Received 12 April 2020; revised 25 July 2020; accepted 2 September 2020. Date of publication 14 September 2020;
date of current version 8 October 2020.

Digital Object Identifier 10.1109/OAJPE.2020.3023916

Reinforcement Learning for Building Energy Optimization Through Controlling of Central HVAC System

JUN HAO[1] (Student Member, IEEE), DAVID WENZHONG GAO[1] (Senior Member, IEEE), AND JUN JASON ZHANG[2] (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80210 USA
[2]School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China

CORRESPONDING AUTHOR: D. W. GAO (wenzhong.gao@du.edu)

This work was supported in part by the U.S. National Science Foundation under Grant 1711951, in part by the Key Technologies Research and Development Program of Tianjin City under Grant 17ZXRGGX00170, in part by the Corporate Research and Development Program of State Grid of China under Grant SGZJDK00DYJS1900232, in part by the Science and Technology Program of Central China Branch of State Grid Corporation of China under Grant 521400180005, and in part by the National Key Research and Development Program of China under Grant SQ2018AAA010127.

ABSTRACT This paper presents a novel methodology to control HVAC system and minimize energy cost on the premise of satisfying power system constraints. A multi-agent architecture based on game theory and reinforcement learning is developed so as to reduce the cost and computational complexity of the microgrid. The multi-agent architecture comprising agents, state variables, action variables, reward function and cost game is formulated. The paper fills the gap between multi-agent HVAC systems control and power system optimization and planning. The results and analysis indicate that the proposed algorithm is beneficial to deal with the problem of "curse of dimensionality" for multi-agent microgrid HVAC system control and speed up learning of unknown power system conditions.

INDEX TERMS Game theory, reinforcement learning, multi-agent system, HVAC control, cost minimization.

I. INTRODUCTION

ABOUT 30% – 40% of global energy is used by buildings [1]. Within this sector, academic and commercial buildings are accounted for the highest energy consumption [2], [3]. Heating, ventilation, and air conditioning systems (HVAC) are considerably equipped in those facilities and use tremendous amount of electricity [4]. Meanwhile, HVAC system can also impact the indoor working performance. Thus, the control and management of HVAC will be one of the crucial elements of future microgrid [5], [6]. From management perspective, the inspection and maintenance work for central HVAC system will be scheduled at owner's decision-making. From power system aspects, future distribution power grid will be incorporated with more variable renewable energy. As a result, the uncertainties and complexities of power grid condition will be increased, making the traditional centralized optimization and management algorithms computationally expensive and infeasible. The existing mutil-agent HVAC systems control methodologies focus mainly on the building side [7]. This paper fills the gap between optimizing of HVAC systems and planning of distribution power system by proposing an advanced algorithm based on game theory and multi-agent reinforcement learning.

In recent years, the applications of artificial intelligence and multi-agent system to HVAC system and power systems have been investigated. A review about the framework, approaches, concepts and potential value of MAS technology to power industry was presented in [8], [9]. It should be noted that multi-agent system (MAS) is a good solution to deal with super complex, diverse and scattered problem that may occur in the scenarios of HVAC system control and management. Various multi-agent methodologies have already been developed in the literature to enhance the control and autonomous operations of microgrids [9]–[11]. Among the existing multi-agent solutions for microgrids, two aspects are neglected so that the HVAC system's optimization and management become inaccurate and less efficient. The first aspect is that the current methodologies are not compatible with individually and separately controlled HVAC systems taking influences of power systems into consideration [7]. Also most of the approaches are based on centralized framework and require the exchange of all kinds of information through local network [12]. In this scenario, the communication network would require large investment and become complex. Therefore, a multi-agent approach with less communication among agents will be needed for power system planning and HVAC system control and management. The second aspect is the efficiency of the policy exploration. As illustrated in other articles [13], agents' actions are based on the negotiation and communication within the same microgrid. While in reality, the agents are in a dynamic and ever-changing power grid. Take the building manager choosing HVAC system setting as an example, the aftereffect of the action may be unpredictable due to the chain effect of distributional locational marginal pricing (DLMP) and other players' decisions. Therefore, an effective and efficient policy exploration approach is vital for agents in a microgrid.

Implementing machine learning or artificial intelligence into HVAC control system is an effective way to enable the controllers with the ability to learn and improve their decision-making. Reinforcement Learning (RL) was implemented in a wide range of power system economic problems [14]–[16], which proves the capability and potential of RL. In this paper, we propose a novel multi-agent reinforcement learning algorithm to optimize the control of HVAC and planning of power system. Each agent controls one central HVAC system in one of the buildings. Our proposed learning mechanism can study and analyze the relationship between independent HVAC controlling systems. Agents are greedy and tend to maximize their profits within the power system constrains. The proposed approach allows to update and enhance the knowledge about the best actions for HVAC system control and scheduling under different weather conditions. The proposed framework enables the system to learn from the stochastic power grid, weather and human activity data and the approach can also exploit the historical experiences to suggest the optimal HVAC system settings.

The development and testing of our proposed methodology relies on models of the energy system that properly account for multi-level dynamic behavior. The aim for our research is to establish a useful and cost-saving mechanism of HVAC scheduling in ever-changing distribution power grid and environments. This paper is the extension and further research of [17]. The paper is organized as the following: the analytical models that simulate the environment and the objective function are introduced in Sec II. In Sec III, a multi-agent game is defined to find the optimal control strategy for each player. Sec IV demonstrates our proposed multi-agent RL to simplify the multi-agent game's computational complexity. Experimental results are presented in Sec V to compare the systems with or without multi-agent RL. The conclusion is made in Sec VI. In this paper, the term ''reward'' and ''payoff'' will be used interchangeably.

