EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Grid-Interactive Hyperscale Data Centers: Deep Reinforcement Learning for Joint Workload–Cooling Scheduling to Enable Demand Response and Renewable Integration

Youchun Qiu[1,*]

[1]Luzhou Vocational & Technical College, Sichuan 646000, China

Abstract

Driven by artificial intelligence and cloud computing, hyperscale data centers are becoming one of the fastest-growing electrical loads worldwide and are increasingly recognized as a new class of flexible loads capable of supporting demand response (DR) and the integration of variable renewable energy (VRE). However, their two principal control levers—IT workload scheduling and cooling system operation—have traditionally been managed in a decoupled manner, leaving both energy efficiency and demand-side flexibility under-exploited. This paper proposes a deep reinforcement learning (DRL) framework that jointly co-schedules computing and thermal resources so that a hyperscale data center can operate as a grid-interactive flexible load. We formulate the joint problem as a constrained Markov Decision Process and develop an actor-critic algorithm combining Deep Deterministic Policy Gradient with a safety shield mechanism to guarantee thermal constraint satisfaction during both training and deployment. A high-fidelity digital twin simulation environment enables safe Sim-to-Real training. Extensive experiments demonstrate that the proposed approach reduces total electricity consumption by 18-25% compared to baseline controllers, cuts thermal violations by over 90%, and maintains service level agreement compliance, while broadening the controllable power envelope of the facility to provide a technical basis for participating in DR programs and aligning data-center power profiles with renewable generation. The framework bridges IT-side and facility-side control and supports the evolution of hyperscale data centers from passive electricity consumers toward active, grid-interactive participants in renewable-penetrated power systems.

Received on 20 August 2025; accepted on 01 May 2026; published on 11 August 2026

Keywords: Data Center Flexibility, Demand Response, Renewable Energy Integration, Grid-Interactive Operation, Deep Reinforcement Learning, Joint Workload–Cooling Scheduling

doi:10.4108/ew.14440

1. Introduction

The proliferation of artificial intelligence (AI) and cloud computing has triggered an unprecedented surge in computational demand, positioning data centers as critical infrastructure for the digital economy. According to Masanet et al. [1], global data center electricity consumption reached approximately 205 terawatt-hours (TWh) in 2018, representing about 1% of worldwide electricity use. Recent comprehensive reviews on dynamic thermal environment management technologies highlight the rapid growth of data center energy demands and the critical need for advanced cooling solutions [2]. This exponential growth presents both environmental and economic challenges, as data centers become significant contributors to global carbon emissions. Beyond their role as computing infrastructure, hyperscale data centers are now widely regarded as a new class of large-scale, controllable electrical loads within modern power systems. Their dual control levers—computing workload placement and cooling setpoints—endow them with intrinsic demand-side flexibility, positioning them as natural candidates for participating in demand response programs and for absorbing variable renewable energy. Unlocking this flexibility, however, requires moving beyond the traditional decoupled control of IT and cooling subsystems.

*Corresponding author. Email: qiu_tea@163.com

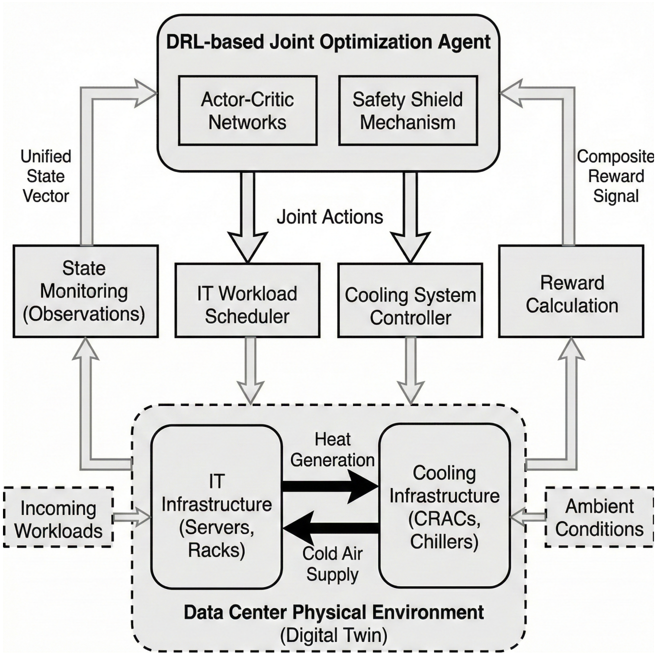

Figure 1. System architecture of the proposed DRL-based joint optimization framework. The agent observes states from both IT and cooling systems and outputs coordinated control actions

