Incentive Temperature Control for Green Colocation Data Centers via Reinforcement Learning

Rongrong Wang[1], Duc Van Le[1], Jikun Kang[2], Rui Tan[1], Xue Liu[2]

[1]School of Computer Science & Engineering, Nanyang Technological University, Singapore
[2]School of Computer Science, McGill University, Canada

***Abstract*—Increasing supply air temperatures is a rule-of-thumb approach to reduce cooling energy usage of data centers (DCs). However, colocation DCs are short of incentive programs to move tenants from the current over-cooling strategy despite the expanding allowable temperature ranges of the computing equipment. This paper considers an essential incentive mechanism, in which the DC operator offers monetary incentives to offset tenants' electricity payments. We propose an encoder-embedded multi-agent reinforcement learning solution to let the operator agent and tenant agents collaboratively find their policies for deciding the incentives and supply air temperatures, respectively, which are coupled in determining the DC's total cooling power usage. The solution does not require the cooling power model, which is complex and in general unavailable in practice. Moreover, as each tenant agent learns in the other tenants' latent state spaces defined by their pre-trained variational autoencoders, only encoded tenants' states are exchanged, thereby mitigating information leakage concerns. Extensive trace-driven evaluation and comparison with three baselines show that our solution effectively incentivizes tenants to move from the over-cooling strategy and achieves substantial cooling power savings.**

***Index Terms*—Colocation Data Centers, Temperature Control, Incentive, Multi-Agent Learning, Reinforcement Learning**

I. Introduction

The data center (DC) industry provides the infrastructures for cloud computing and has been growing. As the DC sector uses significant energy (2% [1] to 7% [2] of a country's total electricity usage), increasing DC energy efficiency is crucial to the carbon emission management and sustainability goals. DCs have two broad business models of *enterprise* and *colocation*. An enterprise DC is a system entirely owned and operated by a single entity. Differently, a colocation DC hosts the information technology (IT) systems of multiple tenants through quality of service (QoS) agreements on the space, power, cooling, and communication supplies. In the endeavors of improving energy efficiency, compared with enterprise DCs, colocation DCs involve one extra but intricate factor, i.e., the engagement with the tenants. To make sense, the endeavor needs to be beneficial to both the colocation DC operator and the tenants with different objective functions.

Following the above principle, the existing studies [3], [4] have proposed incentive programs to encourage tenants to reduce IT power usage. However, in air-cooled colocation DCs, how to incentivize the tenants to control their supply air temperatures in a collaborative effort to reduce the power usage of the complex cooling infrastructure still remains as an open problem. In today's practices, air-cooled DCs often adopt low supply air temperatures for large thermal safety margins, which, however, lead to high energy overheads (e.g., around 40% of DC energy used for cooling [5]). At the same time, the American Society of Heating, Refrigeration and Air-Conditioning Engineers (ASHRAE) has been working on extending the recommended allowable temperature ranges of IT equipment in its published *de facto* industry standard [6]. Pilot trials [7] in enterprise DCs have also suggested the feasibility of adopting higher supply air temperatures and the resulted cooling energy saving potential of up to 37%. Therefore, effective incentive programs to encourage the tenants to move from the current over-cooling strategy are desirable for greening colocation DCs.

In this paper, we consider an essential incentive mechanism, in which the colocation DC operator offers monetary incentives to offset the electricity payments of a tenant. Each tenant aims at minimizing the expected net payment by choosing its supply air temperature periodically based on its latest IT load. The operator, while being driven to reduce the DC's total power usage due to say the imposed carbon tax or social responsibility, aims at maximizing the expected revenue by choosing the giveaway incentives based on the tenants' IT loads and their chosen supply air temperatures. If the operator gives positive incentives for temperatures higher than the original over-cooling temperatures, the DC's total cooling power usage will reduce.

However, the design of the operator's and tenants' decision-making policies in the above incentive mechanism faces two major challenges. First, the operator and the tenants are coupled through a complex cooling infrastructure, which usually involves two stages of individual computer room air conditioners (CRACs) as the front stage and the shared chilled water system as the back stage. The IT loads and supply air temperatures of all tenants jointly affect the DC's total cooling power usage and there is no simple split for attributing to the tenants [8]. As such, the design of a tenant's policy on a sole basis is unlikely effective. Moreover, deriving analytical solutions to a game-theoretic formulation of the problem is not promising, due primarily to the unavailability of holistic and accurate models of DCs' cooling systems in practice. Second, intuitively, information transparency among the tenants coupled through the cooling system is beneficial for each tenant to make informed decision and can contribute to the

efficiency of the incentive mechanism. However, the sharing of tenants' local states (i.e., IT loads) for the transparency may leak critical information, e.g., type of computation and vulnerable time for further attacks such as power attack [9].