II. ESTABLISHMENT OF COMPUTING SYSTEM

In this section, we introduce several analytical models to create and simulate the learning environment for the proposed multi-agent game and multi-agent RL. Those models will help us to measure the loss of working efficiency, and the cost of energy in monetary values.

A. HUMAN WORK PERFORMANCE MODEL

Through controlling the temperature, HVAC system can affect the indoor personnels' working efficiency [18], [19]. Temperature is the crucial feature of indoor environment. Unsuccessful control of temperature can result in low efficiency, sick building symptoms, etc. In this paper, the working productivity ξ is referred to express the correlations between working efficiency and monetary value [20], and the model can be expressed as

$$\begin{aligned}\xi &= g(T_{in})\\ &= 0.1647524 \cdot (\frac{5}{9} \cdot (T_{in} - 32))\\ &\quad - 0.0058274 \cdot (\frac{5}{9} \cdot (T_{in} - 32))^2\\ &\quad + 0.0000623 \cdot (\frac{5}{9} \cdot (T_{in} - 32))^3 - 0.4685328 \quad (1)\end{aligned}$$

where ξ denotes the working efficiency which is determined by inside temperature T_{in}. The temperature setting should be among the bracket $T_l \leqslant T_{in} \leqslant T_u$, where T_l and T_u are the lower and upper limits, respectively. It should be pointed out that, although according to [21], the ideal temperature range for university buildings is between 68° F and 74° F, the temperature settings in our study is relaxed to 64° F and 79° F for research purposes.

For simplicity representation, $\xi_{k,t}$ and $x_{k,t}$ are the work efficiency and indoor temperature in building k at time t, respectively, (1) can be formulated as

$$\xi_{k,t} = g(x_{k,t}). \quad (2)$$

$\mathbf{x}_t$ is the control variable in this paper.

B. THE NEURAL NETWORK BASED ENERGY CONSUMPTION PROFILE MODELS

To build the learning environment for RL, the other crucial model is to simulate the HVAC system's energy consumption corresponding to buildings' indoor temperatures. In this manuscript, eQUEST is used [22] to generate the dataset to train the neural network models that predict the HVAC systems' consumption. eQUEST can provide all-inclusive simulations about central HVAC systems and appropriate assumptions for lighting and plug-in loads according to the size and type of the simulated buildings. The energy consumption calculated by eQUEST is based on various factors including outdoor temperature, story, architecture, etc. Those various factors can be used as key features to train the neural network models. The hourly report from eQUEST can provide enough data for training and testing the neural network

TABLE 1. **Partial simulation results of Ritchie center on July** 1*st* **from** 15 : 00 **to** 20 : 00**.**

Hour	T$_{in}$ (°F)	T$_{out}$ (°F)	Energy (BTU)
15	64	94	1.08×10^7
16	64	92	1.31×10^7
17	64	92	1.70×10^7
18	64	93	2.03×10^7
19	64	89	2.22×10^7
20	64	85	2.19×10^7

models (NNM). And those NNMs will be used to forecast the power consumption for each building in same distribution power grid. Table 1 shows selected simulation results of the hourly energy consumption on July 1*st* 2016 at the Ritchie center, the recreation center at University of Denver (DU). Hour column indicates the time of one day, T$_{in}$ and T$_{out}$ are the building temperature setting and the outside dry-bulb temperature, respectively. Energy indicates the energy consumed by the building corresponding to the temperature setting. Apparently, the increase and decrease of power usage would influence the utility cost.

One year simulated data corresponding to every indoor temperature setting within the control bracket $T_l \leqslant T_{in} \leqslant T_u$ is generated for those buildings in our tested campus power grid. A three-layer feed-forward neutral network with sigmoid activation function is trained for every bus in the power grid. There are 10 neurons in hidden layer. Levenberg-Marquardt backpropagation algorithm [23] is implemented to train those models. Neural network can be easily generalized to train energy prediction models for buildings with different sizes. Neural network can also handle unbalanced weather data compared with other machine learning techniques like SVM. The inputs are time, indoor temperature, and outdoor temperature, and the output is energy consumption. In this paper, the dataset are randomly divided into three sections for training, validation and testing purposes. Each section takes up 75%, 10% and 15% of dataset, respectively. After testing, the prediction models' R-values are all above 0.92, which are good enough to build the learning environment for our proposed RL algorithm. We denote $e_{k,t}$ as the energy consumptions of building k at time t and can be expressed as,

$$\begin{aligned} e_{k,t} &= h_k(x_{k,t}, t, T_{out,t}) \\ \mathbf{e}_t &= H(\mathbf{x}_t, t, T_{out,t}) \end{aligned} \tag{3}$$

where $e_{k,t}$ is a function of indoor temperature $x_{k,t}$, time t and outdoor temperature $T_{out,t}$, $\mathbf{e}_t = [e_{1,t}\ e_{2,t} \cdots e_{n,t}]^\intercal$.

C. DISTRIBUTION LOCATIONAL MARGINAL PRICING FOR DU CAMPUS GRID

In the previous sections, we have already introduced two kinds of analytical models that help to build the learning environment. In this section, the last component of the learning environment will be addressed.