Within a typical hyperscale data center, the cooling infrastructure consumes approximately 30-40% of total energy [3]. The Power Usage Effectiveness (PUE), defined as the ratio of total facility power to IT equipment power, remains at 1.5 or higher for many legacy data centers, indicating that substantial energy is wasted on cooling and other auxiliary systems. While industry leaders like Google and Microsoft have achieved PUE values approaching 1.1 through advanced cooling technologies [4], the majority of existing facilities operate far below this efficiency frontier. This gap represents a significant opportunity for energy optimization through intelligent control strategies.

The fundamental challenge in data center energy optimization lies in the decoupled control paradigm that separates IT workload management from cooling system operation [5]. Traditional IT schedulers focus primarily on Quality of Service (QoS) metrics such as throughput and latency, without considering the thermal implications of their decisions. These schedulers typically employ round-robin, first-fit, or best-fit algorithms that optimize for computational efficiency while ignoring the spatial distribution of heat generation. Conversely, cooling controllers employ conservative strategies based on predefined temperature setpoints, often leading to over-cooling that wastes substantial energy. Facility managers typically configure Computer Room Air Conditioning (CRAC) units to maintain uniform supply temperatures across the data hall, regardless of actual thermal loads. This separation prevents the exploitation of synergies between workload placement and thermal management. Crucially, this decoupled paradigm not only sacrifices energy efficiency but also masks the data center's potential as a grid-interactive flexible load, since neither subsystem alone exposes a coordinated, predictable power envelope to the grid.

Recent advances in deep reinforcement learning (DRL) have demonstrated remarkable success in controlling complex physical systems without requiring explicit mathematical models [6]. The seminal work by DeepMind showed that DRL can achieve superhuman performance in game playing and subsequently demonstrated its applicability to real-world control problems. In the context of data center optimization, Wan et al. proposed SafeCool, a model-based reinforcement learning approach for safe and energy-efficient cooling management [7]. Building on this success, Li et al. [8] proposed an end-to-end DRL framework for data center cooling optimization, achieving 11% energy savings in simulation studies. Recent work has explored event-driven approaches [9] and digital twin frameworks combined with machine learning for thermal modeling [10]. However, most existing approaches focus solely on cooling control without jointly optimizing IT workload scheduling, leaving significant optimization potential unexplored.

In this paper, we propose a comprehensive DRL framework for joint optimization of IT workload scheduling and cooling system control, as illustrated in Figure 1. Our approach treats the data center as an integrated cyber-physical system where computational workloads and thermal dynamics are intrinsically coupled. The main contributions of this work are fourfold. First, we develop a unified mathematical model that captures the complex interactions between IT workloads, server thermal dynamics, and cooling system behavior in hyperscale data centers. This model incorporates heat recirculation effects through a thermal influence matrix that captures the spatial coupling between servers. Second, we formulate the joint optimization problem as a constrained Markov Decision Process (MDP) and design a reward function that balances energy efficiency, thermal safety, and SLA compliance. Third, we propose a DRL algorithm based on the actor-critic architecture with a safety shield mechanism that ensures constraint satisfaction during both training and deployment. The safety shield employs a lightweight thermal predictor to filter potentially dangerous actions before execution. Fourth, we construct a high-fidelity digital twin simulation environment and demonstrate through extensive experiments that our approach achieves 18-25% energy savings compared to state-of-the-art baselines while maintaining thermal safety.

2. Related Work

Data center energy optimization has attracted significant research attention over the past decade, driven by both economic and environmental concerns. Traditional approaches to cooling optimization rely on model predictive control (MPC) or rule-based strategies. Recent work by Lu et al.

[11] proposed an alternative reinforcement learning control strategy for data center air-cooled HVAC systems. These methods typically require accurate physical models of thermal dynamics, which are difficult to obtain due to complex airflow patterns, heat recirculation effects, and time-varying environmental conditions. Moore et al. [12] pioneered the use of computational fluid dynamics (CFD) simulations to predict thermal behavior in data centers, but the computational cost of CFD models limits their applicability for real-time control. Li et al. [13] developed a data-driven subspace predictive control method for air-cooled data center thermal modeling and optimization.

