Available online at https://jcupt.bupt.edu.cn

The Journal of China Universities of Posts and Telecommunications

April 2026, 33(2): 15 – 30

Attention-driven electrical-computational cooperative scheduling approach via hierarchical multi-agent reinforcement learning

Li Tong[1], Yang Chao[2](✉), Sun Feng[1], Li Shujin[3], Ma Xiaoning[4]

1. Electric Power Research Institute, State Grid Liaoning Electric Power Co., Ltd., Shenyang 110006, China
2. State Grid Liaoning Electric Power Co., Ltd., Shenyang 110006, China
3. State Grid Liaoning Management Training Center, Shenyang 110102, China
4. Materials Branch, State Grid Liaoning Electric Power Co., Ltd., Shenyang 110103, China

Abstract

The surge of data center loads and high renewable penetration threatens power system stability, requiring joint power-computing scheduling. This paper presents an attention-driven hierarchical multi-agent proximal policy optimization (H-MAPPO) method to solve the coordination issue. A Transformer-based fine-grained task-computing-power coupling model is built to map diverse computing tasks to grid loads. The grid-computing interaction is formulated as a Stackelberg game via bi-level Markov decision process (MDP). The upper agent conducts global dispatch with proximal policy optimization (PPO), and lower agents use attention-based coordination algorithm PPO (ACA-PPO) for distributed task scheduling. Tests on German European network of transmission system operators for electricity (ENTSO-E) grid data and Google cluster traces reveal H-MAPPO attains 98.36% profit of the optimal mixed integer quadratic programming (MIQP), shortens solving time from 342.6 s to 2.5 s, and lifts task completion rate by 21.6% versus hierarchical independent proximal policy optimization (H-IPPO). The hierarchical attention-aided framework supports scalable real-time power-computing collaborative scheduling for low-carbon grids.

Keywords computing and power collaboration, adjustable resources, double-layer optimization, optimize scheduling

1 Introduction

With the rapid expansion of the digital economy, the surge in data center computing loads and the inherent volatility of renewable energy have created severe supply-demand imbalances in power systems. Collaborative scheduling of electricity and computing systems has thus become critical for grid stability[1-2]. However, existing research faces significant challenges in modeling and optimization efficiency.

Firstly, most existing studies rely on coarse-grained power usage effectiveness (PUE)-based or aggregate load models, where data centers are treated as indivisible loads. Such models cannot accurately capture the task-level heterogeneity and the nonlinear coupling between computing services and electricity demand, thus limiting the precision of collaborative scheduling[3-4]. Secondly, the collaborative system involves massive decision variables across generation,

Received date: 22-01-2026
Corresponding author: Yang Chao, E-mail: yangchaoneu@163.com
DOI: 10.19682/j.cnki.1005-8885.2026.1002

transmission, and resource allocation. Traditional methods like MIQP often require hours to solve high-dimensional problems, failing to meet the second-level response requirements of the electricity market[5-6]. Thirdly, while deep reinforcement learning (DRL) offers adaptive capabilities, existing frameworks often lack effective hierarchical coordination mechanisms to model the leader-follower interaction between grid dispatch and computing response[7-8].

To address these challenges, an attention-aware collaborative scheduling method based on H-MAPPO, namely hierarchical multi-agent reinforcement learning, is proposed in this paper. The main contributions are as follows.

1) This paper constructs a fine-grained task-computing-power coupling model. A service-oriented representation system is established to bridge heterogeneous computing service characteristics with physical power constraints, thereby enabling accurate mapping from service tasks to grid-perceptible loads.

2) This paper formulates a bi-level collaborative optimization framework for the power-computing coordination problem. The problem is modeled as a Stackelberg game formalized via a bi-level MDP, balancing the global economic optimality of the power system with the local flexibility of computing nodes.

3) This paper proposes a multi-head attention embedded H-MAPPO algorithm. A hierarchical framework is designed in which macro-dispatch is performed by the upper-level agent, and dynamic node correlations are captured by lower-level agents through ACA-PPO, namely an attention-based coordination algorithm.

The rests of this paper are organized as follows. Sect. 2 reviews the related work regarding electricity-computing collaboration and the application of reinforcement learning in scheduling problems. Sect. 3 constructs the fine-grained task-computing-power coupling model and establishes the mathematical formulation for the collaborative optimization problem. Sect. 4 details the proposed H-MAPPO framework, including the hierarchical game mechanism and the attention-based ACA-PPO algorithm. The simulation results and performance analysis are illustrated in Sect. 5. Finally, Sect. 6 concludes the paper and outlines future research directions.

2 Related work

2.1 Unified modeling of computing-electricity systems

Accurate system modeling is the prerequisite for effective collaboration. Existing research on computing load and its coupling with power systems demonstrates a trend evolving from macro statistics to micro mechanisms. Early macro models were predominantly based on statistical methods such as autoregressive integrated moving average (ARIMA)[9-10], treating data centers as black boxes focusing solely on total power consumption[11], thus lacking explanatory power for structural variations in internal loads[12]. Subsequent PUE models linked information technology (IT) power consumption to total power consumption[13-14]; however, by treating IT equipment as an indivisible entity, they failed to reflect the nonlinear relationship between the value of computing tasks and energy consumption[15-16]. Further hardware-level models attempted to establish functional relationships between power consumption and graphics processing unit (GPU) utilization[17]. Nevertheless, research indicates that task type[18] and software parameter configuration[19] are the fundamental drivers of energy fluctuations, which are difficult to represent accurately relying solely on hardware utilization. Although recent studies have proposed unified multi-party representation models to address the heterogeneity of distributed renewable energy systems[20], establishing a fine-grained link between specific computing tasks and power consumption remains challenging.

In summary, existing research still lacks an end-to-end fine-grained mapping from computing tasks to grid nodes. Existing models fail to capture the software-defined characteristics and elasticity of the tasks themselves, thereby limiting the precision of collaborative optimization. In view of this, this paper constructs a service-oriented task-computing-power' coupling model, aiming to systematically integrate task

characteristics with hardware performance to lay the foundation for realizing task-level collaborative optimization.

2.2 Electricity-computing collaborative optimization

Collaborative optimization of electricity-computing is primarily divided into two categories: traditional mathematical programming and artificial intelligence methods. Traditional methods represented by mixed integer linear programming (MILP) can theoretically obtain the global optimum[14]. However, when facing large-scale heterogeneous resources and dual uncertainties of generation and load, their computational complexity grows exponentially, and frequent re-solving makes it difficult to meet the real-time response requirements of the power market[21-22]. In contrast, DRL has become a research hotspot due to its model-free and adaptive characteristics. Li et al.[7] and Wu et al.[20] utilized deep deterministic policy gradient (DDPG) to optimize the operation of high-proportion renewable energy grids, but the single-agent paradigm faces a severe curse of dimensionality when dealing with large-scale electricity-computing systems[23]. Although multi-agent reinforcement learning (MARL) mitigates scalability issues, direct application to strongly coupled systems often leads to convergence difficulties due to environmental non-stationarity[24]. Existing improvement schemes, such as hierarchical communication[25] or parameter sharing[21], mostly lack formal modeling of the hierarchical interaction between macro scheduling and micro response, making it difficult to balance global economy with local flexibility.

In terms of algorithm selection, regarding the power system which demands extremely high safety and stability, although off-policy algorithms like DDPG and soft actor-critic (SAC) possess high sample efficiency, they are sensitive to hyperparameters and prone to training oscillation, making it difficult to ensure the smooth execution of scheduling strategies[26-27]. In contrast, PPO limits the policy update step size through a clipping mechanism, demonstrating superior performance in robustness and convergence stability[28-29], thus rendering it more suitable for macro-economic dispatch. For the computing system, considering the distributed collaborative nature of task scheduling, adopting the PPO framework based on the centralized training decentralized execution (CTDE) paradigm allows for both the utilization of global information to optimize strategies and the maintenance of the independent execution capabilities of nodes[30].

In summary, MARL provides a novel approach for solving the collaboration problem of electricity-computing systems. However, current research often addresses specific problems in isolation, lacking an integrated optimization framework capable of organically fusing corresponding solutions. Therefore, this paper proposes the H-MAPPO hierarchical DRL optimization framework, aiming to systematically address multiple challenges including non-stationarity, hierarchical coordination mechanisms, and heterogeneous information filtering, and to construct a holistic integrated optimization framework.