We implement the distribution locational marginal pricing model [24], [25] within the University of Denver campus power grid to not only calculate the energy cost but also to reflect the state transitions and the corresponding rewards when HVAC setting is changed.

$$\underset{p_j^b, p_i^g}{\arg\max} s = \sum_{j=1}^{N} (c_j - p_j^b) \cdot q_{c_j} - \sum_{i=1}^{M} (p_i^g - u_i) \cdot q_{u_i}$$

$$\text{s.t.} \sum_{i=1}^{M} q_{u_i} - \sum_{j=1}^{N} q_{c_j} - L_P(V, \theta) = 0 \tag{4}$$

$$\sum_{i=1}^{M} Q_{u_i} - \sum_{j=1}^{N} Q_{c_j} - L_Q(V, \theta) = 0 \tag{5}$$

$$f_j(V, \theta) \leqslant f_j^{Max} \tag{6}$$

$$q_{u_i}^{MIN} \leqslant q_{u_i} \leqslant q_{u_i}^{MAX} \tag{7}$$

$$Q_{u_i}^{MIN} \leqslant Q_{u_i} \leqslant Q_{u_i}^{MAX} \tag{8}$$

$$V_i^{MIN} \leqslant V_i \leqslant V_i^{MAX} \tag{9}$$

where s denotes system social surplus that is obtained from our DLMP calculation, N is the number of buildings in smart grid and j is the building index; M is the total number of electricity suppliers and i is the generator index; c_j stands for the building bid price for each power generation and u_i represents the offer price from each power generation; p_j^b is the distribution locational marginal price at each building j, and p_i^g stands for the distribution locational marginal price at supply bus i; q_{c_j} is the power demand at building j; q_{u_i} is the power supply from bus i; V and θ are voltage magnitude and angle at each bus, respectively; f_j stands for the power flow at jth line, which is limited by f_j^{max} A; q_{u_i} is the active power output from each power source and the maximum capacity $q_{u_i}^{MAX}$ MW, while Q_{u_i} is the reactive power output from the corresponding energy generation and the maximum capacity $Q_{u_i}^{MAX}$ MVar; V_i stands for the voltage magnitude of the ith bus with power injection, in this case study V_i^{MIN} pu and V_i^{MAX} pu; and $L_P(V, \theta)$ and $L_Q(V, \theta)$ are the total active power loss and reactive power loss in the smart gird, respectively. To simplify the formulation, we denote the DLMPs $\mathbf{p}_t = [p_{1,t}\ \ p_{2,t} \ldots p_{n,t}]^\intercal$ as a nonlinear function $\Gamma(\cdot)$ of energy consumption $\mathbf{e}_t$ as

$$\mathbf{p}_t = \Gamma(\mathbf{e}_t) \tag{10}$$

And according to (3), $\Gamma(\cdot)$ can be expanded as a function of the control variable $\mathbf{x}_t$

$$\mathbf{p}_t = \Gamma(H(\mathbf{x}_t, t, T_{out,t})). \tag{11}$$

D. OVERALL MULTI-AGENT PAYOFF

With all the analytical models that have been discussed in this paper, the learning environment can finally be modeled for the proposed MARL algorithm. After taking an action, the rewards that agents receive at the state transition comprise two parts: the energy cost, calculated by the production of the end-use energy and the corresponding DLMP; the reduction of work productivity, calculated by the production of the amount of occupants and the cost of efficiency reduction per person [17]. The state is under the influence of the condition of the power system, the working efficiency, the amount of indoor personnels and the energy consumption. Taking an action would definitely transit into a new state. However, agent's state may be affected by other agents' actions through chain effects in power systems such as DLMP fluctuation, line congestion, etc.

(12) defines the rewards ψ_t at time t of the agent in distribution power grid. Because of the DLMPs, occupants, and energy usages, every agent has its own reward function and the state is different though the formulation for every single agent looks the same.

$$\begin{aligned}\psi_t &= \sum_{k=1}^{n}[p_{k,t} \cdot e_{k,t} + w \cdot \alpha(1 - \xi_{k,t}) \cdot o_{k,t}] \qquad (12)\\ &= \mathbf{p}_t \cdot \mathbf{e}_t + w \cdot \alpha(\mathbf{1} - \boldsymbol{\xi}_t) \cdot \mathbf{o}_t\\ &= \Gamma[H(\mathbf{x}_t, t, T_{out,t})] \cdot H(\mathbf{x}_t, t, T_{out,t})\\ &\quad + w \cdot \alpha[1 - g(\mathbf{x}_t)] \cdot \mathbf{o}_t\\ &= \Psi(\mathbf{x}_t, t, T_{out,t}, \mathbf{o}_t) \qquad (13)\end{aligned}$$

where ψ_t is the overall cost at time t, w is the weight for the efficiency component which is set to be 0.1 in this paper, α is the hourly saving for each personnel when the working productivity is 1, and $\mathbf{o}_t = [o_{1,t}\ o_{2,t} \cdots o_{n,t}]^\intercal$ where $o_{k,t}$ is the number of occupants in building k at time t.

It should be noted that, $T_{out,t}$ can be acquired from weather predictions and the number of occupants $\mathbf{o}_t$ is estimated by the schedules of buildings. Therefore, t, $T_{out,t}$ and $\mathbf{o}_t$ are not unknown variables at a give time. The overall payoff is decided by the indoor temperature settings $\mathbf{x}_t$. The objective is to optimize indoor temperature setting $\hat{\mathbf{x}}_t$ at time $\mathbf{t}$. $\mathbf{t}$ is a vector of 24 hours. Therefore, the problem can be expressed as

$$\begin{aligned}\hat{\mathbf{x}}_t = \ &\arg\min_{\mathbf{x}_t} \Psi(\mathbf{x}_t)\\ &\text{s.t. } T_l \leqslant x_{k,t} \leqslant T_u \qquad (14)\end{aligned}$$

III. MULTI-AGENT COST GAME

Based on the analytical models in Sec.II, the payoff for each indoor HVAC system setting can easily be calculated at any time step. The cost game solves a finite N person game. The multi-agent game is designated to select the optimal strategy in order to maximize the payoff within a certain smart grid. The players are the buildings' managers that represent their own benefits. The number of strategies for each player is finite, which is 16 in our paper.