Machine learning approaches have gained traction as alternatives to physics-based models. Zohdi [10] developed a digital-twin and machine-learning framework for precise heat and energy management, showing that data-driven models can achieve accuracy comparable to CFD simulations at a fraction of the computational cost. However, supervised learning approaches require labeled training data that may not capture all operating conditions, and they cannot adapt to changing system dynamics. Reinforcement learning addresses these limitations by learning control policies through direct interaction with the environment. Li et al. [8] proposed an end-to-end DRL algorithm for cooling control based on the Deep Deterministic Policy Gradient (DDPG), achieving 11% cooling cost savings on the EnergyPlus simulation platform. Their approach demonstrated the feasibility of using continuous-action RL for HVAC control, but focused exclusively on cooling setpoints without considering workload scheduling. Mahbod et al. [14] achieved 15-20% energy savings using model-free reinforcement learning with careful reward shaping.

Energy-aware workload scheduling has been extensively studied in cloud computing and high-performance computing contexts. Beloglazov et al. [15] proposed virtual machine consolidation techniques that reduce the number of active servers to minimize energy consumption. Their work demonstrated that intelligent workload placement can achieve significant energy savings, but did not consider the thermal implications of consolidation. When multiple virtual machines are consolidated onto a single server, the resulting heat concentration can create thermal hotspots that require increased cooling effort. Zhang et al. [16] conducted a comprehensive review of cooling technologies for data centers, identifying the need for integrated approaches that consider both IT and cooling systems. Their analysis showed that the potential for energy savings is greatest when workload scheduling and cooling control are optimized jointly. Recent work on LLM workload scheduling [17] addresses the unique challenges posed by AI training workloads.

The joint optimization of IT workload scheduling and cooling control has received increasing attention in recent years. Chi et al. [18] proposed a multi-agent deep reinforcement learning approach where separate agents control workload distribution and cooling setpoints, coordinated through a shared reward function. Their experiments showed that joint optimization achieves 15% greater energy savings than optimizing either subsystem independently. Wang et al. [19] developed a physics-guided safe reinforcement learning framework that incorporates thermal constraints directly into the learning algorithm. Their approach uses a graph neural network to model thermal dependencies between servers, enabling the agent to reason about heat recirculation effects. Biemann et al. [20] presented a reinforcement learning framework for HVAC control that achieved significant energy reductions. Mu et al. [21] developed a large-scale data center cooling control system via sample-efficient reinforcement learning. The systematic review by Kahil et al. [22] analyzed studies on reinforcement learning for data center energy optimization, concluding that the field is maturing rapidly, but significant challenges remain in deployment to production systems.

Several recent works have addressed the challenge of safe deployment and optimization. Zhao et al. [23] proposed an optimal data center energy management approach using hybrid quantum-classical methods. Yang et al. [24] proposed deep reinforcement learning techniques for joint optimization of data center operations. Rostami et al. [25] developed thermal-aware workload distribution methods that account for equipment reliability and temperature variation costs. Hogade et al. [26] presented a game-theoretic framework for geo-distributed data centers that considers both energy costs and network transfer costs. Wang et al. [27] addressed safe reinforcement learning with thermal constraint guarantees. Mu et al. [28] explored multi-fidelity model-based reinforcement learning for adaptive cooling, while Ji et al. [29] introduced transformer-based predictive control. Zhu et al. [30] demonstrated digital twin-enabled reinforcement learning for real-time optimization, and Liao et al. [31] addressed edge data center energy cost saving. Liu et al. [32] developed carbon-aware scheduling for sustainable operations.

Our work differs from prior approaches in several key aspects. First, we explicitly model the thermal influence matrix to capture heat recirculation effects, enabling the agent to learn optimal workload placement patterns that minimize cooling requirements. Second, we incorporate a safety shield mechanism that guarantees thermal constraint satisfaction during both training and deployment, addressing a critical barrier to real-world adoption. Third, we employ a Sim-to-Real training strategy with behavioral cloning from historical operational data, providing a safe pathway from simulation to production deployment.