3 System modeling and framework

3.1 Architecture of electricity-computing collaboration system (ECCS)

The architecture of the ECCS proposed in this paper is shown in Fig. 1. This architecture takes the collaborative optimization center as the hub, achieving closed-loop control of the power subsystem and the computing subsystem through bidirectional information interaction. The power side covers uncontrollable renewable energy units, controllable traditional units, energy storage systems, and rigid base loads; the computing side is modeled as flexible load clusters capable of flexible response, containing heterogeneous computing resources and multi-category task queues. Based on global situational awareness, the collaboration optimization center guides the computing side to reschedule tasks by formulating regulatory electricity prices. Combined with the dispatch of unit output and energy storage on the physical side, this achieves the ECCS optimization of the entire system.

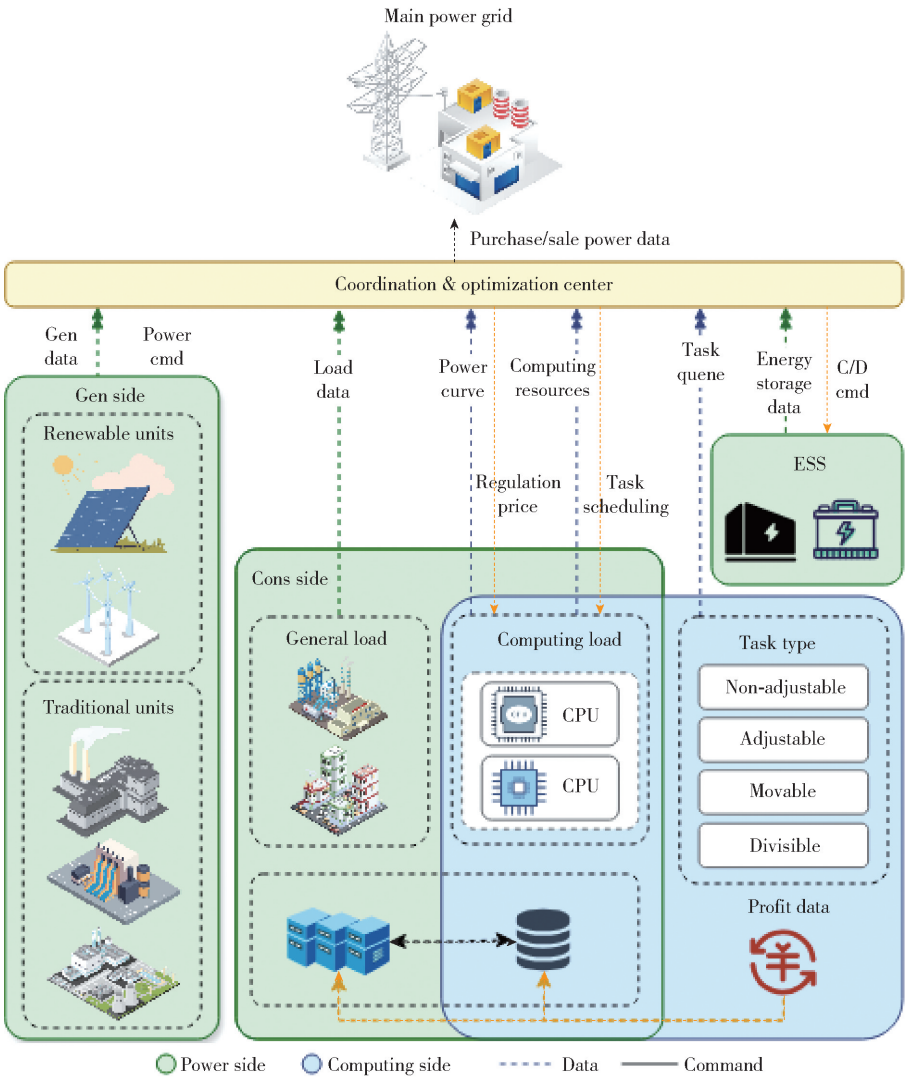

Fig. 1 Architecture of ECCS

3.2 ECCS representation model

This section establishes a unified quantitative model for heterogeneous resources within the collaborative architecture.

For the computing node, the model focuses on task-resource matching. Service requests are decomposed into tasks R, where each task i is formalized as a tuple $R_i = [R_{\text{sta},i}, R_{\text{cal},i}, R_{\text{pri},i}, R_{\text{occ},i}]$ standardized by 500 GFLOPS and 100 GB units. To capture flexibility, a shiftability matrix $\boldsymbol{D}_C = [d_{c_i,c_j}]_{m\times m}$ is defined, where $d_{c_i,c_j} = 1$ denotes time-shiftable tasks responsive to price signals, and $d_{c_i,c_j} = 0$ denotes rigid tasks. These tasks are processed by servers S, characterized by capacity limits and power states $[S_{\text{sta},j}, S_{\text{cal},j}, P_{S_j}^{\text{full}}, P_{S_j}^{\text{idle}}]$, and supported by energy storage E for power smoothing.

For the power node, the model quantifies the regulation potential of supply and demand. The supply side comprises generating units G following pre-dispatched plans P_{Gl}^{pre}. The demand side consists of loads L defined by feasible power ranges and regulation limits. Their response capabilities are unified by a regulation matrix $\boldsymbol{D}_L$, where $\boldsymbol{d} = [d_{l,r}]_{L\times R}$, $d_{l,r} \in \{0, 1, -1\}$ corresponds to fixed, shiftable, and curtailable loads, respectively.

3.3 ECCS mapping model

To accurately characterize the complex energy consumption features of heterogeneous computing resources when executing different types of tasks, and

to overcome the limitation of traditional physical formulas in capturing non-linear software-hardware coupling, this paper proposes a data-driven mapping model $\mathscr{F}_{\Phi}(\cdot)$ based on a Transformer deep neural network.

First, a multi-dimensional feature input space is constructed. The input feature vector $\boldsymbol{x}_t$ at time t is defined as consisting of resource occupation status, task software features, and environmental parameters:

$$\boldsymbol{x}_t = [\boldsymbol{U}_t, \boldsymbol{S}_t, \boldsymbol{E}_t] \tag{1}$$

where $\boldsymbol{U}_t$ represents the real-time aggregated load vector of various hardware resources within the node; $\boldsymbol{S}_t$ characterizes the software features of the currently running task set; and $\boldsymbol{E}_t$ contains external variables such as ambient temperature that affect cooling power consumption.

Second, a time-series energy consumption prediction model is established. The multi-head self-attention mechanism in the Transformer architecture is utilized to capture the long-term and short-term dependencies of load fluctuations to predict the basic power consumption P_{IT}^t of IT equipment:

$$P_{\mathrm{IT}}^t = T_{\Phi}(x_{t-w}, \ldots, x_t) \tag{2}$$

where $x_{t-w}, \ldots, x_t$ is the historical feature sequence of length w, and Φ represents the network weight parameters obtained through offline training.

Finally, combining the PUE of the data center and the regulation amount of the energy storage system, the total load $P_{C,\mathrm{all}}^t$ presented by the computing node to the power grid is obtained:

$$P_{C,\mathrm{all}}^t = P_{\mathrm{IT}}^t P^{\mathrm{PUE}}(P_{\mathrm{IT}}^t) + \sum_{h=1}^{v}(P_{c,h}^t - P_{d,h}^t) \tag{3}$$

where $P^{\mathrm{PUE}}(\cdot)$ is a non-linear fitting function regarding the IT load, used to correct the energy consumption changes of auxiliary equipment such as cooling systems.

As shown in Fig. 2, the Transformer-based ECCS mapping model consists of three main stages. The input layer collects and integrates data from multiple sources, such as computing resource usage and task information, into a standardized format. The processing layer uses a Transformer structure to analyze these inputs, employing an attention mechanism to identify the complex relationships between computing tasks and equipment energy consumption. Finally, the output layer generates the prediction results, providing both the basic power required by the computing hardware and the total system energy consumption, which includes the necessary cooling power.

