation capacities that cause low-performance scores, which also leads to a decrease in the total revenue. Because B4-20% can better balance regulation capacity and performance scores, it outperforms B4-10% and B4-30%. The heuristic method B3 outperforms the three equal-proportional strategies in terms of both regulation revenue and performance score. However, it is worse than the two RL methods because it is hard for PSO to find the global optimal solution. The two RL-based methods (B1 and B2) significantly outperform the other benchmarks. They have lower score violations and higher regulation revenue. This is because the well-trained agents in the two DRL methods can predict the signal, and determine whether to participate in the market according to different operating states. This strategy fully exploits the building's thermal inertia by providing intermittent services. The average daily revenue of proposed method B1 is the highest among all the benchmarks; its daily score violations are also the smallest. The above results validate the superiority of the proposed method B1 in the regulation capacity offering problem.

E. Building Temperatures and Operating Power of DCS

Fig. 11 gives all the buildings' indoor temperature deviations on a typical summer day, when DCS provides regulation capacity based on the proposed method B1. The daily ambient temperature fluctuates within 28∼32 °C. Meanwhile, buildings' set temperatures distribute in 20∼23 °C, and buildings' required comfortable temperature ranges are all set as ±1 °C, as the two blue lines show. When the DCS does not participate in the regulation market, buildings' indoor temperature deviations are generally small and close to zero. It can be seen from Fig. 11 that the proposed method B1 can guarantee all the buildings' temperature comforts within the required range. It is noted that the indoor temperature deviations in different buildings are similar at night (i.e., 22:00∼8:00), while they are diverse during the daytime. This is because uncertain people

This article has been accepted for publication in IEEE Transactions on Smart Grid. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSG.2022.3220732

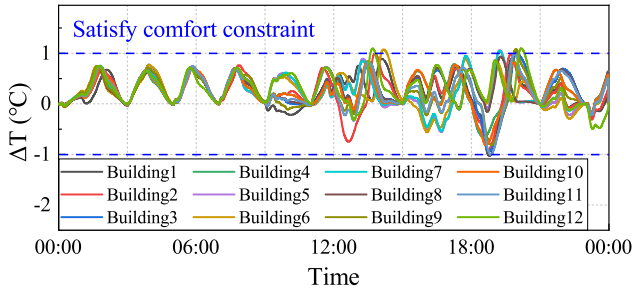


Fig. 11: Indoor temperature deviations based on strategy B1.

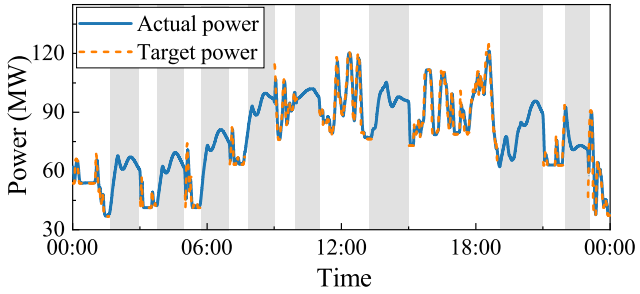


Fig. 12: The operating power of DCS based on strategy B1.

flows bring variational heat loads in the daytime, which causes a significant impact on indoor temperatures.

Fig. 12 shows DCS's operating power consumption during the service. In each hour, if DCS participates in the market and provides regulation capacity, there is a target power for DCS to follow the real-time signals. By contrast, if DCS does not provide regulation services, there is no target power for DCS to follow and it will try to maintain the building's set indoor temperatures. It can be seen from Fig. 12 that after providing the regulation service for a period, the proposed strategy B1 tends to wait for some time. As a result, the DCS participates in the market for only 15 hours, and the longest continuous time to provide regulation service is 4 hours. This is because the buildings' total thermal inertia is limited and it may get fully utilized after a couple of hours' continuous regulation. Fig. 12 also shows that the DCS's actual power can follow the target power well. It proves that the DCS can provide high-quality regulation services with high-performance scores.