3. System Model

We consider a typical air-cooled data center consisting of N servers organized in racks and rows, and M Computer Room Air Conditioning (CRAC) units. The data center follows the standard hot-aisle/cold-aisle layout where server fronts face cold aisles supplied with chilled air, and server backs exhaust hot air into hot aisles that return to the CRAC units. The

Table 1. Summary of Notation

Symbol	Description
N, M	Number of servers and CRAC units
$P_{i,t}^{IT}$	IT power consumption of server i at time t (W)
$u_{i,t}$, $f_{i,t}$	CPU utilization and frequency of server i
$T_{i,t}^{in}$, $T_{i,t}^{out}$	Inlet and outlet temperature of server i (°C)
T_t^{sup}, V_t^{air}	CRAC supply temperature and airflow rate
$\mathbf{D}$	Thermal influence matrix ($N \times N$)
λ_t	Workload arrival rate at time t
COP	Coefficient of Performance of cooling system
P_{idle}, P_{peak}	Server idle and peak power consumption
ρ, C_p	Air density and specific heat capacity

system operates in discrete time steps $t \in \{0, 1, \ldots, T\}$ with an interval Δt (typically 5-15 minutes), which balances the need for responsive control against the thermal time constants of the physical system. Table 1 summarizes the key notation used throughout this paper.

External workloads arrive according to a non-homogeneous Poisson process with rate λ_t, exhibiting the diurnal patterns typical of cloud services with peak loads during business hours and reduced activity during nights and weekends. Each job k is characterized by a tuple $\langle r_k, d_k, c_k \rangle$, where r_k denotes resource requirements (CPU cores, memory), d_k represents the deadline or latency sensitivity, and c_k indicates the computational complexity. The workload scheduler must allocate incoming jobs to servers while balancing multiple objectives, including response time, energy consumption, and thermal constraints.

The instantaneous power consumption of server i is modeled as a function of CPU utilization $u_{i,t}$ and operating frequency $f_{i,t}$. Following the widely-adopted linear power model validated by extensive empirical studies [15], we express server power as:

$$P_{i,t}^{IT}(f_{i,t}, u_{i,t}) = P_{idle} + u_{i,t} \cdot (P_{peak}(f_{i,t}) - P_{idle}) \quad (1)$$

where P_{idle} is the idle power consumption representing the baseline energy draw of an active but unloaded server, and $P_{peak}(f_{i,t})$ is the peak power at operating frequency $f_{i,t}$. Modern servers support Dynamic Voltage and Frequency Scaling (DVFS), which allows the operating frequency to be adjusted based on workload demand. Based on CMOS circuit physics, dynamic power scales cubically with frequency: $P_{peak}(f) = \kappa \cdot f^3 + C$, where κ and C are hardware-specific constants that can be determined through power profiling [33]. This relationship creates opportunities for energy savings by running workloads at reduced frequencies when deadlines permit. Recent work has also explored cooperative operation optimization methods for data center alliances considering diversified flexibility [34].

The thermal dynamics of the data center are governed by heat transfer principles. According to the first law of thermodynamics, the heat generated by server i must be carried away by the airflow passing through its chassis. Under steady-state conditions, the outlet temperature is given by:

$$T_{i,t}^{out} = T_{i,t}^{in} + \frac{P_{i,t}^{IT}}{\rho \cdot C_p \cdot V_{i,t}^{flow}} \quad (2)$$

where $\rho \approx 1.2$ kg/m^3 is the air density at typical data center conditions, $C_p \approx 1005$ J/(kg·K) is the specific heat capacity of air, and $V_{i,t}^{flow}$ is the volumetric airflow rate through the server. This equation shows that higher-power servers or reduced airflow result in a greater temperature rise across the server.

A critical challenge in data center thermal management is heat recirculation, where hot exhaust air from some servers re-enters the inlet of other servers rather than returning directly to the CRAC units. This phenomenon creates thermal hotspots that can cause equipment failures and reduce computational performance through thermal throttling [35]. We model this coupling using a thermal influence matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ [12]:

$$\mathbf{T}_t^{in} = T_t^{sup} \cdot \mathbf{1} + \mathbf{D} \cdot \mathbf{P}_t^{IT} \quad (3)$$

where D_{ij} represents the thermal influence coefficient from server j to server i, quantifying how much of the heat generated by server j contributes to the inlet temperature of server i. The matrix $\mathbf{D}$ captures the complex airflow patterns in the data center, including recirculation through gaps in floor tiles, over-rack airflow, and mixing in the hot aisle. In our DRL approach, we do not explicitly estimate $\mathbf{D}$; instead, the neural network implicitly learns this nonlinear mapping from historical sensor data, enabling adaptation to facility-specific airflow patterns.