This paper proposes an encoder-embedded multi-agent reinforcement learning (MARL) solution to address the above two challenges collectively. Specifically, the MARL system consists of an operator agent (OA) and multiple tenant agents (TAs) which interact with each other and the environment (i.e., the cooling system) iteratively over time to learn their optimal policies. The OA runs the deep deterministic policy gradient (DDPG) learner and the TAs run their respective encoder-embedded multi-agent DDPG (MADDPG) learners. As MARL is model-free (i.e., it does not require the holistic and accurate model of the environment), it is suitable for the colocation DCs' adoption. To address the information leakage concern, each TA builds a variational autoencoder (VAE) for its IT load data and uses the VAE's encoder to mask its local state before sharing with other TAs. As such, each TA's MADDPG learner operates in the latent state spaces of all other TAs and can still capture the near-optimal policies.

We conduct extensive real-trace-driven simulations in EnergyPlus [10] to evaluate the effectiveness of our proposed incentive mechanism in comparison with three baselines. The simulation results show that our proposed incentive mechanism can effectively move the tenants from the over-cooling strategy and achieve about 35% saving in total cooling power.

II. Related Work

Incentive Colocation DCs: The existing incentive programs designed for colocation DCs mainly focus on the demand response (DR). The study in [11] proposes an incentive approach for a hierarchical DR problem, where the relationship between the operator and tenants is formulated as a Stackelberg game. In [4], a market-oriented incentive program is proposed to encourage tenants to share their idle servers and form a public resource pool. The work of [12] proposes the MesPP program to maximize energy reduction considering the incentive budget constraint of the operator. Different from these studies that focus on incentivizing the tenants to reduce IT loads, this work incentivizes the tenants to raise supply air temperatures to reduce the cooling power usage.

MARL for DC Control: MARL-based DC control employs multiple agents to make decisions and improve the DC performance. In [13], to improve energy efficiency, two DDPG agents learn to cooperatively control IT and cooling systems, respectively. In [14], MARL is applied to address the renewable energy demand-supply matching problem for a set of geo-distributed DCs. The work in [15] applies MARL for distributed cooling control, where each agent learns the optimal cooling policy to minimize the cooling power of a CRAC unit. In [16], the multi-agent Q-learning is applied for task scheduling. In the above MARL approaches, the agents are trained to optimize a global objective function and share information for coordination. Differently, we consider a multi-agent system where each agent optimizes its specific objective.

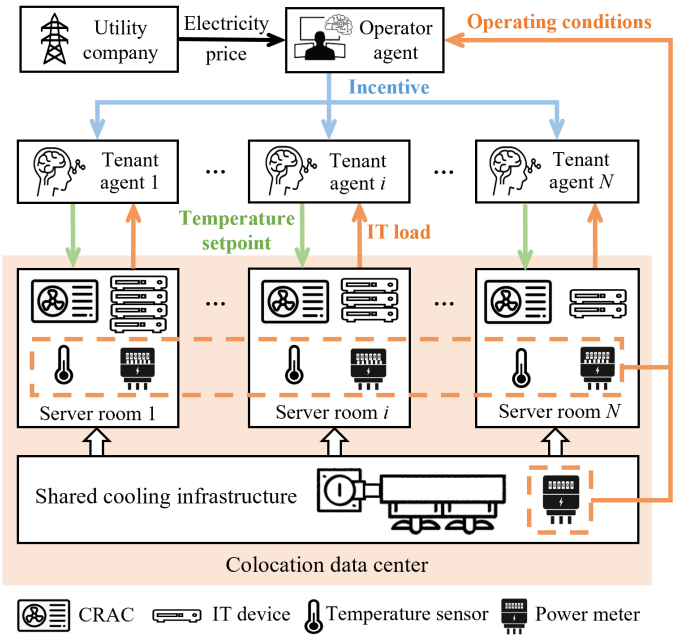


Fig. 1. Colocation DC system model.

Moreover, the tenant agents apply the VAEs to mask their local states before shared to other tenant agents. It is more challenging for the agents to learn from the masked data.

III. System Model & Problem Formulation

A. Colocation DC System

Fig. 1 illustrates a colocation DC that serves multiple tenants by leasing a server room to each tenant to house IT equipment. The DC operator purchases electricity from the utility company and manages the DC facility. Let $\mathcal{N} = \{1, ..., N\}$ denotes the set of N tenants. A tenant $i \in \mathcal{N}$ decides the supply air temperature setpoint T_i of the CRAC unit in the room. The operator controls a two-stage cooling system [17]. The first stage consists of CRAC units in all rooms and each unit transfers the heat from the IT equipment to the shared cooling system. In the second stage, the shared cooling system dissipates the heat collected from all server rooms to the atmosphere. The operator charges each tenant based on the IT power usage measured by the power meter in its room. Generally, the cooling power usage decreases with the supply air temperature setpoints of the server rooms. To improve the DC energy efficiency, the operator can encourage the tenants to raise their temperature setpoints by offering monetary incentives. Meanwhile, each tenant considers its technical constraints in determining the temperature setpoint for its room.