A N person, finite non co-operative multi-agent game can be formulated as the following [17]:

$$\psi(x_{i,t}) = (\mathbf{N}, \{\mathbf{S}^{i,t}\}_{i \in N}, \{\delta^{i,t}\}_{i \in N}) \qquad (15)$$

where $\mathbf{N}$ is the number of players at time t, $\mathbf{S}^i$ is the i_{th} players' finite pure strategy set at time t and δ^i is the payoff set corresponding to various pure strategies at time t. If the game has an optimal solution, there is at least one Nash equilibrium (NE), which means that one player could not get a better payoff than the optimal strategy at NE if the game reaches a Nash equilibrium and the other players are playing according to their Nash equilibrium strategies.

$$\delta^{i,t}(\alpha^*) \geqslant \delta_j^{i,t}(\alpha^{*-i}, \alpha^i), \quad \forall i \in N, \ \forall \alpha^i \in \Sigma^i \qquad (16)$$

where $*$ means the best response strategy at NE, $-i$ denotes the players other than i_{th} player, and Σ^i is the mixed strategy space for player i. By assuming ψ^i is the optimal HVAC system setting for player i, then we can transfer the game problem into an nonlinear optimization problem for the each player in smart grid.

$$\begin{aligned}&(\psi^{i,t})\\ &\min \gamma^{i,t} - \delta^{i,t}(\alpha) \qquad (17)\\ &\text{s.t. } \delta^{i,t}(\alpha^{-i,t}, s_j^{i,t}) - \gamma^{i,t} \leqslant 0, \quad \forall j = 1, \ldots, m^i \qquad (18)\\ &\sum_{j=1}^{m^i} \alpha_j^{i,t} = 1 \qquad (19)\\ &\alpha_j^{i,t} \geqslant 0 \quad \forall j = 1, \ldots, m^i \qquad (20)\end{aligned}$$

where $\gamma^{i,t}$ is assumed as the optimal payoff corresponding to the best response strategy, m^i is the number of strategies, $(\alpha^{-i,t}, s_j^{i,t})$ denotes the player i's strategies set while the others' strategy sets are expressed as $\alpha^{-i,t}$ at time t. According to [26], after applying KKT condition, we can obtain that a Nash equilibrium of game (15) can be transformed into a problem of equalities and inequalities.

Lemma 1: A necessary and sufficient condition for game ψ to have a Nash equilibrium strategy set α is

$$\begin{aligned}&\gamma^{i,t} - \delta^{i,t}(\alpha) = 0 \quad \forall i \in N \qquad (21)\\ &\delta^{i,t}(\alpha^{-i,t}, s_j^{i,t}) - \gamma^{i,t} \leqslant 0, \qquad (22)\\ &\qquad \forall j = 1, \ldots, m^i, \quad \forall i \in N\\ &\sum_{j=1}^{m^i} \alpha_j^{i,t} = 1, \quad \forall i \in N \qquad (23)\\ &\alpha_j^{i,t} \geqslant 0 \quad \forall j = 1, \ldots, m^i, \ \forall i \in N \qquad (24)\end{aligned}$$

Form (21), we can obtain that for every player in the same smart grid their best response strategy is at Nash equilibrium. (22),(23),(24) are the equality and inequality constraints for optimization and (22) means that no mix strategy combination would result in better results than best response. Therefore, we can obtain that the optimal solution of nonlinear HVAC controlling problem is the strategy at Nash equilibrium α.

Theorem 1: A necessary and sufficient condition for α^* to a Nash equilibrium of game Ψ is that it is an optimal solution *of the following minimization problem*

$$
\begin{aligned}
&(\Psi)\\
&min \sum_{i\in N} \gamma^{i,t} - \delta^{i,t}(\alpha)\\
&s.t.\ \ \delta^{i,t}(\alpha^{-i,t}, s_j^{i,t}) - \gamma^{i,t} \leqslant 0,\\
&\qquad \forall j = 1, \dots, m^i, \quad \forall i \in N\\
&\qquad \sum_{j=1}^{m^i} \alpha_j^{i,t} = 1, \quad \forall i \in N\\
&\qquad \alpha_j^{i,t} \geqslant 0 \quad \forall j = 1, \dots, m^i,\ \forall i \in N
\end{aligned}
$$

The optimal value of $\min \sum_{i\in N} \gamma^{i,t} - \delta^{i,t}(\alpha)$ should be 0. The value of $\gamma^{i,t}$ at the optimal point gives the expected payoff of the player i at time t.

IV. MULTI-AGENT REINFORCEMENT LEARNING FOR ENERGY COST GAME

The multi-agent game depicted in Sec. III is constrained by the computational complexity. When the number of indoor temperature control strategy or the number of player increase, the computational complexity increases exponentially. This shortcoming makes the multi-agent game consume long time though modern computational capability of CPU or GPU is much better than the past. Hence, there is a need to implement another algorithm to simplify the game strategy set and adapt to the constantly changing environment. Even if the number of strategies or the number of players increase, the computational time should increase linearly. Therefore, we implement the Multi-Agent Reinforcement Learning into the multi-agent game to simplify the strategy set.

In our study, discounted reward is implemented to calculate the rewards when the actions (strategy) are taken by players. x_s denote the state of the system before the s^{th} transition. The discounted reward from state i can be defined as:

$$\zeta_i = \lim_{k\to\infty} E[\sum_{s=1}^{k} \tau^{s-1} r(x_s, \pi(x_s), x_{s+1}) \mid x_1 = i] \tag{25}$$

where τ denotes the discount factor, and $0 \leq \tau \leq 1$, an alternative expression of (25) is:

$$
\begin{aligned}
\zeta_i = E[r(x_1, \pi(x_1), x_2) + \tau r(x_2, \pi(x_2), x_3)\\
+\tau^2 r(x_3, \pi(x_3), x_4) + \cdots]
\end{aligned}
\tag{26}
$$