The cooling system removes heat from the data center to maintain safe operating temperatures. The total cooling power P_t^{cool} consists of fan power for air circulation and chiller power for heat rejection:

$$\begin{aligned} P_t^{cool} &= P_t^{fan} + P_t^{chiller} \\ &= \eta_{fan} \cdot (V_t^{air})^3 + \frac{\sum_{i=1}^{N} P_{i,t}^{IT}}{COP(T_t^{sup}, T_t^{amb})} \end{aligned} \quad (4)$$

Following the fan affinity laws, fan power scales cubically with airflow rate, making airflow reduction a powerful lever for energy savings. The chiller power depends on the cooling load and the Coefficient of Performance (COP), which itself depends on the supply temperature setpoint T_t^{sup} and ambient outdoor conditions T_t^{amb}. Higher supply temperatures improve COP according to the Carnot efficiency relationship, thereby reducing chiller energy consumption [36]. However, this creates a fundamental trade-off: raising supply temperature saves cooling energy but increases the risk of thermal violations, particularly for servers in locations with poor airflow.

4. Problem Formulation

We formulate the joint optimization problem as a constrained Markov Decision Process defined by the tuple

$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. The MDP formulation provides a principled framework for sequential decision-making under uncertainty, enabling the agent to learn optimal control policies through interaction with the environment.

The state vector $s_t \in \mathcal{S}$ must capture all information necessary for optimal decision making. Based on the system model, we define:

$$s_t = \left[\mathbf{u}_t, \mathbf{T}_t^{in}, T_t^{sup}, V_t^{air}, T_t^{amb}, Q_t, h_t, P_{t-1}^{total}\right] \quad (5)$$

where $\mathbf{u}_t = [u_{1,t}, \ldots, u_{N,t}]^\top$ is the vector of CPU utilizations across all servers, $\mathbf{T}_t^{in}$ is the vector of inlet temperatures, Q_t represents the pending job queue length, h_t encodes the hour of day to capture diurnal patterns, and P_{t-1}^{total} provides feedback on the previous power consumption. For large-scale data centers with thousands of servers, directly processing high-dimensional temperature vectors becomes computationally challenging. We address this by employing a convolutional neural network (CNN) to extract spatial features from the thermal map, treating the server layout as a 2D grid image where pixel intensities represent temperatures, similar to the machine learning approach for thermal modeling proposed by Zohdi [10]. This approach enables the agent to recognize spatial patterns such as hotspots and temperature gradients.

The action vector $a_t \in \mathcal{A}$ comprises both IT and cooling control variables:

$$a_t = \left[\mathbf{w}_t, \Delta T_t^{sup}, \Delta V_t^{air}\right] \quad (6)$$

where $\mathbf{w}_t \in \mathbb{R}^N$ represents workload allocation weights determining how incoming jobs are distributed across servers, and ΔT_t^{sup}, ΔV_t^{air} are incremental adjustments to cooling setpoints. We use incremental actions rather than absolute values to improve control stability and reduce the action space magnitude [37]. The workload allocation weights are passed through a softmax function to ensure they sum to one, and jobs are distributed proportionally to these weights among servers with sufficient available capacity.

The optimization objective is to minimize total energy consumption while maintaining service level agreements and thermal safety constraints:

$$\min_{\pi} J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} \gamma^t \cdot P_t^{total}\right] \quad (7)$$

subject to thermal safety constraints requiring that server inlet temperatures remain below the ASHRAE-recommended threshold: $T_{i,t}^{in} \leq T_{safe} = 27°\text{C}$ for all servers and time steps. Additionally, the system must maintain sufficient processing capacity to handle incoming workloads without excessive queueing: $\sum_{i=1}^{N} \mu_i(t) \geq \lambda_t$, where $\mu_i(t)$ is the processing rate of server i. Control variables must also respect physical bounds on supply temperature ($15°\text{C} \leq T_t^{sup} \leq 25°\text{C}$) and airflow rate ($0 \leq V_t^{air} \leq V_{max}$).