B. Problem Formulation

Time is divided into identical intervals referred to as the *control period*. The beginning time instant of a time interval is called a *time step*. At every time step, each tenant decides its temperature setpoint. Then, the operator gives tenants monetary incentives. We formulate the following two optimization

problems specifying the objectives and constraints of tenants and the operator, respectively.

Tenant's Payment Minimization: Let $L_{i,k}$ denote the IT power usage of tenant i at the time step k. The operator charges tenant i by $\eta\rho_k L_{i,k}$, where η is the constant electricity unit price applied by the operator, and ρ_k is the real-time power usage effectiveness (PUE) of the DC. The value of ρ_k is measured and published by the operator. By participating in the incentive program, tenant i receives the non-negative monetary incentive $b_{i,k}$ from the operator for adopting the temperature setpoint $T_{i,k}$ in the next control period. Then, the net payment of tenant i is calculated by $u_{i,k} = \eta\rho_k L_{i,k} - b_{i,k}$. The tenant aims to decide the temperature setpoints to minimize its long-term average net payment subject to the temperature constraint of its IT equipment. This tenant's payment minimization problem, denoted by OPT-T, is formulated as:

$$\min_{T_{i,k},\forall k} \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} u_{i,k}, \quad \text{s.t.} \quad T_{\min,i} \le T_{i,k} \le T_{\max,i}, \quad (1)$$

where $T_{\min,i}$ and $T_{\max,i}$ are the minimum and maximum allowable supply air temperature setpoints of tenant i's IT equipment, respectively.

Operator's Revenue Maximization: The operator controls the cooling system to maintain the temperature setpoints required by all tenants. Let $\mathbf{T}_k = [T_{1,k}, ..., T_{N,k}]$ and $\mathbf{L}_k = [L_{1,k}, ..., L_{N,k}]$ denote the temperature and IT power vectors at the time step k, respectively. We model the total DC power usage by $P_k^{\text{DC}} = f(\mathbf{T}_k, \mathbf{L}_k)$, where $f(\cdot)$ is a non-linear function of $\mathbf{T}_k$ and $\mathbf{L}_k$. The problem formulation does not require a specific model for $f(\cdot)$. We adopt the model from [8] to generate the ground truth for the evaluation experiments in this paper. Let $\mathbf{b}_k = [b_{1,k}, ..., b_{N,k}]$ denote the vector of monetary incentives that the operator gives the tenants based on their IT power usages and temperature setpoints at the time step k. Then, the revenue of the operator for running the colocation DC is calculated by $v_k = \sum_{i=1}^{N} u_{i,k} - \mu P_k^{\text{DC}}$, where μ is the constant electricity unit price from the utility company. The operator's objective is to decide the incentives which maximize its long-term expected revenue. This operator's revenue maximization problem, denoted by OTP-O, is formulated as:

$$\max_{\mathbf{b}_k,\forall k} \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} v_k, \quad \text{s.t.} \quad b_{i,k} \ge 0, i = 1, \ldots, N. \quad (2)$$

C. Challenges

Solving OPT-T and OPT-O analytically is challenging due to the complex combined impact of the tenants' temperature setpoints and IT power usages on the DC power usage. Specifically, the operator gives a higher incentive to a tenant if this tenant adopts a higher temperature, while each tenant adopts the highest allowable temperature according to its IT equipment's thermal specification. However, the server power in general increases with the temperature [18]. Thus, adopting the highest allowable temperatures may not be the tenants' best policy since it may lead to higher net payment. In this paper, we design an MARL-based incentive mechanism to let the operator and the tenants learn their respective policies for determining the tenants' incentives and temperature setpoints to solve the OPT-T and OPT-O optimization problems.

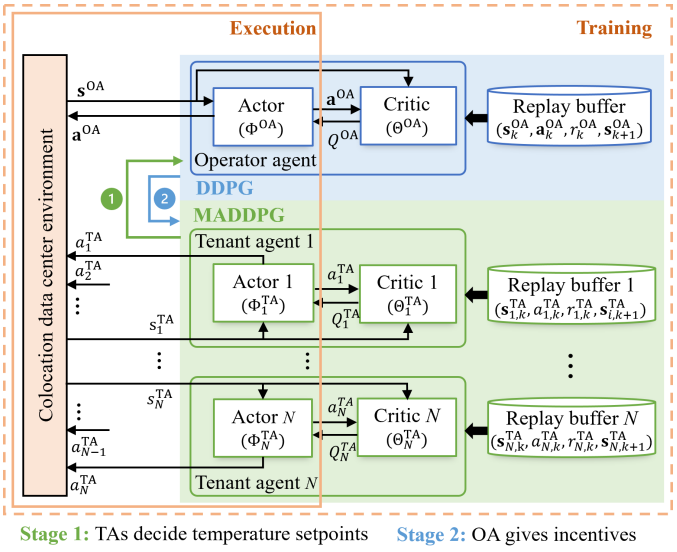


Fig. 2. Structure of proposed MARL framework.