In (25) and (26), τ is used to discount the rewards for later actions, and τ can be expressed as a function of μ:

$$\tau = (\frac{1}{1+\mu})^{s-1} \tag{27}$$

where μ denotes the rate of interest. When $\mu > 0$, we can ensure that $0 \leq \tau \leq 1$. When μ approaches $+\infty$, τ approximately equals to 0. If μ equals 0, τ is 1. (25) can be formulated as:

$$\zeta_i = \lim_{k\to\infty} E[\sum_{s=1}^{k} (\frac{1}{1+\mu})^{s-1} r(x_s, \pi(x_s), x_{s+1}) \mid x_1 = i] \tag{28}$$

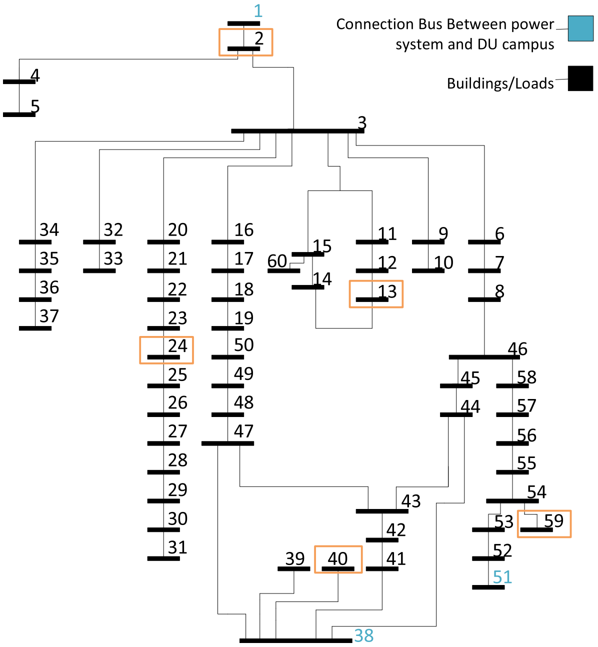

FIGURE 1. The network topology of DU campus grid.

With the definition of discounted rewards, we can calculate the reward of every action and implement a multi-agent RL algorithm that can help to reduce the computational complexity for the proposed game. The learning process of RL algorithm requires the updating of rewards every time the system transitions into a new state [27]. Like in other research that relates to RL algorithm, we define the constantly updating quantities as Q factors as well [28]. So $Q(i, a)$ is used to denote the reward quantity for state i and action a.

The reward that is calculated in the transition is denoted as feedback. The feedback is used is to update the Q-factors for the evaluation of actions (strategies) in the former state. Generally speaking if the value of a feedback is good, the Q-factor of that action is increased, otherwise, the the Q-factor of that action is decreased. Therefore, the system is analyzed and controlled in real time. In each state visited, some action is opted out and the system is ready to proceed to next state. The "state" in our context is the power system condition at the specific time when all the agents have decided their actions and start to consume new amount of energy. Since at a specific time point, the number of indoor occupant is fixed, the factors that affect the choice of HVAC setting are the energy consumption and the utility price. When the action is selected or changed, the DLMP will be influenced. Then, the system enters a new state.

In our study, we choose the discounted reward multi-agent RL for the proposed multi-agent game to reduce computational complexity [29]. The generalized steps for discounted reward multi-agent RL can be expressed as follows:

- Step 1 (Input and Initiation): Set the Q-factors to 0:

$$Q(i, a) \leftarrow 0, \quad \forall i, and, \ \forall a \in A(i) \tag{29}$$

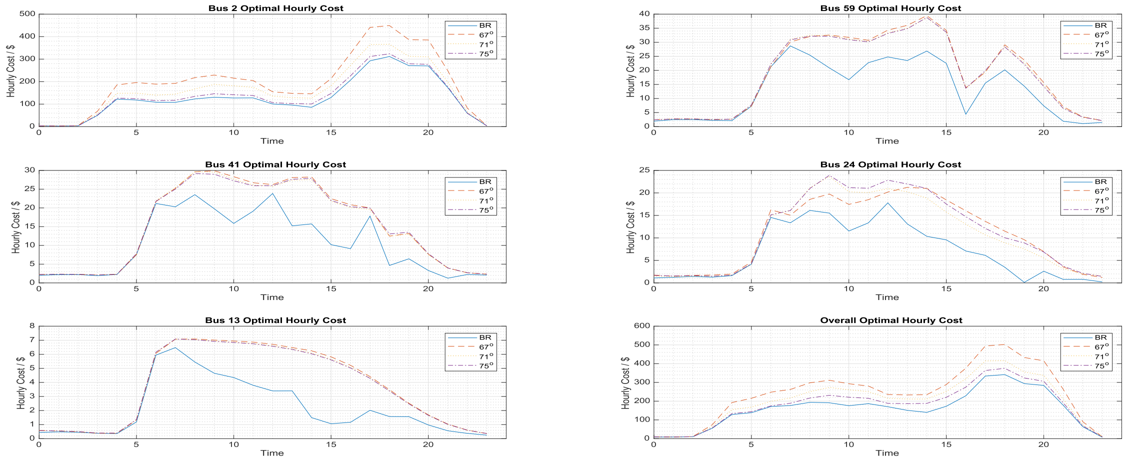

FIGURE 2. The summer day results comparison.

$A(i)$ denotes the set of actions in state i. In our case, the number of action equals to the number of strategy in the multi-agent energy game.

- Step 2 (Q-factor Update): Let $\mid A(i) \mid$ denote the number of actions in set $A(i)$. Hence, the probability of action a is selected in state i as $\frac{1}{|A(i)|}$. $r(i, a, j)$ denotes the transition reward. The algorithm for updating $Q(i, a)$ is defined as:

$$Q(i, a) \leftarrow (1 - \alpha^k)Q(i, a) + \alpha^k[r(i, a, j) + \tau \max_{b \in A(j)} Q(j, b)], \quad (30)$$

α^k defines the learning rate in the k^{th} iteration. Set $k = 1$ when transition to a new state. *Itmax* denotes the maximum number of iteration, and should be set to a large number. The computation of α^k will be discussed later. τ denotes the discount factor in multi-agent RL.