We design the immediate reward r_t to reflect the optimization objective while incorporating soft constraint penalties that guide the agent away from constraint violations:

$$\begin{aligned} r_t = -\Big(\alpha \cdot \frac{P_t^{total}}{P_{base}} + \beta \cdot \sum_{i=1}^{N} \max(0, T_{i,t}^{in} - T_{safe})^2 \\ + \xi \cdot \frac{Q_t}{Q_{max}}\Big) \end{aligned} \quad (8)$$

where P_{base} is a normalization constant based on baseline power consumption, and α, β, ξ are weighting coefficients that balance energy efficiency against thermal safety and SLA compliance. The quadratic penalty for thermal violations ensures that more severe overheating receives exponentially larger penalties, strongly discouraging the agent from approaching dangerous operating regions. The queue penalty encourages the agent to maintain sufficient processing capacity.

5. Proposed Method

We propose an actor-criticbased DRL algorithm with three key components: a feature extraction network for processing high-dimensional state inputs, an actor-critic architecture with twin critics for stable policy optimization, and a safety shield mechanism for constraint satisfaction. Figure 2 illustrates the overall framework, and Algorithm 1 presents the complete training procedure.

The state encoder processes heterogeneous inputs using a hybrid neural network architecture. A convolutional network processes the 2D thermal map to extract spatial features such as hotspot patterns and temperature gradients. The CNN consists of three convolutional layers with 32, 64, and 128 filters respectively, each followed by batch normalization and ReLU activation. The convolutional features are flattened and concatenated with scalar inputs (ambient temperature, queue length, time features) processed through a separate multi-layer perceptron. The combined representation passes through two fully-connected layers with 256 and 128 units to produce the final state encoding $\phi(s_t)$.

The actor network $\pi_\theta(s)$ maps states to deterministic actions, outputting both continuous cooling setpoint adjustments and workload allocation weights. For continuous actions, we use tanh activation scaled to the appropriate ranges. For workload allocation, we apply softmax normalization to ensure the weights sum to one. During training, we add Ornstein-Uhlenbeck noise for exploration, with the noise scale decaying over the course of training to enable fine-tuning of the policy [37].

The critic networks $Q_{\omega_1}(s, a)$ and $Q_{\omega_2}(s, a)$ estimate the expected cumulative reward for state-action pairs. Following the Twin Delayed DDPG (TD3) algorithm [38], we maintain two critics and use the minimum of their estimates to compute target values, mitigating the overestimation bias that can destabilize training. The critics share the state encoder with the actor but have separate output heads, each consisting of two fully-connected layers that process the concatenation of state features and actions.