IV. MARL-based Incentive Mechanism

A. Markov Game

We formulate a Markov game (MG) [19] consisting of N tenant agents and one operator agent. The learning processes of agents are modeled as Markov decision processes (MDPs).

1) Tenant Agents (TAs): The state, action, and reward functions of the TA are defined as follows.

State: At the time step k, the state of TA i is a vector of IT loads of all server rooms denoted by $\mathbf{s}_{i,k}^{\text{TA}} = [L_{1,k}, \ldots, L_{N,k}]$. Under this formulation, all TAs exchange IT load information.

Action: Based on the observed state $\mathbf{s}_{i,k}^{\text{TA}}$, the TA i decides an action $a_{i,k}^{\text{TA}}$ which is the supply air temperature setpoint $T_{i,k}$ selected from the range of $[T_{\min,i}, T_{\max,i}]$ for its server room in the next control period. In practice, due to the granularity of CRAC's temperature control (e.g., 1°C), the TA can only choose discrete values for $a_{i,k}^{\text{TA}}$.

Reward: Upon receiving the actions of all TAs at the time step k, the operator decides an incentive $b_{i,k}$ for each tenant and controls the cooling system to maintain the required temperature setpoints in server rooms for the next control period. Given $\mathbf{s}_{i,k}^{\text{TA}}$ and $a_{i,k}^{\text{TA}}$, the reward of TA i is $r_{i,k}^{\text{TA}} = -\eta\rho_k L_{i,k} + b_{i,k}$, which is the negative of its net payment.

2) Operator Agent (OA): The state, action, and reward formulations of the OA are presented as follows.

State: At time step k, the operator state $\mathbf{s}_k^{\text{OA}}$ is a vector consisting of IT loads and temperature setpoints of all TAs, and is denoted by $\mathbf{s}_k^{\text{OA}} = [L_{1,k}, \ldots, L_{N,k}, T_{1,k}, \ldots, T_{N,k}]$.

Action: The operator action $\mathbf{a}_k^{\text{OA}}$ is the monetary incentives given to all tenants, which is given by $\mathbf{a}_k^{\text{OA}} = [b_{1,k}, ..., b_{N,k}]$. Each component $b_{i,k}$ is selected from a range of $[0, b_{\max}]$, where $b_{\max}$ is the upper-bound value of the incentive.

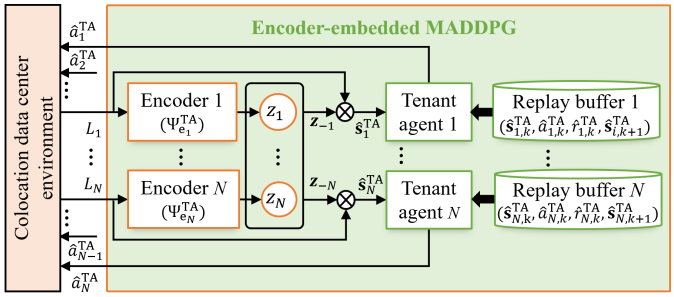


Fig. 3. Structure of encoder-embedded MADDPG.

Reward: The reward r_k^{OA} is the operator revenue and is given by $r_k^{\mathrm{OA}} = \sum_{i=1}^{N}(\eta\rho_k L_{i,k} - b_{i,k}) - \mu P_k^{\mathrm{DC}}$, where $\sum_{i=1}^{N}(\eta\rho_k L_{i,k} - b_{i,k})$ is the total payment of all TAs, and μP_k^{DC} is the DC power consumption payment.

B. Design of MARL-based Incentive Mechanism

We propose an MARL-based incentive mechanism to solve the above MG between OA and TAs. Specifically, as illustrated in Fig. 2, the OA adopts DDPG [20] to learn the incentive policy, while the TAs use MADDPG [21] to learn their temperature setpoint management policies.

1) DDPG Design of OA: The OA is formulated in the actor-critic framework. At the time step k, the actor network parameterized by Φ^{OA} takes the state $\mathbf{s}_k^{\mathrm{OA}}$ to make the action $\mathbf{a}_k^{\mathrm{OA}}$ using the policy $\pi^{\mathrm{OA}}(\mathbf{s}_k^{\mathrm{OA}}|\Phi^{\mathrm{OA}})$. The critic network with parameters Θ^{OA} provides the Q-value estimation $Q^{OA}(\mathbf{s}_k^{\mathrm{OA}}, \mathbf{a}_k^{\mathrm{OA}}|\Theta^{\mathrm{OA}})$. While interacting with the environment, the experience $(\mathbf{s}_k^{\mathrm{OA}}, \mathbf{a}_k^{\mathrm{OA}}, r_k^{\mathrm{OA}}, \mathbf{s}_{k+1}^{\mathrm{OA}})$ is stored into a replay buffer D^{OA}. To stabilize the training process, the target actor and target critic are used, which are parameterized by Φ'^{OA} and Θ'^{OA}, respectively. The detailed design of the DDPG's actor and critic networks can be found in [20].