- Step 3 (Termination Check): Increase k by 1. Set $i \leftarrow j$, when $k < Itmax$, return to Step 1 if $Q(i, a)$ can be updated. Otherwise, proceed to Step 4.
- Step 4 (Outputs): For each state i, select the action a^* so that the corresponding $Q(i, a^*)$ achieves the optimal value.

The learning rate α^k should be positive value and satisfy $\alpha^k < 1$. Otherwise, the system would not converge or may even diverge. The learning rate for the discounted reward reinforcement learning is a function of k and have to meet the condition in [30]. In our research, the learning rate step size is expressed as:

$$\alpha^k = \frac{C}{D + k} \quad (31)$$

where $C = 90$ and $D = 100$ in our tests to ensure the system is stable.

Therefore, we can formulate our discounted reward multi-agent RL for multi-agent game as follows:

- Step 1 (Input and Initiation): According to (15) and (29), set the Q-factors to 0:

$$Q(i, s) \leftarrow 0, \forall i, and, \forall s \quad (32)$$

$S(i)$ denotes the set of strategy in game Ψ. For each agent, the number of action equals to the number of strategy in the proposed multi-agent game. α^k defines the learning rate in the k^{th} iteration. Set $k = 1$ when transition to a new state. *Itmax* denotes the maximum number of iteration, and should be set to a large number. In our research, the $Itmax = 10000$.

- Step 2 (Q-factor Update): Let $\mid S(i) \mid$ denotes the number of actions in set $S(i)$. Hence, the probability of strategy s is selected in state i as $\frac{1}{|S(i)|}$. $\delta(i, a, j)$ denotes the transition reward of the corresponding strategy. The algorithm for updating $Q(i, a)$ is defined as:

$$Q(i, a) \leftarrow (1 - \alpha^k)Q(i, a) + \alpha^k[\delta(i, s, j) + \tau \max_{b \in S(j)} Q(j, b)], \quad (33)$$

It should be noted that $\max_{b \in S(j)} Q(j, b)$ equals the optimal social cost $\gamma^{i,t}$ in game Ψ. Therefore, we can transform (33) into:

$$Q(i, a) \leftarrow (1 - \alpha^k)Q(i, a) + \alpha^k[\delta(i, s, j) + \tau\gamma(i)], \quad (34)$$

where $\gamma(i)$ denotes the optimal payoff for the multi-agent energy game.

- Step 3 (Termination Check): Increase k by 1. Set $i \leftarrow j$, when $k < Itmax$, then return to Step 1. Otherwise, proceed to Step 4.
- Step 4 (Outputs): For each state i, select the strategy s^* so that the corresponding $Q(i, a^*)$ achieves the optimal value.
- Pop up the best two strategy according to the optimal $Q(i, a^*)$ to play the energy game and get rewards from the game.

Reward function is individual to each agent. Different agents can receive different rewards for the same state transition. Instead,it keeps a vector of estimates, which give the future reward of action a_k, depending on the joint action a_{-k} played by the other agents. During learning, the agent selects

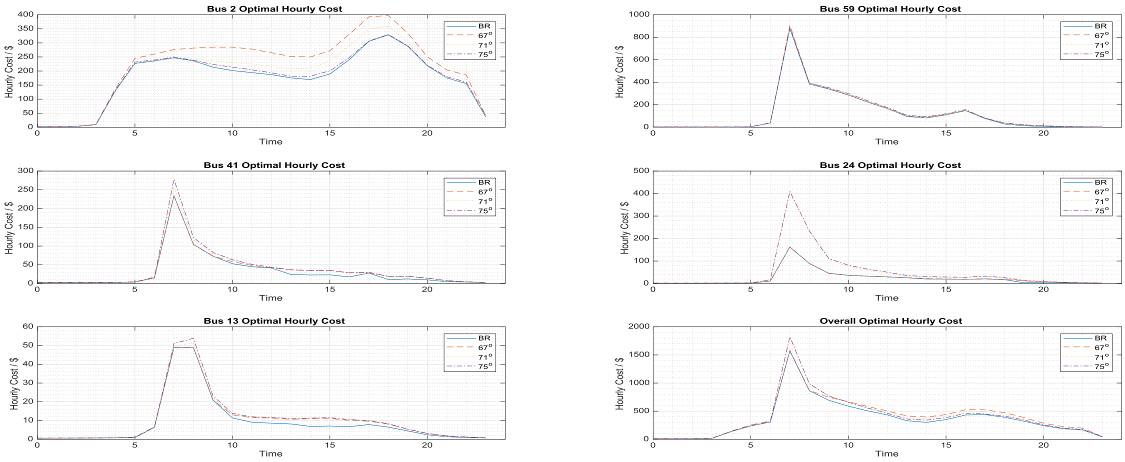

FIGURE 3. The winter day results comparison.