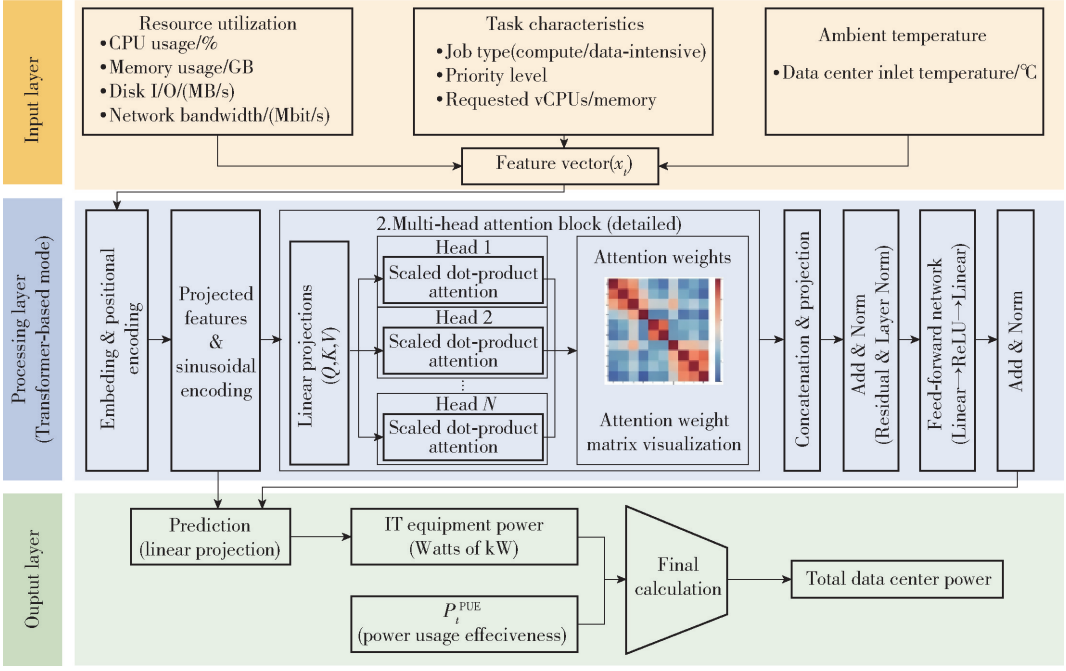

Fig. 2 Structure of the Transformer-based ECCS mapping model

To enable generalization across different computing infrastructure scenarios, the mapping model follows a three-stage adaptability framework: 1) input normalization: resource features are scaled to a unified range, allowing the model to accommodate heterogeneous hardware capacities without retraining; 2) task-type embedding: task characteristics are encoded as learnable embeddings, enabling the model to capture software-level energy signatures that vary across application scenarios; 3) transfer learning calibration: when deployed in a new data center configuration, the model can be rapidly adapted using only a small calibration dataset, significantly reducing the cost of redeployment. This framework ensures that the proposed coupling model is not tied to a specific testbed but can be flexibly applied to diverse electricity-computing integration scenarios.

3.4 Power system optimization model

The primary goal of the power system operation is to maximize the net operating profit over the scheduling horizon $\mathcal{T}$. This objective is formulated as the total electricity sales revenue derived from computing loads minus the aggregate operational costs, including thermal generation costs and energy storage degradation costs. The optimization problem is formulated as follows:

$$\max \sum_{t\in T}\left(\sum_{m\in M}\lambda_t P^C_{m,t}-\sum_{i\in G}C_i(P_{i,t})-\sum_{k\in S}C_k(P^S_{k,t})\right) \tag{4}$$

The thermal generation cost $C_i(P_{i,t})$ for unit i is modeled as a quadratic fuel cost function to reflect the non-linear relationship between fuel consumption and power output:

$$C_i(P_{i,t})=a_i(P_{i,t})^2+b_iP_{i,t}+c_i \tag{5}$$

where a_i, b_i, c_i are the cost coefficients of the unit i. The energy storage cost $C_k(P^S_{k,t})$ accounts for the degradation and maintenance expenses, which are modeled based on the power throughput:

$$C_k(P^S_{k,t})=K_{k,1}|P^S_{k,t}|+K_{k,2}E_{k,\text{rated}} \tag{6}$$

s.t.

$$P_{i,\min}u_{i,t}\leqslant P_{i,t}\leqslant P_{i,\max}u_{i,t};\quad \forall i\in G,t\in T \tag{7}$$

$$-R^{\text{down}}_i\leqslant P_{i,t}-P_{i,t-1}\leqslant R^{\text{up}}_i;\quad \forall i\in G,t\in T \tag{8}$$

$$0\leqslant P^{\text{RE}}_{j,t}\leqslant P^{\text{RE,max}}_{j,t};\quad \forall j\in R,t\in T \tag{9}$$

$$-P^{c,\max}_k\leqslant P^S_{k,t}\leqslant P^{d,\max}_k;\quad \forall k\in S,t\in T \tag{10}$$

$$E_{k,t}=E_{k,t-1}-\eta_kP^S_{k,t}\Delta t \tag{11}$$

$$E^{\min}_k\leqslant E_{k,t}\leqslant E^{\max}_k \tag{12}$$

$$\sum_{i\in G}P_{i,t}+\sum_{j\in R}P^{\text{RE}}_{j,t}+\sum_{k\in S}P^S_{k,t}=\sum_{m\in M}P^C_{m,t}+P^{\text{Base}}_t,\quad \forall t\in T \tag{13}$$

$$u_{i,t}\in\{0,1\};\quad \forall i\in G,t\in T \tag{14}$$

where $K_{k,1}$ represents the variable operation and maintenance coefficient, and $K_{k,2}$ represents the fixed capacity cost coefficient.

The constraints listed above ensure the physical security and feasibility of the system operation. Specifically, Eq. (7) defines the operational boundaries for thermal units determined by the binary commitment state $u_{i,t}$ constrained in Eq. (14). Eq. (8) restricts the output variation rate to prevent mechanical fatigue. Eq. (9) constrains renewable output based on availability. The flexibility of the energy storage system (ESS) is modeled in Eq. (10) to Eq. (12), bounding charging/discharging power and enforcing state of charge (SOC) continuity. Finally, Eq. (13) enforces the power balance between supply and demand.

3.5 Computing system optimization model

The computing system aims to maximize its total service profit. This objective reflects the active response of the computing side to the grid's price signals λ_t. The profit function is defined as the revenue from computing services minus the electricity costs and the operational wear-and-tear costs, formulated as follows:

$$\max \sum_{t\in T}\sum_{m\in M}(R^{\text{comp}}_m(t)-\lambda_tP^C_{m,t}-C^{\text{op}}_m(t)) \tag{15}$$

The service revenue $R^{\text{comp}}_m(t)$ is determined by the value of tasks executed at the current time slot:

$$R^{\text{comp}}_m(t)=\sum_{i\in J}v_ix_{i,m,t} \tag{16}$$

where v_i represents the service value of task i. The operational cost $C^{\text{op}}_m(t)$ captures the wear-and-tear caused by resource switching, modeled as a penalty on utilization fluctuation:

$$C^{\text{op}}_m(t)=\beta_m\|\boldsymbol{U}_{m,t}-\boldsymbol{U}_{m,t-1}\|^2_2 \tag{17}$$

s. t.

$$\sum_{m\in M}\sum_{t\in T} x_{i,m,t}=1;\quad \forall i\in J \tag{18}$$

$$a_i x_{i,m,t}\leqslant t x_{i,m,t}\leqslant (d_i-\tau_i)\cdot x_{i,m,t};\quad \forall i\in J, m\in M, t\in T \tag{19}$$

$$\sum_{i\in J} r_i x_i, m, t\leqslant \boldsymbol{R}_m;\quad \forall m\in M, t\in T \tag{20}$$

$$P^C_{m,t}\leqslant P^{\max}_m;\quad \forall m\in M, t\in T \tag{21}$$

$$x_{i,m,t}\in 0,1 \tag{22}$$

where $\boldsymbol{U}_{m,t}$ is the resource utilization vector and β_m is the penalty coefficient.

The scheduling strategy is strictly bounded by both quality of service (QoS) and physical hardware limitations. Specifically, Eq. (18) enforces the atomicity of task execution, ensuring that each task i is scheduled exactly once across all available nodes and time slots. Eq. (19) defines the valid time window for task processing, guaranteeing that the execution start time falls between the task's arrival time and the latest allowable start time determined by the deadline minus the duration, thus strictly satisfying the timeliness requirement. On the hardware side, Eq. (20) represents the multi-dimensional resource capacity constraint, ensuring that the aggregated demand for GPU, and memory of concurrent tasks, namely the sum of their resource vectors, does not exceed the node's physical configuration. Additionally, Eq. (21) introduces a general physical safety constraint, limiting the real-time power consumption of the computing node within its rated power cap $P^{\max}_m$ to prevent overload. Finally, Eq. (22) defines the binary nature of the decision variable.

3.6 Bi-level optimization formulation

This collaborative optimization problem is modeled as a Stackelberg game and formalized using MDP, namely a bi-level MDP. The upper-level MDP corresponds to the macro-dispatch of the power system, acting as the leader, while the lower-level MDP models the micro-scheduling of computing nodes, acting as followers.