2) MADDPG Design of TAs: The MADDPG framework enables multiple DDPG agents to make decisions and interact within the environment. Specifically, each TA i has an actor and a critic with Φ_i^{TA} and Θ_i^{TA} as the model parameters, respectively, as well as their target network copies parameterized by $\Phi_i'^{\mathrm{TA}}$ and $\Theta_i'^{\mathrm{TA}}$, respectively. At the time step k, the TA i observes the state $\mathbf{s}_{i,k}^{\mathrm{TA}}$ and takes the action $a_{i,k}^{\mathrm{TA}}$ using the policy $\pi_i^{\mathrm{TA}}(\mathbf{s}_{i,k}^{\mathrm{TA}}|\Phi_i^{\mathrm{TA}})$. The Q-value estimation from the critic is given by $Q_i^{\mathrm{TA}}(\mathbf{s}_{i,k}^{\mathrm{TA}}, a_{i,k}^{\mathrm{TA}}|\Theta_i^{\mathrm{TA}})$. The replay buffer D_i^{TA} stores the experience $(\mathbf{s}_{i,k}^{\mathrm{TA}}, a_{i,k}^{\mathrm{TA}}, r_{i,k}^{\mathrm{TA}}, \mathbf{s}_{i,k+1}^{\mathrm{TA}})$. Using the mini-batch samples from D_i^{TA}, the policy gradient and loss function of TA i are calculated following the MADDPG algorithm [21] to update network parameters. Moreover, the Gumbel-Softmax estimator [22] is applied for each TA to address the non-differentiable issue due to its discrete action space.

3) Training Process: At the beginning, each TA i gets the state $\mathbf{s}_{i,k}^{\mathrm{TA}}$ from the environment and takes the action $a_{i,k}^{\mathrm{TA}}$ based on the policy $\pi_i^{\mathrm{TA}}(\mathbf{s}_{i,k}^{\mathrm{TA}}|\Phi_i^{\mathrm{TA}})$. After all TAs execute actions $a_{1,k}^{\mathrm{TA}}, \ldots, a_{N,k}^{\mathrm{TA}}$, the OA obtains the state $\mathbf{s}_k^{\mathrm{OA}}$ and executes the action $\mathbf{a}_k^{\mathrm{OA}}$ given by the policy $\pi^{\mathrm{OA}}(\mathbf{s}_k^{\mathrm{OA}}|\Phi^{\mathrm{OA}})$. Then, TAs are rewarded by $r_{1,k}^{\mathrm{TA}}, \ldots, r_{N,k}^{\mathrm{TA}}$, respectively. The OA receives the reward r_k^{OA}. The environment then moves to the next decision-making step and gives the states $\mathbf{s}_{1,k+1}^{\mathrm{TA}}, \ldots, \mathbf{s}_{N,k+1}^{\mathrm{TA}}$ and $\mathbf{s}_{k+1}^{\mathrm{OA}}$. By collecting the transitions, replay buffers are built for each TA i as D_i^{TA} and the OA as D^{OA} for training.

C. MARL with Data Masking

In §IV-B, each tenant reveals its IT load to all other tenants, which may raise information leakage concerns. Specifically, a tenant may be compromised unconsciously by an external adversary. As a result, the external adversary may exfiltrate the IT load data of all tenants. A server room's continuous high IT load may indicate critical services and imply a high-value target of cyber attacks. Moreover, the adversary can analyze the IT load data to infer the type of running applications, which may be the privacy of the tenant.

We propose VAE-based data masking to mitigate the tenants' information leakage concern. VAE is a neural network consisting of two parts: an encoder and a decoder, which are connected through the latent space representation. In our approach, each TA i trains a VAE using its own IT load data and then uses the encoder of the trained VAE, denoted by $\Psi_{e_i}^{TA}$ to mask its IT load data before sharing with other TAs.

After TA i receives the masked data from other TAs, which is denoted by $\mathbf{z}_{-i,k} = [z_{1,k}, ..., z_{i-1,k}, z_{i+1,k}, ..., z_{N,k}]$, this TA takes the action $\hat{a}_{i,k}^{\mathrm{TA}}$ based on the state $\hat{\mathbf{s}}_{i,k}^{\mathrm{TA}} = [L_{i,k}, \mathbf{z}_{-i,k}]$. By interacting with the environment, the experience $(\hat{\mathbf{s}}_{i,k}^{\mathrm{TA}}, \hat{a}_{i,k}^{\mathrm{TA}}, \hat{r}_{i,k}^{\mathrm{TA}}, \hat{\mathbf{s}}_{i,k+1}^{\mathrm{TA}})$ is stored in the replay buffer $\hat{D}_i^{\mathrm{TA}}$ for training. Fig. 3 illustrates the structure of the encoder-embedded MADDPG.