From the above experiments, we can obtain several operating insights for the DCS operation when it provides regulation services to the market. First, DCS indeed possesses the flexibility to participate in the market and it can provide approximately 20% of its power as the capacity offering. Then, to obtain more revenue, DCS tends not to provide regulation services for a continuous long time. Finally, the capacity offering in the night (i.e., 22:00-8:00) is more stable than that in the daytime (i.e., 8:00-22:00), and it may probably be easier to determine the night capacity.

F. Discussion on Additional Cost for Regulation Service

It is possible that controlling DCS to provide regulation services may lead to additional costs, e.g., extra power consumption and hardware degradation, which are not considered in our objective Eq. (14). In this subsection, we conduct primary additional analysis to show that these losses could be negligible compared with the regulation revenue.

Based on the DCS operation data in one week, we compare the power consumption results under two different modes in Table II, i.e., the mode with regulation service and without regulation service. It can be seen that, the maximum deviation of the power consumption is around ± 65 MWh. Compared with the daily power consumption, the maximum deviation proportion is only 4.30%, which is quite small. Furthermore, considering the power consumption differences could be positive or negative values in different days, the average power consumption difference during a week is only -13.2 MWh (i.e., only 0.7% compared with the average daily power consumption). The main reason is that the DCS operating power always fluctuates around the original baseline power curve to maintain the required comfortable indoor temperatures. Hence, the additional power consumption cost is negligible.

In addition, the compressor is the most important component in a DCS. When the DCS provides regulation services, both the inbuilt regulation function and time interval of the compressor are the same with the DCS's original operating mode. The only difference is the regulation target, which adds a new power target to reflect the gap between the actual and promised power. So the adjustment signals for the DCS compressor are still sent within the same time interval, i.e., the regulation times of the compressor do not change. Moreover, both the system time delay and ramp rate are considered to limit the frequent adjustment of DCS power. This can ensure that the power is always adjusted within the designed acceptable range. Therefore, we assume that providing regulation service has a negligible impact on a DCS lifetime in this paper. Nevertheless, the proposed method can be readily extended to consider this additional operational cost if necessary, which could be a future research direction.

V. Conclusion

This paper proposes an hour-ahead capacity offering strategy for DCS to provide regulation services. To cope with the uncertain market signals and buildings' random cooling demands, the problem is formulated as a MDP. Then, to address the challenge from the complex physical model, we adopt a model-free DRL algorithm, i.e., DDPG, to seek the effective policy of the MDP. Furthermore, to decrease the probability of aggressive explorations in the training process, an intrinsic-motivated method based on the pseudo-count is proposed to improve the training efficiency. The proposed intrinsic motivated DRL framework can combine with different DRL algorithms to improve the agent's training efficiency, and be applied in other decision-making problems in power systems, such as the dispatch of electric vehicles, and batteries.

Numerical results show that the well-trained agent can strategically offer DCS's regulation capacities properly to maximize the total cumulative revenue, while satisfying the performance requirement from the market and the comfort requirement from buildings. The proposed method improves the training efficiency significantly and is "safer" compared with traditional DRL methods, because it can obtain a same

TABLE II: The comparison of DCS power consumption with and without providing regulation services

Days	Power consumption without regulation	Power consumption with regulation	Power consumption differences
Day 01	1508.64 MWh	1443.84 MWh	-64.8 MWh (-4.30%)
Day 02	1651.20 MWh	1682.88 MWh	-31.68 MWh(1.92%)
Day 03	71.65 MWh	72.80 MWh	1.15 MWh(1.61%)
Day 04	1719.6 MWh	1886.16 MWh	-65.28 MWh(-3.35%)
Day 05	1189.2 MWh	1225.44 MWh	36.24 MWh (3.05%)
Day 06	1560.48 MWh	1439.04 MWh	-49.92 MWh (-3.32%)
Day 07	1658.88 MWh	1649.76 MWh	-9.12 MWh (-0.55%)
Average	**1595.28 MWh**	**1582.08 MWh**	**-13.2 MWh(-0.70%)**

or better controller through less constraint violations. However, the proposed method can only decrease the constraint violation by improving the training efficiency, which is suitable for problems with soft constraints. It cannot achieve the strict safety of the hard constraint that requires zero violations, which will be our future work.