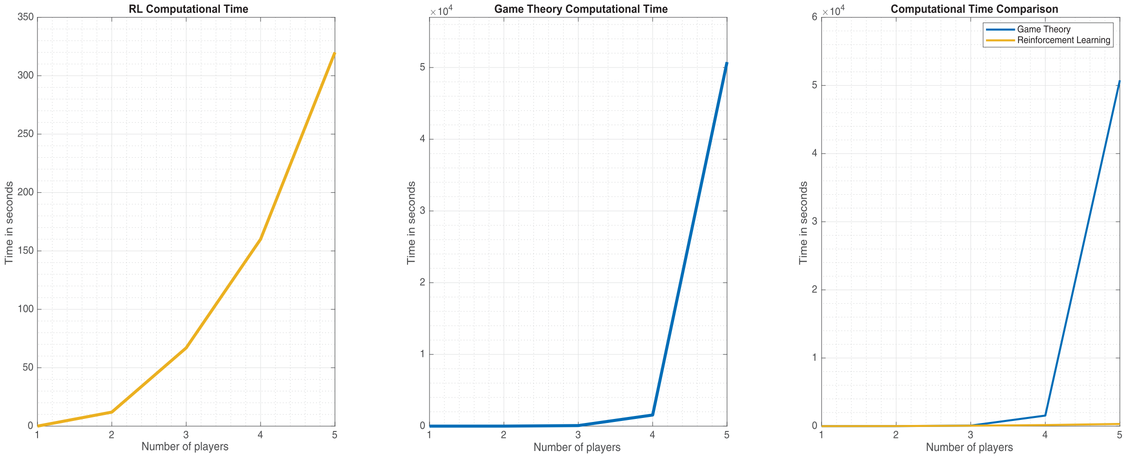

FIGURE 4. The computational time comparison.

an action and then needs to observe the actions taken by other agents, in order to update the appropriate Q(s,a)value.

The multi-agent reinforcement learning help to refine the strategy set in (15), and transform it as following:

$$\psi(x_{i,t}) = (\mathbf{N}, \{\mathbf{A}^{i,t}\}_{i\in N}, \{\delta^{i,t}\}_{i\in N}) \tag{35}$$

where A is the set of two strategies selected by multi-agent reinforcement learning.

V. RESULTS

In this section, bus 2, bus 59, bus 41, bus 24, and bus 13 are selected to demonstrate the proposed methodology and are highlighted in Fig.1. The results in Fig. 2 and Fig. 3 show that the proposed hybrid architecture can minimize the operational cost of each building. No other control strategy can realize a better operational cost which means that the novel multi-agent system can reach the global optimal point as the original game depicted in Sec III. And the algorithm is capable of optimizing under different weather conditions.

Fig. 4 shows the computational complexity comparison between the two proposed methodologies. It shows that when the number of players in a game increases, the time complexity of the reinforcement learning based game can be approximated as a linear function, while the time complexity of the game theory can be approximated as an exponential function. When there are just two or three players in a game, the computation time for the proposed methodology in Sec III is smaller than the methodology described in Sec IV. As the number of players increases, the reinforcement learning based algorithm shows its undefeatable advantage compared with the algorithm based on game theory solely.

VI. CONCLUSION

In this paper, we propose a multi-agent energy cost game and multi-agent reinforcement learning with discounted Q factor hybrid architecture. To achieve our goal, several analytical models are introduced and investigated in this paper. In terms of optimization results, the proposed hybrid system can achieve the same optimal results as the original energy cost game algorithm. However, from computational complexity perspective, the hybrid methodology can deal with the "curse of dimensionality" and make the computational load decrease dramatically. The simulation results prove the performance of our proposed architecture.

APPENDIX I. PROOF OF THEOREM 1

In light of Lemma 1, the feasible set of (Ψ) is nonempty as for every finite non co-operative game a Nash equilibrium exists. Further let S be the feasible region for (Ψ), then we have:

$$\min_{(\alpha,\gamma^1,\ldots,\gamma^n)\in S} \sum_{i\in N}(\gamma^i - \sum_{j=1}^{m^i} \delta^i(\alpha^{-i}, s^i_j)) \geqslant 0. \tag{36}$$

Thus, if α^* is a Nash equilibrium it is feasible for (Ψ), and from (1),

$$\sum_{i\in N}(\gamma^{i*} - \delta^i(\alpha^*)) = 0 \tag{37}$$

yielding that α^* is an optimal solution of (Ψ).

Conversely, suppose $(\alpha, \gamma^{1*}, \ldots, \gamma^{n*})$ is an optimal solution of (Ψ) then it satisfies (22) to (24).

By virtue of (22), $\sum_{i\in N}(\delta^i(\alpha^{-i*}, s^i_j))$.

But according to the existence theorem of Nash equilibrium, there must exist at least one $(\alpha, \gamma^1, \ldots, \gamma^n)$ feasible for (Ψ) with $\sum_{i\in N}(\gamma^i - \delta^i(\alpha)) = 0$. So for $(\alpha, \gamma^{1*}, \ldots, \gamma^{n*})$ to be a global minimum for (Ψ),

$$\sum_{i\in N}(\delta^i(\alpha^*) - \gamma^{i*}) = 0 \tag{38}$$

Consequently γ^* is an Nash equilibrium of game ψ, on account of Lemma 1. The payoff δ^{i*} is obviously the optimal expected payoff to player i.

We see that the problem of computing a Nash equilibrium of ψ reduces to that of solving the optimization problem (ψ) with optimal value zero.

REFERENCES

[1] U. UKGBC and G. B. Council, "Carbon reductions in existing non-domestic buildings," Green Building Council, London, U.K., Tech. Rep., 2011.

[2] H. Decc and E. S. S. Consultation, "Department of energy and climate change," Nat. Heat Map, London, U.K., Tech. Rep. 10D/719, 2013, pp. 12–19. [Online]. Available: http://tools.decc.gov.uk/nationalheatmap/

[3] M. H. Chung and E. K. Rhee, "Potential opportunities for energy conservation in existing buildings on university campus: A field survey in Korea," *Energy Buildings*, vol. 78, pp. 176–182, Aug. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778814003181

[4] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy Buildings*, vol. 40, no. 3, pp. 394–398, Jan. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778807001016

[5] S. Wu, K. Neale, M. Williamson, and M. Hornby, "Research opportunities in maintenance of office building services systems," *J. Qual. Maintenance Eng.*, vol. 16, no. 1, pp. 23–33, Mar. 2010.