V. Performance Evaluation

A. Experiment Setup

We use EnergyPlus 9.5 [10] to simulate the physical processes of a colocation DC with six server rooms (i.e., $N = 6$). The cooling system configuration follows the implementations in [8]. We use the IT power traces of the Blue Waters dataset [23] to simulate the tenants' IT power usages with diverse usage patterns. In the simulations, we set $T_{\min,i}$ = 20°C and $T_{\max,i}$ = 30°C for all tenants $i \in \mathcal{N}$. The control period is 10 minutes. The electricity unit prices charged by the power grid and the DC operator are set to $\mu = 1$ and $\eta = 1.8$, respectively. The maximum incentive $b_{\max}$ is 15.

The latent dimension of each VAE is 10 and the encoder has two hidden layers with 64 and 32 ReLU neurons, respectively. The actor and critic networks of the OA and TAs have the same architecture with three hidden layers, each consisting of 128 ReLU neurons. We use the Adam optimizer with a learning rate of 0.001. The discount factor and the soft update parameter are 0.99 and 0.01, respectively. The mini-batch size is 64, and the replay buffer size is 50,000.

B. Training Performance

First, we evaluate the training performance of the following two variants of our proposed approach. (1) *Masked*: Each TA uses a trained VAE's encoder to mask its IT power state before sharing it with other TAs for decision-making. (2) *Unmasked*: Each TA reveals its IT power state to all other TAs without data

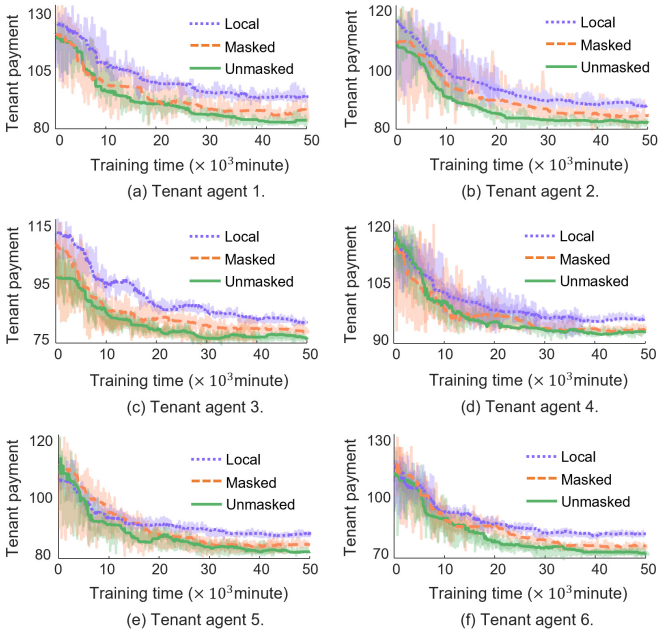


Fig. 4. The net payments of the six tenant agents during the training.

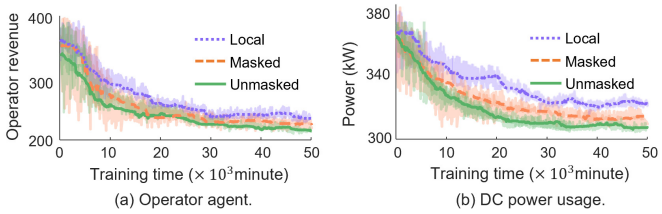


Fig. 5. Training performance of the operator agent.

masking. We also compare with a baseline approach, called *Local*, where each TA makes the decision solely based on its local IT power state.

Fig. 4 shows the net payment traces of six TAs during training. With three approaches, the net payments of all TAs decrease in the first 3,000 time steps and then become flat. Moreover, the Masked and Unmasked approaches mostly have lower payments than those of the Local approach. This indicates that the IT load information exchange helps TAs in learning optimal temperature management policies to reduce the net payments. In addition, the Masked and Unmasked approaches generate mostly similar tenant payments after the training period of 3,000 time steps. This implies that the data masking by the VAE's encoders does not prevent the TAs from learning the near-optimal temperature management policies.

Fig. 5 presents the operator revenue and the DC power usage during training. With three approaches, the operator revenue and the DC power usage decrease over training and become flat after 3,000 time steps. Moreover, compared with the other two approaches, the Local approach overall has higher operator revenue and DC power usage. This is because the tenants under the Local approach give more payments for taking lower temperature setpoints. Moreover, the Masked and Unmasked approaches generate similar operator revenues and DC power usages. The reason is that TAs can learn the near-

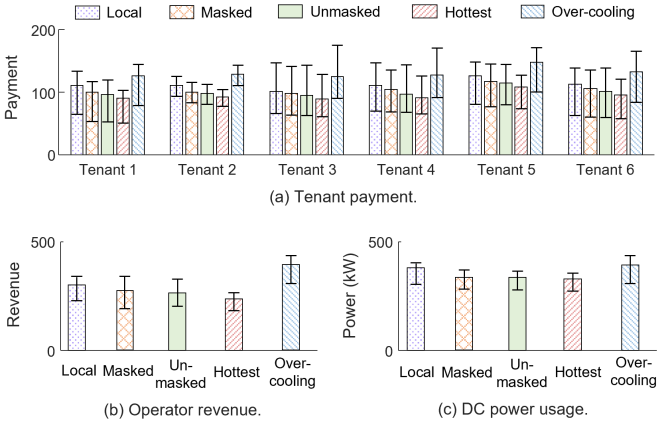


Fig. 6. Execution performance of 1 day. The bar, upper cap, and lower cap represent the average, maximum, and minimum values, respectively.