Although all the experiments are implemented in the simulation environment, it can still prove that the proposed method outperforms the traditional ones in the same environment. Considering the gap between the simulation and reality, how to deploy the DRL controller to combine with the real-world system will be our next work.

REFERENCES

[1] A. Q. Al-Shetwi, M. Hannan, K. P. Jern, M. Mansur, and T. Mahlia, "Grid-connected renewable energy sources: Review of the recent integration requirements and control methods," *J. Clean. Prod.*, vol. 253, p. 119831, 2020.

[2] B. Mohandes, M. S. E. Moursi, N. Hatziargyriou, and S. E. Khatib, "A review of power system flexibility with high penetration of renewables," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 3140–3155, 2019.

[3] X. Zhang, M. Pipattanasomporn, T. Chen, and S. Rahman, "An IoT-based thermal model learning framework for smart buildings," *IEEE Internet Things J.*, vol. 7, pp. 518–527, Nov. 2019.

[4] M. Cai, M. Pipattanasomporn, and S. Rahman, "Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques," *Appl. Energy*, vol. 236, pp. 1078–1088, 2019.

[5] Y. Xiao, Q. Su, F. S. S. Bresler, R. Carroll, J. R. Schmitt, and M. Olaleye, "Performance-based regulation model in pjm wholesale markets," in *2014 IEEE PES General Meeting*, pp. 1–5, 2014.

[6] Y. Li, Y. Rezgui, and H. Zhu, "District heating and cooling optimization and enhancement – towards integration of renewables, storage and smart grid," *Renew. Sust. Energ. Rev.*, vol. 72, pp. 281–294, 2017.

[7] L. Fabietti, T. T. Gorecki, F. A. Qureshi, A. Bitlislioğlu, I. Lymperopoulos, and C. N. Jones, "Experimental implementation of frequency regulation services using commercial buildings," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1657–1666, 2018.

[8] S. Buffa, M. H. Fouladfar, G. Franchini, I. Lozano Gabarre, and M. Andrés Chicote, "Advanced control and fault detection strategies for district heating and cooling systems—a review," *Appl. Sciences*, vol. 11, no. 1, p. 455, 2021.

[9] S. J. Cox, D. Kim, H. Cho, and P. Mago, "Real time optimal control of district cooling system with thermal energy storage using neural networks," *Appl. Energy*, vol. 238, pp. 466–480, 2019.

[10] M. Sameti and F. Haghighat, "Hybrid solar and heat-driven district cooling system: Optimal integration and control strategy," *Solar Energy*, vol. 183, pp. 260–275, 2019.

[11] L. Hao, M. Wei, F. Xu, X. Yang, J. Meng, P. Song, and Y. Min, "Study of operation strategies for integrating ice-storage district cooling systems into power dispatch for large-scale hydropower utilization," *Appl. Energy*, vol. 261, p. 114477, 2020.

[12] J. Hong, H. Hui, H. Zhang, N. Dai, and Y. Song, "Distributed control of large-scale inverter air conditioners for providing operating reserve based on consensus with nonlinear protocol," *IEEE Int. Things J.*, 2022. Early Access.

[13] H. Hui, P. Yu, H. Zhang, N. Dai, W. Jiang, and Y. Song, "Regulation capacity evaluation of large-scale residential air conditioners for improving flexibility of urban power systems," *Int. J. Electr. Power Energy Syst.*, vol. 142, p. 108269, 2022.

[14] D. Xie, H. Hui, Y. Ding, and Z. Lin, "Operating reserve capacity evaluation of aggregated heterogeneous tcls with price signals," *Appl. Energy*, vol. 216, pp. 338–347, 2018.