The upper-level optimization is modeled as a tuple $M^{\text{upper}}=(S^{\text{upper}}, A^{\text{upper}}, P^{\text{upper}}, R^{\text{upper}}, \gamma^{\text{upper}})$. The state s^{upper}_t encapsulates the global operating status of the grid. It is defined as:

$$s^{\text{upper}}_t=(P_{i,t}, P^{\text{RE}}_{j,t}, E_{k,t}, P^{\text{Base}}_t, P^{C,\text{total}}_t, \lambda_{t-1}, t) \tag{23}$$

where $P_{i,t}, P^{\text{RE}}_{j,t}, E_{k,t}$ denote the status of thermal units, renewable generation, and energy storage, respectively. P^{Base}_t is the conventional base load, and $P^{C,\text{total}}_t=\sum_{m\in M} P^C_{m,t}$ represents the aggregated computing load observed by the grid. λ_{t-1} is the price history, and t encodes the time information.

The action a^{upper}_t includes the dispatch commands and price signals:

$$a^{\text{upper}}_t=(\{\Delta P_{i,t}\}_{i\in G}, \{P^S_{k,t}\}_{k\in S}, \lambda_t) \tag{24}$$

where $\Delta P_{i,t}$ represents the ramp adjustment for thermal units and is used to update the corresponding unit output, $P^S_{k,t}$ is the ESS scheduling power, and λ_t is the real-time electricity price broadcast to the lower level.

The reward r^{upper}_t is designed to maximize the net profit while strictly penalizing safety violations:

$$r^{\text{upper}}_t=\left(\sum_{m\in M}\lambda_t P^C_{m,t}-\sum_{i\in G}C_i(P_{i,t})-\sum_{k\in S}C_k(P^S_{k,t})\right)-\sum_{c\in C}\omega_c V_{\phi_i}(s^{\text{upper}}_t) \tag{25}$$

The first part represents the exact economic objective. The second term imposes soft penalties with weight ω_c for violating physical constraints defined in $V_{\phi_i}(\cdot)$.

The lower level is formulated as a Dec-POMDP, namely a decentralized partially observable MDP, where each computing node $m\in M$ acts as an independent agent.

Due to privacy and communication limits, agent m only observes local information o^m_t:

$$o^m_t=(Q^m_t, \boldsymbol{U}^m_t, \lambda_t, Z^m_t) \tag{26}$$

where Q^m_t is the length of the local task queue, $\boldsymbol{U}^m_t$ is the real-time resource utilization vector, λ_t is the price signal received from the upper level, and Z^m_t is the state of the local battery.

The agent executes a composite action $a^m_t=(x^m_t, p^{\text{ups},m}_t)$ to schedule tasks and local energy resources. Here, x^m_t determines the task allocation and $p^{\text{ups},m}_t$ controls the local storage to arbitrage against price λ_t.

The reward r^m_t balances the economic objective with service level assurance (SLA):

$$r^m_t=(R^{\text{comp}}_m(t)-\lambda_t P^C_{m,t}-C^{\text{op}}_m(t))+\beta_1\|\boldsymbol{U}^m_t\|_2-\beta_2 D^m_t \tag{27}$$

where the first part aligns with the profit maximization

goal including operational costs. The term $\beta_1 \| \boldsymbol{U}_t^m \|_2$ serves as a reward shaping term to encourage high resource utilization. The term D_t^m represents the SLA violation penalty, quantifying the number of tasks exceeding deadlines or causing queue overflows at time t.

4 Algorithm design

The electricity-computing collaborative optimization problem constitutes a Stackelberg game. This hierarchical relationship aligns with the hierarchical reinforcement learning (HRL) paradigm. However, standard multi-agent methods struggle to capture complex collaborative requirements among computing nodes. Therefore, a multi-head attention mechanism is integrated into the lower-level MAPPO framework, and ACA-PPO, namely the attention-based coordination algorithm based on PPO, is proposed. This enables computing nodes to dynamically attend to relevant peers for global perspective decision-making. The overall interaction flow is shown in Fig. 3.

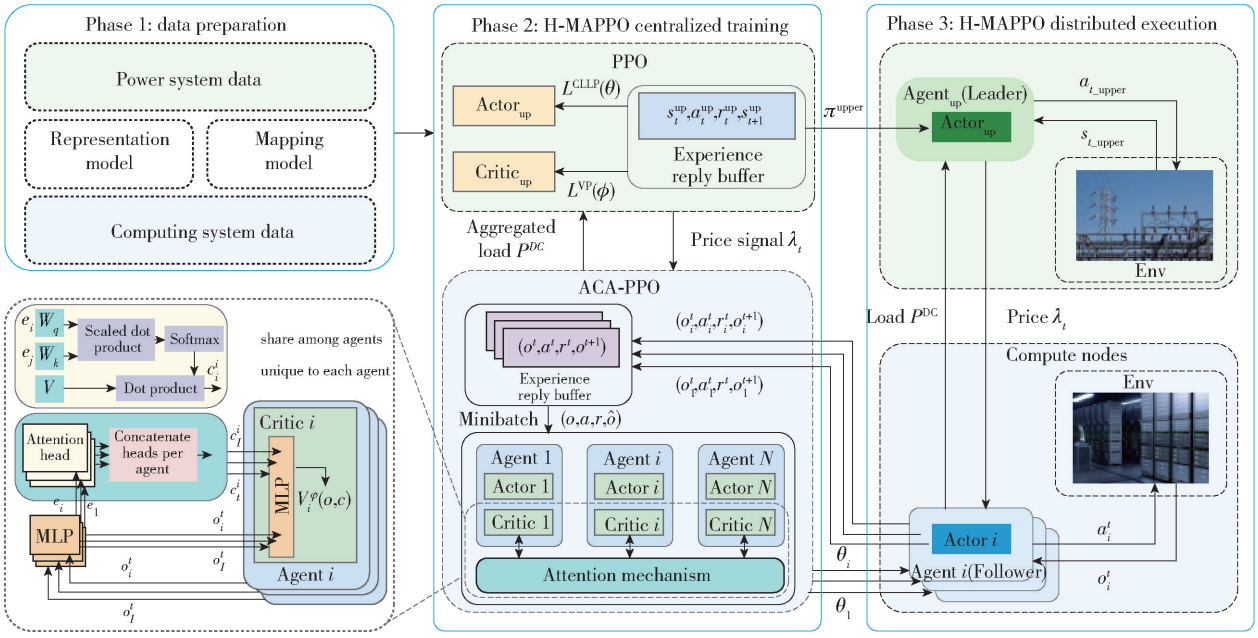

Fig. 3 Interaction flow of the H-MAPPO framework

4.1 Leader: macro-scheduling with PPO

The upper-level power system employs PPO[31] to formulate macro-economic strategies. PPO is chosen for its stability in high-dimensional continuous action spaces. Its objective function maximizes the clipped surrogate objective:

$$L(\theta) = E_t[\min(r_t(\theta)A_t, \mathrm{C}_{\mathrm{clip}}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\cdot A_t)] \tag{28}$$

where $r_t(\theta)$ is the probability ratio and A_t is the advantage estimate. As PPO is a standard baseline, its specific pseudocode is omitted, and the focus is placed on the innovative lower-level design.

4.2 Follower: micro-collaboration with ACA-PPO

The lower-level computing system employs the proposed ACA-PPO algorithm based on the CTDE paradigm. By introducing a multi-head attention mechanism[32], the decision a_t^i of agent i depends on both local observation o_t^i and a context c_t^i aggregated from other agents $a_t^i \sim \pi_{\theta_i}(\cdot \mid o_t^i, c_t^i)$. The attention entropy regularization is introduced to encourage exploration. The core update process is detailed in Algorithm 1.

Algorithm 1 ACA-PPO update module

Input: buffers B_i, network params θ_i, ϕ_i, attention weights W_{attn}

function ACA-PPO_Update ($\{B_i\}_{i=1}^N, \{\theta_i, \phi_i\}_{i=1}^N, W_{\text{attn}}$)
for $k = 1$ **to** K^{lower} **do**
  for $i = 1$ **to** M **in parallel do**
    Sample batch B_i from B_i; Calculate advantage $\hat{A}_t^i$ using (o_t^i, c_t^i)
    $L_t^{\text{CLIP}} \leftarrow \mathbb{E}_t[\min(r_t^i(\theta)\hat{A}_t^i, C_{\text{clip}}(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon)\cdot \hat{A}_t^i)]$
    $L_t^{\text{VF}} \leftarrow \mathbb{E}_t[(V_{\phi_i}(o_t^i, c_t^i) - \hat{R}_t^i)^2]$;
    $L_t^{\text{ATT}} \leftarrow \mathbb{E}_t\left[-\sum_j \alpha_t^{ij} \log \alpha_t^{ij}\right]$
    Update: $\theta_i \leftarrow \theta_i + \alpha \nabla(L_t - \lambda L_t^{\text{ATT}})$; $\phi_i \leftarrow \phi_i - \alpha \nabla L_t^{\text{VF}}$
  end for
    Update shared attention weight matrix W_{attn}
end for
return: Updated $\{\theta_i, \phi_i\}_{i=1}^N, W_{\text{attn}}$
end function

4.3 Overall training process

The H-MAPPO training follows the leader-follower protocol. In each epoch, the framework collects trajectories via environment interaction and then sequentially updates the upper and lower policies. The complete process is presented in Algorithm 2.