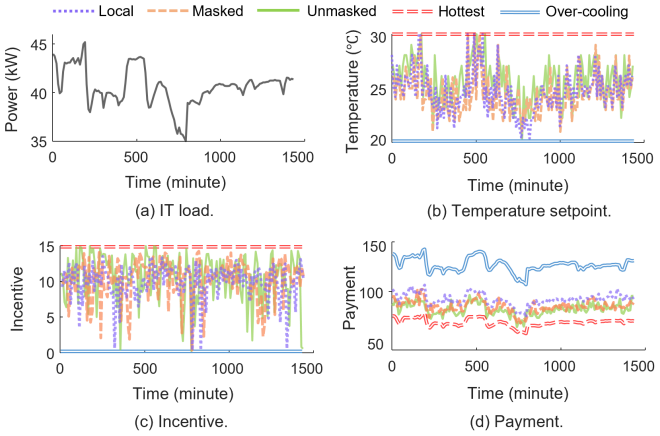


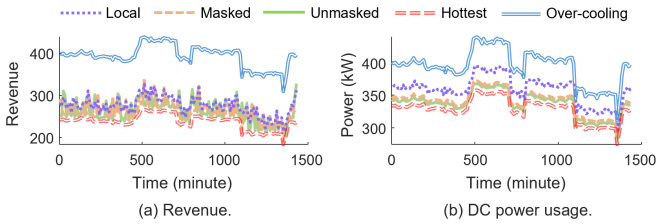
Fig. 7. Traces of the IT load, incentive, temperature setpoint, and net payment of the tenant 3 over an execution period of 1 day.

optimal temperature management policies with masked data. Therefore, the TAs select similar temperature setpoints and the OA offers similar incentives under these two approaches.

C. Execution Performance

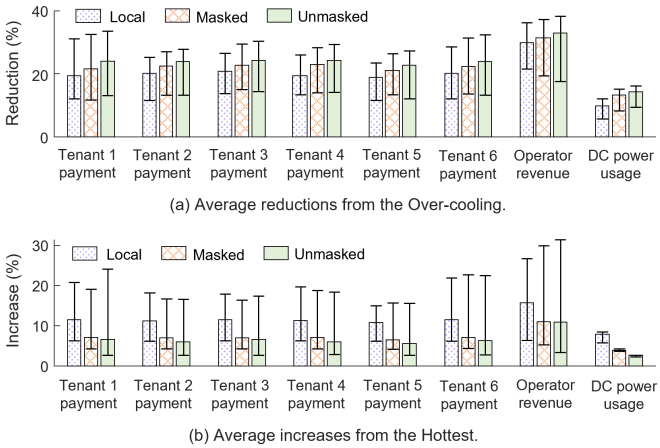
We compare the execution performance of three MARL-based incentive approaches with the following two baseline approaches. *(1) Hottest*: All tenants always select the highest temperature setpoint of 30°C and the operator offers the maximum incentive of $b_{\max}$ to tenants. *(2) Over-cooling*: The tenants always select the lowest temperature setpoint of 20°C and receive no incentives from the operator. There is no IT load information exchange among TAs in both baseline approaches.

Fig. 6 shows the 1-day execution results of the average tenant net payments, operator revenue, and DC power usage of various approaches. According to Fig. 6(a), the Masked and Unmasked approaches generate similar tenant payments, operator revenue, and DC power usage, which are lower than those with the Local approach. Specifically, the average net payments of six tenants with the Local approach are higher than those of the Ummasked approach by 8.75%, 7.33%, 5.74%, 9.89%, 9.04%, and 8.68%, respectively. Similarly, as shown in Figs. 6(b) and (c), compared with the Unmasked approach, the Local approach leads to 7.52% and 8.06%



(a) Revenue. (b) DC power usage.

Fig. 8. Traces of the operator revenue and DC power usage over 1 day.



(a) Average reductions from the Over-cooling.

(b) Average increases from the Hottest.

Fig. 9. Comparison with the baseline approaches over a 10-days' period.

higher operator revenue and DC power usage, respectively. Moreover, the Over-cooling approach generates the highest tenant payment, operator revenue, and DC power usage.

Fig. 7 shows the IT power usage, incentive, temperature setpoint, and net payment traces of 1-day execution of tenant 3. With the Local, Masked, and Unmasked approaches, tenant 3 is incentivized to change temperature setpoints in response to IT load dynamics for higher incentives. Fig. 8 shows the corresponding operator revenue and DC power usage.