[6] Y. Zhang *et al.*, "Distributed electrical energy systems: Needs, concepts, approaches and vision," *Acta Automatica Sinica*, vol. 43, no. 9, pp. 1544–1554, 2017.

[7] M. W. Ahmad, M. Mourshed, B. Yuce, and Y. Rezgui, "Computational intelligence techniques for HVAC systems: A review," *Building Simul.*, vol. 9, no. 4, pp. 359–398, 2016.

[8] S. D. J. McArthur *et al.*, "Multi-agent systems for power engineering applications—Part I: Concepts, approaches, and technical challenges," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1743–1752, Nov. 2007.

[9] C.-X. Dou and B. Liu, "Multi-agent based hierarchical hybrid control for smart microgrid," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 771–778, Jun. 2013.

[10] M. Pipattanasomporn, H. Feroze, and S. Rahman, "Multi-agent systems in a distributed smart grid: Design and implementation," in *Proc. IEEE/PES Power Syst. Conf. Expo.*, Mar. 2009, pp. 1–8.

[11] L. Hernandez *et al.*, "A multi-agent system architecture for smart grid management and forecasting of energy demand in virtual power plants," *IEEE Commun. Mag.*, vol. 51, no. 1, pp. 106–113, Jan. 2013.

[12] F.-D. Li, M. Wu, Y. He, and X. Chen, "Optimal control in microgrid using multi-agent reinforcement learning," *ISA Trans.*, vol. 51, no. 6, pp. 743–751, Nov. 2012.

[13] W. Saad, Z. Han, H. V. Poor, and T. Başar, "Game theoretic methods for the smart grid," 2012, *arXiv:1202.0452*. [Online]. Available: http://arxiv.org/abs/1202.0452

[14] R. Ragupathi and T. K. Das, "A stochastic game approach for modeling wholesale energy bidding in deregulated power markets," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 849–856, May 2004.

[15] J. Duan, Z. Yi, D. Shi, C. Lin, X. Lu, and Z. Wang, "Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC–DC microgrids," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5355–5364, Sep. 2019.

[16] S. Wang *et al.*, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst.*, early access, Apr. 23, 2020, doi: 10.1109/TPWRS.2020.2990179.

[17] J. Hao, Y. Lee, X. Dai, and J. J. Zhang, "Game-theoretic control of energy consumption through optimizing social costs for future smart grid," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2019, pp. 1–5.

[18] O. Seppanen, W. J. Fisk, and D. Faulkner, "Control of temperature for health and productivity in offices," Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. LBNL-55448, 2004.

[19] J. Hao, X. Dai, Y. Zhang, J. Zhang, and W. Gao, "Distribution locational real-time pricing based smart building control and management," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2016, pp. 1–6.

[20] O. Seppanen, W. J. Fisk, and Q. Lei, "Effect of temperature on task performance in office environment," Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. LBNL-60946, 2006.

[21] G. I. Earthman, "School facility conditions and student academic achievement," UCLA's Inst. Democracy, Educ., Access, Los Angeles, CA, USA, Tech. Rep. wws-rr008-1002, 2002.

[22] J. Hirsch. (2010). *eQUEST: The Quick Energy Simulation Tool*. [Online]. Available: http://www.doe2.com/equest/

[23] J. J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Numerical Analysis*. Dundee, U.K.: Springer, 1978, pp. 105–116.

[24] F. Meng and B. H. Chowdhury, "Distribution LMP-based economic operation for future smart grid," in *Proc. IEEE Power Energy Conf. Illinois*, Feb. 2011, pp. 1–5.

[25] J. Hao, Y. Gu, Y. Zhang, J. Jason Zhang, and D. Wenzhong Gao, "Locational marginal pricing in the campus power system at the power distribution level," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2016, pp. 1–5.

[26] B. Chatterjee, "An optimization formulation to compute Nash equilibrium in finite games," in *Proc. Int. Conf. Methods Models Comput. Sci. (ICM2CS)*, Dec. 2009, pp. 1–5.

[27] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[29] C. J. C. H. Watkins, ''Learning from delayed rewards,'' Ph.D. dissertation, King's College, Cambridge, U.K., 1989.

[30] A. Zilinskas, ''Simulation-based optimization: Parametric optimization techniques and reinforcement learning,'' *Interfaces*, vol. 35, no. 6, p. 535, 2005.

JUN HAO (Student Member, IEEE) received the M.S. degree in electrical and computer engineering, specializing in electric power engineering, from the University of Denver, CO, USA, in 2015, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. He dedicates himself to the research of electric power engineering. He has also participated in a lot of professional activities and serves as a reviewer for several IEEE Transactions.

DAVID WENZHONG GAO (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical and computer engineering, specializing in electric power engineering, from the Georgia Institute of Technology, Atlanta, USA, in 1999 and 2002, respectively. He is currently with the Department of Electrical and Computer Engineering, University of Denver, CO, USA. His current teaching and research interests include renewable energy and distributed generation, microgrid, smart grid, power system protection, power electronics applications in power systems, power system modeling and simulation, and hybrid electric propulsion systems. He is the General Chair of the 48th North American Power Symposium (NAPS 2016) and the IEEE Symposium on Power Electronics and Machines in Wind Applications (PEMWA 2012). He is an Editor of the IEEE Transactions on Sustainable Energy and an Associate Editor of the IEEE Journal of Emerging and Selected Topics in Power Electronics.

JUN JASON ZHANG (Senior Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2003 and 2005, respectively, and the Ph.D. degree in electrical engineering from Arizona State University, USA, in 2008. He is currently a Professor with the School of Electrical Engineering and Automation, Wuhan University. He authored/coauthored over 70 peer reviewed publications. His research interests include the areas of sensing theory, signal processing and implementation, time-varying system modeling, and their applications in intelligent power and energy systems. He is the Technical Co-Chair of the 48th North American Power Symposium (NAPS 2016).