[15] X. Li, W. Li, R. Zhang, T. Jiang, H. Chen, and G. Li, "Collaborative scheduling and flexibility assessment of integrated electricity and district heating systems utilizing thermal inertia of district heating network and aggregated buildings," *Appl. Energy*, vol. 258, p. 114021, 2020.

[16] H. Hui, P. Siano, Y. Ding, P. Yu, Y. Song, H. Zhang, and N. Dai, "A transactive energy framework for inverter-based HVAC loads in a real-time local electricity market considering distributed energy resources," *IEEE Trans. Ind. Informat.*, 2022. Early Access.

[17] J. Cai and J. E. Braun, "A regulation capacity reset strategy for hvac frequency regulation control," *Energy Build.*, vol. 185, pp. 272–286, 2019.

[18] M. B. Anwar and M. O'Malley, "Strategic participation of residential thermal demand response in energy and capacity markets," *IEEE Trans. Smart Grid*, pp. 1–1, 2021.

[19] G. Chen, B. Yan, H. Zhang, D. Zhang, and Y. Song, "Time-efficient strategic power dispatch for district cooling systems considering the spatial-temporal evolution of cooling load uncertainties," *CSEE Journal of Power and Energy Systems*, 2021.

[20] N. Lu, "An evaluation of the hvac load potential for providing load balancing service," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1263–1270, 2012.

[21] X. Yang, J. Liu, J. Zhou, Y. Du, M. Wang, and Y. Chen, "Evaluation of air-conditioning load adjustability based on load plasticity," in *Journal of Physics: Conference Series*, vol. 1920, p. 012055, IOP Publishing, 2021.

[22] P. Wang, D. Wu, and K. Kalsi, "Flexibility estimation and control of thermostatically controlled loads with lock time for regulation service," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3221–3230, 2020.

[23] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2019.

[24] T. Chen, Q. Cui, C. Gao, Q. Hu, K. Lai, J. Yang, R. Lyu, H. Zhang, and J. Zhang, "Optimal demand response strategy of commercial building-based virtual power plant using reinforcement learning," *IET Gener. Transm. Distrib.*, Apr. 2021.

[25] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, "Multi-agent deep reinforcement learning for hvac control in commercial buildings," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 407–419, 2021.

[26] S. Çalışır and M. K. Pehlivanoğlu, "Model-free reinforcement learning algorithms: A survey," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2019.

[27] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2020.

[28] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement learning*, pp. 45–73, Springer, 2012.

[29] PJM Manual 11: Energy Ancillary Services Market Operations (Revision: 177). [Online]. Available: https://www.pjm.com/-/media/documents/manuals/m11.ashx. [Accessed: Nov. 01, 2021].

[30] PJM Manual 12: Balancing Operations (Revision: 45). [Online]. Available: https://www.pjm.com/ /media/documents/manuals/m12.ashx. [Accessed: March. 23, 2022].

[31] A. Potapov and M. Ali, "Convergence of reinforcement learning algorithms and acceleration of learning," *Physical Review E*, vol. 67, no. 2, p. 026706, 2003.

[32] B. Bonet and H. Geffner, "Learning depth-first search: A unified approach to heuristic search in deterministic and non-deterministic settings, and its application to mdps.," in *ICAPS*, vol. 6, pp. 142–151, 2006.

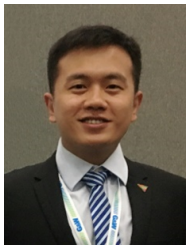
[33] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[34] I. Menache, S. Mannor, and N. Shimkin, "Basis function adaptation in temporal difference reinforcement learning," *Annals of Operations Research*, vol. 134, no. 1, pp. 215–238, 2005.
[35] C. Dimitrakakis and R. Ortner, "Decision making under uncertainty and reinforcement learning," 2018.
[36] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," *Advances in neural information processing systems*, vol. 29, 2016.
[37] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for markov decision processes," *J. Comp. Syst. Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.
[38] K. P. Murphy, "A probabilistic perspective," *Text book*, 2012.
[39] G. Xuan, W. Zhang, and P. Chai, "Em algorithms of gaussian mixture model and hidden markov model," in *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, vol. 1, pp. 145–148, IEEE, 2001.
[40] New District Hengqin: District Cooling and Heating System Technical Guidelines (the 4th Edition), Zhuhai, China, Mar. 2016.
[41] G. Ostrovski, M. G. Bellemare, A. Oord, and R. Munos, "Count-based exploration with neural density models," in *International conference on machine learning*, pp. 2721–2730, PMLR, 2017.
[42] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine learning*, vol. 38, no. 3, pp. 287–308, 2000.
[43] Z. Zhu, K. Lin, and J. Zhou, "Transfer learning in deep reinforcement learning: A survey," *arXiv preprint arXiv:2009.07888*, 2020.