Algorithm 2 H-MAPPO collaborative training framework

Initialize: upper parameters Θ^{up}, lower parameters Θ_i^{low}, attention weights W_{attn}
for epoch $e = 1$ **to** E **do**
  Reset environment and buffers; Obtain initial states s_0^{up}, $\{o_0^i\}$
  for step $t = 0$ **to** $T-1$ **do**
    Upper Action: $a_t^{\text{up}} \sim \pi(s_t^{\text{up}}) \rightarrow$ Extract price λ_t
    Lower Action: $c_t^i \leftarrow \text{Attn}(\{o_t^i\})$, $a_t^i \sim \pi(o_t^i, c_t^i)$ for all i
    Execute: $(\{o_{t+1}^i, r_t^i\}, s_{t+1}^{\text{up}}, r_t^{\text{up}}) \leftarrow$ Env. step$(a_t^{\text{low}}, a_t^{\text{up}})$
    Store transitions in buffers B^{up} and $\{B_i^{\text{low}}\}$
  end for
  $\Theta^{\text{up}} \leftarrow \text{PPO}_{\text{update}}(B^{\text{up}}, \Theta^{\text{up}})$
  $\Theta_i^{\text{low}}, W_{\text{attn}} \leftarrow \text{ACA-PPO}_{\text{update}}(\{B_i^{\text{low}}\}, \Theta_i^{\text{low}}, W_{\text{attn}})$
end for

4.4 Computational complexity analysis

The computational complexity of H-MAPPO is analyzed as follows. For the upper-level PPO, the time complexity per update is $O(B_{\text{up}} d_{\text{up}})$, where B_{up} is the batch size and d_{up} is the dimension of the upper-level state-action space. For the lower-level ACA-PPO, the attention mechanism introduces additional complexity. Given M agents, each with observation dimension d_o, the multi-head attention operation costs $O(M^2 d_h H)$, where d_h is the per-head dimension and H is the number of attention heads. The overall per-update complexity is therefore $O(B_{\text{low}} M^2 d_h H + M d_o)$, where B_{low} is the lower-level batch size.

In comparison, MIQP solves a centralized optimization problem with complexity $O(N_{\text{var}}^3)$ in the worst case, where N_{var} is the number of decision variables. For our problem scale, MIQP requires 342.6 s per decision, while H-MAPPO reduces this to 2.5 s per inference. This reduction of over two orders of magnitude enables real-time scheduling applications.

For the practical application cost and scenario decomposition. Under the default configuration of 4 nodes and 500 episodes, training takes 8.6 h on an Ray Tracing Xtreme (RTX) 4090 GPU with 4.2 GB memory. Inference requires 2.5 s per step, meeting real-time requirements. Communication overhead is from 0.5 to 1.0 MB per step, acceptable in typical data center networks. For larger systems, hierarchical grouping is adopted: nodes are clustered by proximity; local coordination is handled by intra-cluster attention, while inter-cluster coordination is delegated to the upper-level agent. For bandwidth-constrained scenarios, attention reduces communication from $O(M^2)$ to $O(kM)$ with less 5% performance loss.

5 Experiments and analysis

5.1 Evaluation of recognition performance

The simulation covers a 24-hour cycle with 96 time slots, each corresponding to a 15-minute interval. The power system is modeled after a German market scenario using ENTSO-E data[34], comprising 3 thermal units whose quadratic cost coefficients a, b, c defined in Eq. (5) are listed in Table 1, wind and

photovoltaic (PV) farms, and storage. The computing system simulates 4 data centers with heterogeneous clusters, where task workloads follow Poisson arrival and exponential processing distributions derived from the Google Cluster Trace dataset[35].

Table 1 Parameters of thermal power units

Unit	a	b	c	Min/kW	Max/kW	Ramp/(kW·h^{-1})
Unit 1	0.003 4	3	30	10	150	100
Unit 2	0.001 0	10	40	50	375	100
Unit 3	0.001 0	5	70	100	500	200

To validate the effectiveness of H-MAPPO, it is compared against three representative baselines, as detailed in Table 2.

Table 2 Comparison of baseline algorithms and H-MAPPO

Algorithm	Architecture	Description
C-PPO	Single centralized agent controlling all global variables	Validate the necessity of the hierarchical structure
H-IPPO	Upper PPO and lower independent PPO agents	Validate the effectiveness of the attention mechanism
MIQP	Centralized mathematical programming using Gurobi	Provide the theoretical optimal benchmark
H-MAPPO	Upper PPO and lower collaborative ACA-PPO	Proposed method

A bi-level reward structure with penalty weights $\omega_c = \omega_D = 10$ is adopted. Two-layer MLPs with 256 neurons and ReLU activation are used by the DRL agents. Four attention heads with a dimension of 64 are employed by the ACA-PPO module. Training runs for 500 episodes with a learning rate of 1×10^{-4}, discount factor $\gamma = 0.98$, and batch size 64. The key performance indicators (KPIs) include: total system profit, task completion rate, energy efficiency, and solution time.

5.2 Performance analysis of the mapping model

Before the system-level scheduling is evaluated, the accuracy of the proposed Transformer-based coupling model is first validated. It is compared against two baselines: a linear regression model and a long short-term memory (LSTM) model. The prediction accuracy on the test set is shown in Table 3.

Table 3 Comparison of power prediction accuracy

Model	RMSE/W	MAE/W	MAPE/%
Linear PUE	154.2	120.5	12.4
LSTM	85.6	64.3	6.8
Transformer (Ours)	42.1	31.5	3.2

As observed, the traditional linear model fails to capture the non-linear fluctuations caused by heterogeneous tasks, resulting in the highest error. While LSTM improves performance, the proposed Transformer model achieves the lowest root mean square error (RMSE) and mean absolute percentage error (MAPE). This high-fidelity mapping provides a reliable foundation for the subsequent H-MAPPO optimization.

5.3 Training process analysis

Fig. 4 and Fig. 5 respectively show the reward and system operating cost curves of H-MAPPO and the two DRL baseline algorithms during the training process.

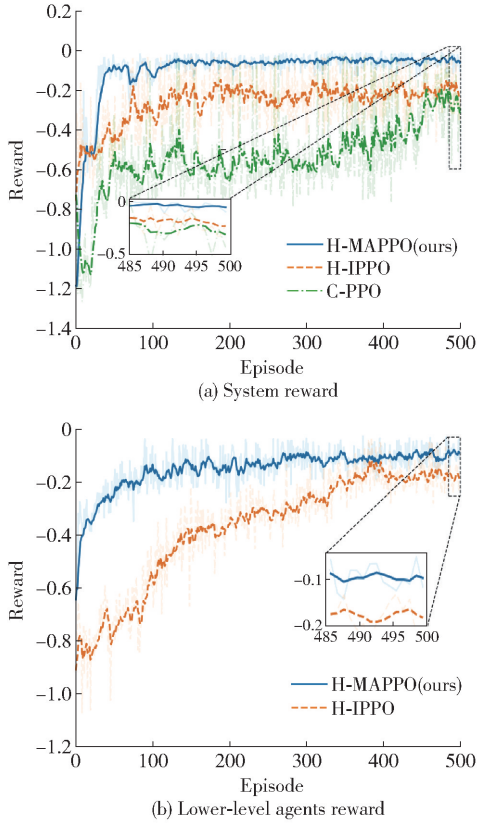

Fig. 4 Learning curves of reward for different methods

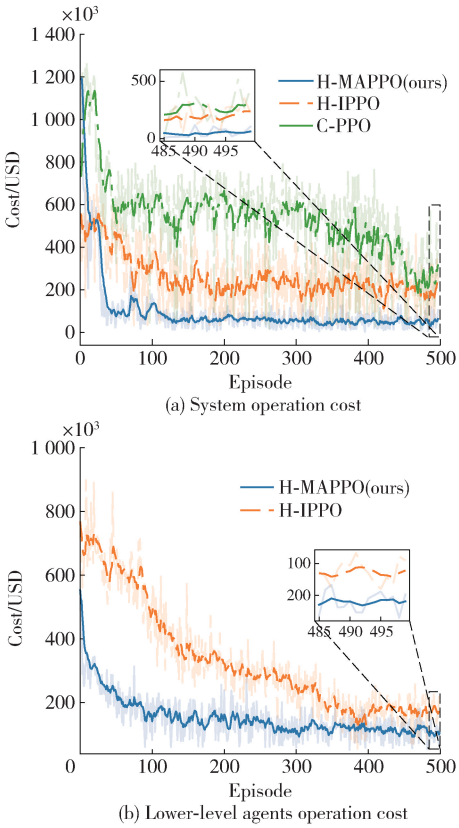

Fig. 5 Learning curves of operating cost for different methods

All curves are average results of multiple training runs, and the shaded areas represent the standard deviation range.