We further compare the performance of all approaches over a longer period of 10 days. From Fig. 9(a), the Masked and Unmasked approaches reduce the average tenant payment by up to 30% from the Over-cooling approach. Moreover, the Local, Masked, and Unmasked approaches move the tenants from the over-cooling strategy and reduce the DC power usage by 8.83%, 13.32%, and 14.26%, respectively. Particularly, the Masked and Unmasked approaches have more reductions in tenant payment and DC power usage than the Local approach and achieve about 35% cooling power reduction from the *Baseline-Hot* approach. From Fig. 9(b), three MARL-based approaches generate higher tenant payments, operator revenue, and DC power usage than the Hottest approach. Specifically, they increase the average tenant payment, operator revenue, and DC power usage by up to 11%, 15%, and 9%, respectively.

VI. Conclusion

In this paper, we propose an incentive mechanism to effectively incentivize the tenants in the colocation DC to raise the supply air temperature setpoints of their server rooms, aiming at maximizing all participants' financial benefits. The proposed mechanism adopts an MARL framework where the operator learns a policy to determine the monetary incentives offered to the tenants, and the tenants learn policies to set the supply air temperatures. Moreover, our data masking method helps alleviate the information leakage concern due to IT load data exchange among tenants. Extensive trace-driven evaluation shows that our proposed approach enables tenants to reduce net payments and the colocation DC to reduce power usage.

Acknowledgement

This project is supported by the National Research Foundation, Singapore, funded under Energy Research Test-Bed and Industry Partnership Funding Initiative, part of the Energy Grid (EG) 2.0 programme.

References

[1] US DoE, "Data centers and servers," 2024.
[2] ASEAN Briefing, "Singapore's data center sector," 2023.
[3] S. Malla and K. Christensen, "A survey on power management techniques for oversubscription of multi-tenant data centers," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–31, 2019.
[4] Y. Wang, F. Zhang, C. Chi, S. Ren, F. Liu, R. Wang, and Z. Liu, "A market-oriented incentive mechanism for emergency demand response in colocation data centers," *Sustainable Computing: Informatics and Systems*, vol. 22, pp. 13–25, 2019.
[5] J. Cho and Y. Kim, "Improving energy efficiency of dedicated cooling system and its contribution towards meeting an energy-optimized data center," *Applied Energy*, vol. 165, pp. 967–982, 2016.
[6] ASHRAE, *2021 Equipment Thermal Guidelines for Data Processing Environments*, 2021.
[7] D. Van Le, Y. Liu, R. Wang, R. Tan, and L. H. Ngoh, "Air free-cooled tropical data center: Design, evaluation, and learned lessons," *IEEE Trans. Sustain. Comput.*, vol. 7, no. 3, pp. 579–594, 2021.
[8] R. Wang, D. Van Le, R. Tan, Y.-W. Wong, and Y. Wen, "Real-time cooling power attribution for co-located data center rooms with distinct temperatures," in *BuildSys*, 2020.
[9] Z. Xu, H. Wang, Z. Xu, and X. Wang, "Power attack: an increasing threat to data centers." in *NDSS*, 2014.
[10] "Energyplus," https://energyplus.net.
[11] M. N. Nguyen, D. Kim, N. H. Tran, and C. S. Hong, "Multi-stage stackelberg game approach for colocation datacenter demand response," in *APNOMS*, 2017.
[12] C. Chi, K. Ji, A. Marahatta, F. Zhang, Y. Wang, and Z. Liu, "An incentive mechanism for improving energy efficiency of colocation data centers based on power prediction," in *ISCC*, 2020.
[13] C. Chi, K. Ji, P. Song, A. Marahatta, S. Zhang, F. Zhang, D. Qiu, and Z. Liu, "Cooperatively improving data center energy efficiency based on multi-agent deep reinforcement learning," *Energies*, vol. 14, 2021.
[14] H. Wang, H. Shen, J. Gao, K. Zheng, and X. Li, "Multi-agent reinforcement learning based distributed renewable energy matching for datacenters," in *ICCP*, 2021.
[15] D. Chen, J. Wan, L. Li, and C. Liu, "Distributed data center cooling control based on multi-agent reinforcement learning," in *ICFTIC*, 2022.
[16] H. Yu and Y. Xia, "An energy saving control strategy based on multi-agent q-learning algorithm for data center," in *Journal of Physics: Conference Series*, vol. 2517, 2023.
[17] L. Perez-Lombard, J. Ortiz, and I. R. Maestre, "The map of energy flow in hvac systems," *Applied energy*, vol. 88, no. 12, pp. 5020–5031, 2011.
[18] D. Moss and J. H. Bean, "Energy impact of increased server inlet temperature," *APC white paper*, vol. 138, 2009.
[19] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2022.
[20] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *ICML*, 2014.
[21] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *NIPS*, 2017.
[22] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.
[23] NCSA, "Blue waters data sets." 2012, https://bluewaters.ncsa.illinois.edu/data-sets.