Peipei Yu (S'21) received the M.S and B.S. degrees in mathematics from Zhejiang University, Zhejiang, China, in 2019 and 2016, respectively. She is currently working toward the Ph.D. degree at University of Macau, Macao, China. Her research interests include Internet of Things for smart energy, demand response, and reinforcement learning control.

Hongcai Zhang (S'14–M'18) received the B.S. and Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2013 and 2018, respectively. He is currently an Assistant Professor with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macao, China. In 2018-2019, he was a postdoctoral scholar with the Energy, Controls, and Applications Lab at University of California, Berkeley, where he also worked as a visiting student researcher in 2016. His current research interests include Internet of Things for smart energy, optimal operation and optimization of power and transportation systems, and grid integration of distributed energy resources.

Yonghua Song (F'08) received the B.E. and Ph.D. degrees from the Chengdu University of Science and Technology, Chengdu, China, and the China Electric Power Research Institute, Beijing, China, in 1984 and 1989, respectively, all in electrical engineering. He was awarded DSc by Brunel University in 2002, Honorary DEng by University of Bath in 2014 and Honorary DSc by University of Edinburgh in 2019. From 1989 to 1991, he was a Post-Doctoral Fellow at Tsinghua University, Beijing. He then held various positions at Bristol University, Bristol, U.K.; Bath University, Bath, U.K.; and John Moores University, Liverpool, U.K., from 1991 to 1996. In 1997, he was a Professor of Power Systems at Brunel University, where he was a Pro-Vice Chancellor for Graduate Studies since 2004. In 2007, he took up a Pro-Vice Chancellorship and Professorship of Electrical Engineering at the University of Liverpool, Liverpool. In 2009, he joined Tsinghua University as a Professor of Electrical Engineering and an Assistant President and the Deputy Director of the Laboratory of Low-Carbon Energy. During 2012 to 2017, he worked as the Executive Vice President of Zhejiang University, as well as Founding Dean of the International Campus and Professor of Electrical Engineering and Higher Education of the University. Since 2018, he became Rector of the University of Macau and the director of the State Key Laboratory of Internet of Things for Smart City. His current research interests include smart grid, electricity economics, and operation and control of power systems. Prof. Song was elected as the Vice-President of Chinese Society for Electrical Engineering (CSEE) and appointed as the Chairman of the International Affairs Committee of the CSEE in 2009. In 2004, he was elected as a Fellow of the Royal Academy of Engineering, U.K. In 2019, he was elected as a Foreign Member of the Academia Europaea.



Hongxun Hui (S'17–M'20) received both the Ph.D. and B. Eng degrees in electrical engineering from Zhejiang University in 2020 and 2015, respectively. He is currently a Post-doctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. His research interests include modelling and optimal control of demand side resources in smart grid, the electricity market considering demand response, and the uncertainty analysis brought by flexible loads and renewable energies.

Chao Huang (M'17) received the B.Eng. degree in electrical engineering and automation from the Harbin Institute of Technology, China, in 2011, the M.S. degree in intelligent transport system from the University of Technology of Compiegne, France, in 2013, and the Ph.D. degree in systems engineering and engineering management from the City University of Hong Kong, Hong Kong, in 2017. From 2019 to 2021, he was a Postdoctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. He is currently an Associate Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. His research interests include data mining, computational intelligence, and energy informatics.