From the system reward subplot in Fig. 4, it can be seen that the H-MAPPO method proposed in this paper performs best in both convergence speed and final performance. Compared with C-PPO, which suffers from unstable learning and poor convergence due to the excessively large joint state-action space, the hierarchical architectures of H-MAPPO and H-IPPO demonstrate immense superiority. Among the two hierarchical methods, H-MAPPO converges faster than H-IPPO and achieves higher and more stable reward values. This preliminarily validates the effectiveness of the hierarchical collaborative architecture. In the lower-level agents reward subplot, the curve of H-MAPPO is significantly higher than that of H-IPPO throughout the process. This directly and powerfully proves the value of the ACA-PPO collaborative algorithm we designed for the lower level.

It indicates that through the attention mechanism, lower-level agents can learn to collaborate efficiently, thereby obtaining higher self-returns than independent learning and mastering local optimal strategies more quickly.

The cost curves in Fig. 5 present a mirror trend corresponding to the reward curves. H-MAPPO is able to reduce the system operating cost and lower-level agent cost to the lowest level most quickly and stabilize them there, once again proving the superiority of its learning efficiency and final performance.

5.4 Comprehensive performance analysis

To comprehensively evaluate the final performance of the algorithms on the test set, various KPIs were compared, and the results are summarized in Table 4.

Table 4 Comparison of comprehensive performance metrics

Method	Task completion rate/%	Total power/W	Total profit/USD	Energy efficiency	Load imbalance	Peak-to-valley ratio	Solution time/s
No-OP	48.27	877 482	47 429	18.5	0.066 4	1.67	0
C-PPO	49.90	895 338	48 264	18.55	0.034 7	1.86	2.3
H-IPPOt	63.14	1 198 195	66 227	18.09	0.020 6	1.63	2.7
H-MAPPO	76.78	1 126 532	77 404	16.30	0.024 8	1.47	2.5
MIQP	77.80	1 269 124	78 695	16.13	0.018 0	1.38	342.6

From Table 4, it can be seen that the H-MAPPO proposed in this paper performs best on all key benefit indicators. Its total profit (77 404) and task completion rate (76.78%) far exceed those of C-PPO and H-IPPO, and are very close to the theoretical optimal MIQP solution with the values of 78 695 and 77.80%, respectively, reaching 98.36% of the optimal solution. This result fully proves the high

quality of the learned policy of the proposed algorithm. The cumulative profit curve in Fig. 6 intuitively displays the profit growth trend of each algorithm in the scheduling process, which shows that H-MAPPO maintains a continuous lead over other comparison algorithms throughout the entire scheduling cycle.

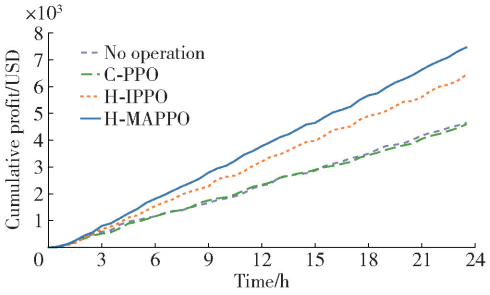

Fig. 6 Comparison of cumulative profit for different scheduling strategies

Comparing H-MAPPO with H-IPPO, the former's total profit is 16.9% higher, and the task completion rate is 21.6% higher. This fully proves the immense value of the lower-level ACA-PPO attention-based collaborative mechanism. Comparing H-MAPPO with C-PPO, the former's total profit is 60.4% higher, validating the significant superiority of the hierarchical architecture over the flat architecture when dealing with such complex problems.

It is worth noting that H-MAPPO's total power consumption is the highest among all methods. This seems contradictory, but combined with its highest task completion rate, the intrinsic logic is: H-MAPPO completes more computing tasks during periods of abundant energy through smarter scheduling, thereby bringing higher profits while the total energy consumption increases accordingly. To measure its efficiency, the energy efficiency metric, namely power consumption per unit profit, is introduced. As shown in Table 4, H-MAPPO's energy efficiency is 16.30 W per profit, which is the most efficient among all DRL methods and closest to MIQP's 16.13 W per profit. Furthermore, regarding the peak-to-valley ratio metric, H-MAPPO (1.47) also performs best among all DRL methods and is closest to MIQP (1.38). This indicates that the proposed method can smooth the load most effectively and make the greatest contribution to the stable operation of the power grid.

To more intuitively demonstrate the impact of different strategies on the load curve, Fig. 7 presents the hourly power consumption comparison of four computing nodes under different strategies.

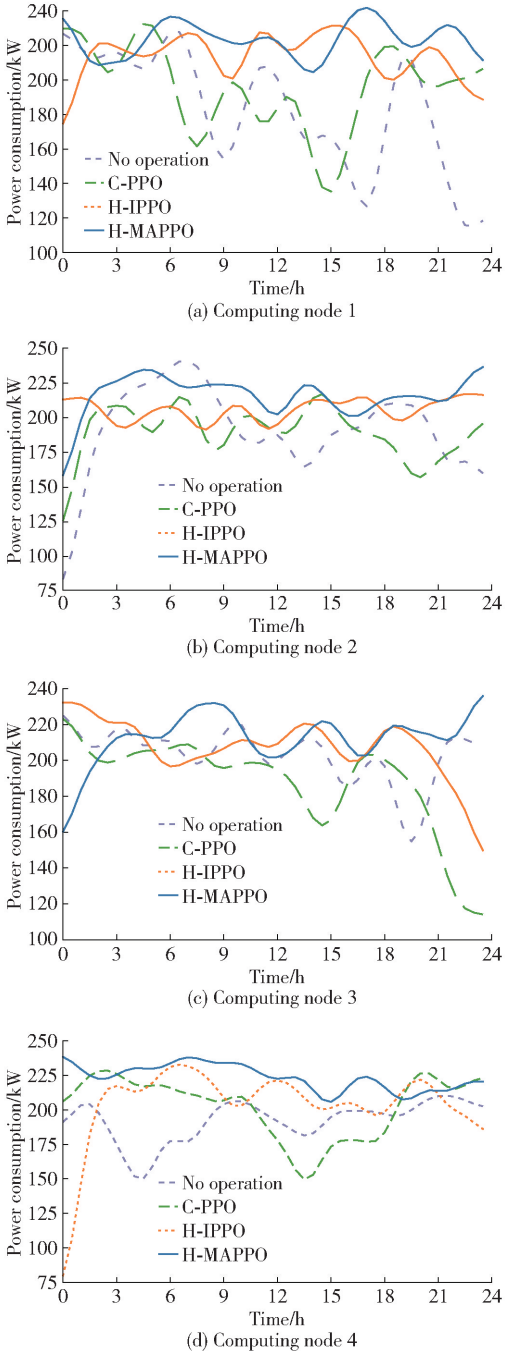

Fig. 7 Comparison of hourly power consumption of each computing node under different scheduling strategies

It can be clearly seen from the figure that the C-PPO strategy exhibits an extremely unstable power

consumption curve on all nodes, with frequent sharp peaks and valleys, reflecting the decision-making difficulty of a flat single agent in complex multi-dimensional control problems. In contrast, both hierarchical methods demonstrate smoother load scheduling capabilities. Among them, the H-MAPPO method proposed in this paper is particularly outstanding. It maintains a relatively highest and most stable power consumption level on all nodes, effectively achieving the valley filling effect. This stable high-power operating state is a direct reflection of H-MAPPO's ability to complete more computing tasks and obtain the highest profit, and it also intuitively explains why it can achieve the lowest peak-to-valley ratio.

The electricity price strategy shown in Fig. 8 also confirms that H-MAPPO has learned a dynamic pricing pattern similar to MIQP to guide the load.

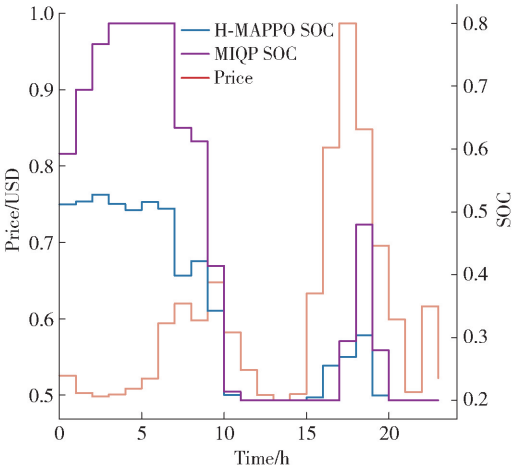

Fig. 8 Comparison of decision electricity prices between H-MAPPO and MIQP

It is also observed that H-MAPPO's load imbalance, with a value of 0.024 8, is slightly higher than that of H-IPPO, with a value of 0.020 6. This reveals an interesting trade-off: the attention-based collaborative mechanism may prioritize optimizing the overall system profit rather than ensuring absolute load balance among computing nodes, allowing some nodes to opportunistically undertake more tasks to seize low-price windows. Finally, in terms of solution time, H-MAPPO only needs 2.5 s to complete a decision, achieving an acceleration of over 100 times compared to MIQP's 342.6 s, highlighting its huge application potential in real-time dynamic decision-making scenarios. Fig. 9 shows the cost comparison between H-MAPPO and MIQP within 24 h, showing that the DRL strategy effectively fits the optimal solution in most periods.

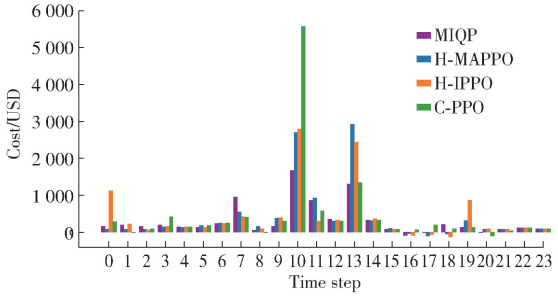

Fig. 9 Hourly comparison of total system operating cost between H-MAPPO and MIQP

5.5 Sensitivity analysis

5.5.1 Penalty weight

To verify the model's sensitivity to the penalty weights in the reward function and to demonstrate the rationality of the parameter selection in the main experiments, a sensitivity analysis on the penalty coefficient ω is conducted in this section, which uniformly representing ω_c in the upper level and ω_D in the lower level. While all other hyperparameters were kept unchanged, the algorithm performance was tested under three penalty-weight settings: 10 as the default value, 20, and 50. The experimental results are shown in Fig. 10 and Fig. 11.

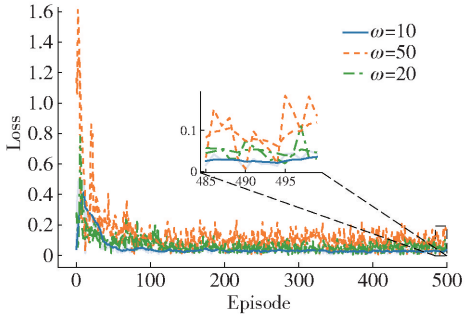

Fig. 10 Training loss curves under different ω

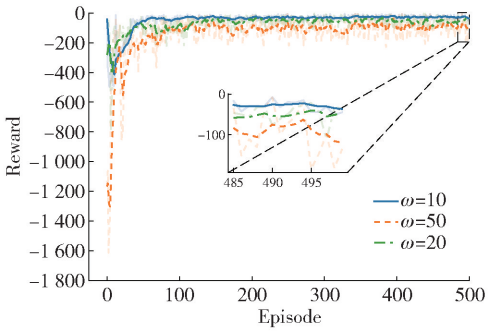

Fig. 11 Total system reward curves under different ω

In Fig. 10, when the penalty weight $\omega = 50$, the loss value in the initial stage of training is the highest and fluctuates most violently. In contrast, for $\omega = 10$, the loss function can decline most quickly and converge to a lower and more stable level. This indicates that an excessively high penalty weight causes unstable gradient updates by imposing huge penalties on violations during the exploration phase, thereby increasing the difficulty of training.

In Fig. 11, this trend is more obvious. The setting of $\omega = 50$ leads to the slowest learning speed of the agent and the worst performance of the finally converged policy. This reveals a key phenomenon: overly severe penalties make agents become overly conservative and afraid to conduct effective exploration to avoid any possibility of violating constraints, thus falling into local optima. Conversely, the setting of $\omega = 10$ performs best in both convergence speed and final reward value, indicating that this value achieves the best balance between guiding agents to comply with constraints and encouraging agents to explore better strategies.

To further verify the training results and evaluate the generalization ability of the final policy, the policy models trained under each ω setting were frozen and evaluated in 30 unseen test scenarios. Table 5 summarizes their average performance on the test set.

Table 5 Performance comparison under different penalty weights

ω	Average total profit/USD	Average completion rate/%	Average peak-to-valley ratio
10	$77\,404 \times 10^3$	76.78	1.47
20	$72\,518 \times 10^3$	73.15	1.51
50	$65\,833 \times 10^3$	68.52	1.59

The test results clearly confirm the conclusions drawn from the training curves. As shown in Table 5, the policy trained using $\omega = 10$ performs best on all key performance indicators: it achieves the highest total system profit and task completion rate, while realizing the lowest peak-to-valley ratio, indicating that its policy has the best comprehensive benefits. This quantitatively proves that an appropriate penalty weight can not only bring a more stable training process but also produce a final policy with strong generalization ability and superior practical performance.

5.5.2 Larger systems

To evaluate the impact of the multi-head attention mechanism, the number of attention heads H in ACA-PPO is varied from 1 to 8 while all other hyperparameters are kept unchanged. As shown in Table 6, when H increases from 1 to 4, the total profit improves from 72 481 to 77 404, and the task completion rate increases from 71.2% to 76.78%. This indicates that multiple attention heads effectively capture diverse coordination patterns among computing nodes. However, further increasing H to 8 yields only marginal gains while increasing training time by approximately 18%. Therefore, $H = 4$ is selected as the default setting to balance performance and computational efficiency.

Table 6 Performance under different attention head counts

Head	Average total profit/USD	Average completion rate/%	Average training time/h
1	$72\,481 \times 10^3$	71.2	7.8
2	$75\,203 \times 10^3$	74.5	8.1
4	$77\,404 \times 10^3$	76.78	8.6
8	$77\,621 \times 10^3$	77.1	10.2

To assess scalability, systems with varying numbers of computing nodes $M = 2, 4, 8, 16$ are simulated. For $M = 8$ and $M = 16$, the hierarchical grouping strategy described in Sect. 4.4 is adopted, where nodes are partitioned into clusters of size 4, full attention is used for intra-cluster coordination, and inter-cluster coordination is delegated to the upper-level agent. As shown in Table 7, H-MAPPO consistently outperforms H-IPPO across all scales. The performance gap widens as M increases, demonstrating that the attention-based

coordination mechanism becomes increasingly valuable in larger systems. The inference time increases from 1.8 s under the smaller-scale setting to 4.2 s under the larger-scale setting, which remains within 5 – 15 min real-time requirement of electricity markets.

Table 7 Scalability comparison under different numbers of computing nodes

Node	H-IPPO profit/USD	H-MAPPO profit/USD	Improvement/%	Inference time/s
2	$68\,512 \times 10^3$	$76\,198 \times 10^3$	11.2	1.8
4	$66\,227 \times 10^3$	$77\,404 \times 10^3$	16.9	2.5
8	$63\,445 \times 10^3$	$74\,128 \times 10^3$	16.8	3.3
16	$58\,923 \times 10^3$	$69\,821 \times 10^3$	18.5	4.2

In summary, through the comprehensive analysis of the training process and test performance, it can be concluded that the selected penalty weight $\omega = 10$ is robust and effective, achieving the optimal trade-off between guiding agents to learn feasible strategies and maximizing system benefits. Additionally, the sensitivity analysis on attention head count shows that $H = 4$ balances performance and training efficiency, which is adopted as the default setting.

6 Conclusions and future work

In this paper, a fine-grained coupling model and a hierarchical multi-agent DRL framework H-MAPPO are combined into the collaborative scheduling of power and computing systems. Through the comprehensive simulation experiments and comparative analysis with baseline algorithms and traditional optimization methods, the following conclusions can be drawn.

1) The task-level electricity-computing coupling model constructed in this study effectively bridges the gap between micro-level task scheduling and macro-level grid load. By establishing an end-to-end quantification channel, this model provides a high-fidelity foundation for collaborative optimization, ensuring that the scheduling decisions are both computationally feasible and grid-friendly.

2) H-MAPPO is effective in solving the complex Stackelberg game between the power grid and multiple computing nodes. By introducing the ACA-PPO algorithm with a multi-head attention mechanism, the framework significantly reduces the decision-making confusion in high-dimensional environments and achieves efficient dynamic coordination. However, the hierarchical structure and the attention mechanism introduce additional computational complexity during training, which necessitates careful hyperparameter tuning to ensure convergence stability.

3) The proposed framework demonstrates superior performance in terms of solution quality and computational efficiency. Compared with the traditional MIQP method, H-MAPPO achieves near-optimal performance, with the total profit reaching 98.36% of the theoretical optimum, while drastically reducing the decision time from 342.6 s at the minute level to 2.5 s at the second level. This acceleration of over 130 times verifies its immense potential for real-time applications in dynamic and uncertain environments.

There are several challenging aspects in future research. The integration of heterogeneous network constraints, such as topology bandwidth and latency of the computing network, into a unified optimization framework is important for achieving deep coordination between energy flow and information flow. Enhancing system robustness against extreme errors in renewable energy forecasting and cyber-attacks through robust optimization or adversarial reinforcement learning needs further study. In addition, considering that computing centers and power grids may belong to different stakeholders, the combination of federated reinforcement learning with the proposed framework is essential to achieve collaboration while protecting data privacy. Finally, scalable decentralized training paradigms, such as graph neural networks, should be explored to address the communication bottlenecks in ultra-large-scale electricity-computing networks.

Acknowledgements

This work was supported by the project of Research on Key Technologies for Multi-Agent-Driven Collaborative Scheduling of Power and Computing Resources (2025YF-105).

References

[1] VERSTRAETEN T, DAEMS P J, BARGIACCHI E, et al. Scalable optimization for wind farm control using coordination graphs. arXiv Preprint, arXiv: 2101.07844, 2021.

[2] BAHI A, OURICI A. Deep reinforcement learning for real-time green energy integration in data centers. arXiv Preprint, arXiv:

2507.21153, 2025.
[3] ZHOU Y C, WEI F C, LI S X, et al. Data center load modeling through optimal energy consumption characteristics: A path to simultaneously enhance energy efficiency and demand response quality. Applied Energy, 2025, 393(C): Article 126095.
[4] GUO H J, YU H, W M, et al. Integrated management of workloads and energy system for data centers. Energy, 2025, 327: Article 136400.
[5] GUO C Y, WANG X, ZHENG Y H, et al. Optimal energy management of multi-microgrids connected to distribution system based on deep reinforcement learning. International Journal of Electrical Power & Energy Systems, 2021, 131: Article 107048.
[6] XU W K, CHU Z D, TENG F. LAPSO: A unified optimization view for learning-augmented power system operations. arXiv Preprint, arXiv: 2505.05203, 2025.
[7] LI J H, ZHANG R C, WANG H, et al. Deep reinforcement learning for optimal power flow with renewables using spatial-temporal graph information. arXiv Preprint, arXiv: 2112.11461, 2021.
[8] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games. Advances in Neural Information Processing Systems, 2022, 35: 24611 – 24624.
[9] WANG S Y, ZHUGE C X, SHAO C F, et al. Short-term electric vehicle charging demand prediction: A deep learning approach. Applied Energy, 2023, 340: Article 121032.
[10] BOX G E P, JENKINS G M, REINSEL G C, et al. Time series analysis: forecasting and control. New York, NY, USA: John Wiley & Sons, 2015.
[11] KOOT M, WIJNHOVEN F. Usage impact on data center electricity needs: A system dynamic forecasting model. Applied Energy, 2021, 291: Article 116798.
[12] MYTTON D, ASHTINE M. Sources of data center energy estimates: A comprehensive review. Joule, 2022, 6(9): 2032 – 2056.
[13] SETYO Z G M, RIJAL H B, AQILAH N, et al. Energy efficiency measurement method and thermal environment in data centers: A literature review. Energies, 2025, 18(14): Article 3689.
[14] XU Y C, CHEN L L, LU Z H, et al. An adaptive mechanism for dynamically collaborative computing power and task scheduling in edge environment. IEEE Internet of Things Journal, 2023, 10(4): 3118 – 3129.
[15] GANDHI A, LEE D, LIU Z H, et al. Metrics for sustainability in data centers. ACM SIGENERGY Energy Informatics Review, 2023, 3(3): 40 – 46.
[16] ZHAO D M, ZHOU J T, ZHAI J d, et al. A reinforcement learning based framework for holistic energy optimization of sustainable cloud data centers. IEEE Transactions on Services Computing, 2025, 18(1): 15 – 28.
[17] RADOVANOVIC A, CHEN B, TALUKDAR S, et al. Power modeling for effective datacenter planning and compute management. IEEE Transactions on Smart Grid, 2022, 13(2): 1611 – 1621.
[18] AL-JUMAILI A H A, MUNIYANDI R C, HASAN M K, et al. Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. Sensors, 2023, 23(6): Article 2952.
[19] LATIF I, NEWKIRK A C, CARBONE M R, et al. Empirical measurements of AI training power demand on a GPU-accelerated node. arXiv Preprint, arXiv: 2412.08602, 2024.
[20] WU J J, ZENG Z, XU S Y, et al. Unified multi-party representation model based power dispatch mechanism for distributed renewable energy systems. The Journal of China Universities of Posts and Telecommunications, 2025, 32(6): 13 – 31.
[21] YANG D F, WANG X Y, SHEN R D, et al. Global optimization strategy of prosumer data center system operation based on multi-agent deep reinforcement learning. Journal of Building Engineering, 2024, 91: Article 109519.
[22] MOHINI C, SHUBHAM G, DHANASHREE J. Optimizing power system observability using GAMS. Proceedings of the IEEE PES 14th Asia-Pacific Power and Energy Engineering Conference (APPEEC'22), 2022, Nov 20 – 23, Melbourne, Australia. Piscataway, NJ, USA: IEEE, 2022: 6p.
[23] KAPOOR A, FREED B, CHOSET H, et al. Assigning credit with partial reward decoupling in multi-agent proximal policy optimization. arXiv Preprint, arXiv: 2408.04295, 2024.
[24] KONG H, XING Q L, WANG Q, et al. ADAC: Actor-double-attention-critic for multi-agent cooperation in mixed cooperative-competitive environments. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(7): 9579 – 9592.
[25] ZHANG Y B, CHEN X, GONG S M, et al. Collective large-scale wind farm multivariate power output control based on hierarchical communication multi-agent proximal policy optimization. Renewable Energy, 2023, 219: Article 119479.
[26] KUMAR H, KOPPEL A, RIBEIRO A. On the sample complexity of actor-critic method for reinforcement learning with function approximation. Machine Learning, 2023, 112(7): 2433 – 2467.
[27] DUAN J L, GUAN Y, LI S E, et al. Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors. IEEE Transactions on neural networks and learning systems, 2022, 33(11): 6584 – 6598.
[28] GU Y, CHENG Y H, CHEN C L P, et al. Proximal policy optimization with policy feedback. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4600 – 4610.
[29] JUNG S W, AN Y Y, SUH B K, et al. Multi-agent deep reinforcement learning for scheduling of energy storage system in microgrids. Mathematics, 2025, 13(12): Article 1999.
[30] OROOJLOOY A, HAJINEZHAD D. A review of cooperative multi-agent deep reinforcement learning. Applied Intelligence, 2023, 53(11): 13677 – 13722.
[31] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms. arXiv Preprint, arXiv: 1707.06347, 2017.
[32] XIAO B C, RAMASUBRAMANIAN B, POOVENDRAN R. Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning. Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS'22), 2022, May 9 – 13, Online. 2022.
[33] ZHAN X Y, ZHU X Y, CHENG P, et al. Data center cooling system optimization using offline reinforcement learning. arXiv Preprint, arXiv: 2501.15085, 2025.
[34] HIRTH L, MÜHLENPFORDT J, BULKELEY M. The ENTSO-E transparency platform: A review of Europe's most ambitious electricity data platform. Applied energy, 2018, 225: 1054 – 1067.
[35] REISS C, WILKES J, HELLERSTEIN J L. Google cluster-usage traces: format + schema. White Paper. Google Inc, 2011.

(Editor: Luo Lang)