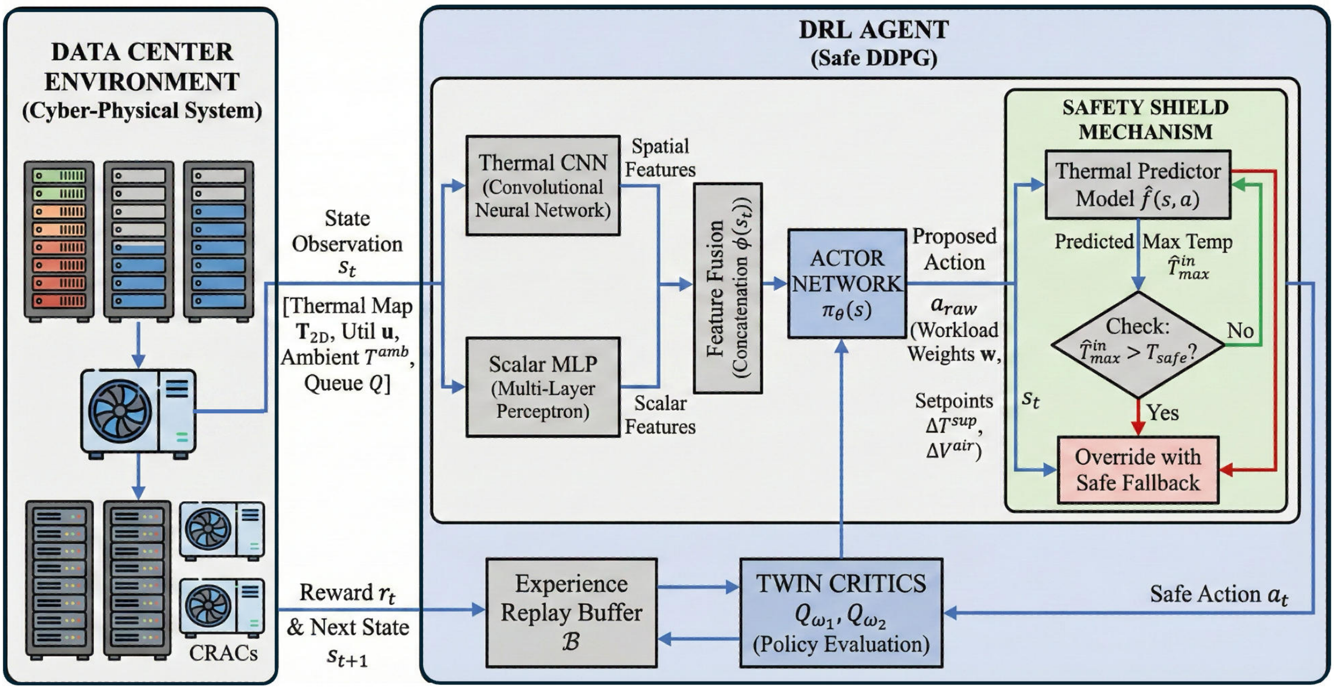

Figure 2. Framework of the proposed Safe DDPG algorithm. The thermal CNN processes spatial temperature maps while the scalar MLP handles other state features. The safety shield filters actions through a thermal predictor

Algorithm 1 Safe DDPG for Joint DC Optimization

Initialize actor π_θ, critics $Q_{\omega_1}, Q_{\omega_2}$
Initialize target networks $\pi_{\theta'}, Q_{\omega'_1}, Q_{\omega'_2}$
Initialize replay buffer $\mathcal{B}$, thermal predictor $\hat{f}$
Pre-train with behavioral cloning: $\theta \leftarrow \arg\min \|\pi_\theta(s) - a_{expert}\|^2$
for episode $= 1$ to E **do**
  Reset environment, observe initial state s_0
  for $t = 0$ to T **do**
    Select action $a_t = \pi_\theta(s_t) + \mathcal{N}_t$
    Predict temperatures: $\hat{\mathbf{T}}^{in}_{t+1} = \hat{f}(s_t, a_t)$
    if $\max_i \hat{T}^{in}_{i,t+1} > T_{safe} - \delta$ **then**
      Override: $a_t \leftarrow a^{safe}_t$
    end if
    Execute a_t, observe r_t, s_{t+1}
    Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{B}$
    Sample mini-batch from $\mathcal{B}$
    Compute target: $y = r + \gamma \min_k Q_{\omega'_k}(s', \pi_{\theta'}(s'))$
    Update critics by minimizing $(Q_{\omega_k}(s,a) - y)^2$
    Update actor by maximizing $Q_{\omega_1}(s, \pi_\theta(s))$
    Soft update targets with τ
  end for
end for

To ensure thermal constraint satisfaction, we implement a safety shield that filters potentially dangerous actions before execution. The shield uses a lightweight thermal predictor $\hat{f}$ trained on historical data to estimate next-step temperatures given the current state and proposed action. If any server is predicted to exceed the safety threshold minus a conservative margin $\delta = 2°$C, the action is overridden with a safe fallback that reduces supply temperature and increases airflow. The thermal predictor is implemented as a small neural network with two hidden layers, trained separately from the RL agent using supervised learning on historical temperature transitions.

Training RL agents directly on production data centers is infeasible due to safety risks and the potential for equipment damage during exploration. We adopt a Sim-to-Real transfer approach consisting of four stages. First, we construct a high-fidelity simulation environment using EnergyPlus [36] integrated with a custom workload generator. The simulator parameters are calibrated using historical operational data including temperature measurements, power consumption, and workload traces. Second, we pre-train the actor network using behavioral cloning from logs of the existing PID-based controller, providing a reasonable initialization that satisfies basic operational constraints [39]. Third, the pre-trained agent undergoes extensive training in the simulator, experiencing millions of time steps including stress-test scenarios such as sudden load spikes, cooling equipment failures, and extreme weather conditions. Fourth, the trained policy is deployed on the production system with the safety shield active, conservative exploration disabled, and human operators monitoring for anomalies.

Table 2. Simulation Parameters

Parameter	Value	Description
N	1000	Number of servers
M	10	Number of CRAC units
P_{idle}	150 W	Server idle power
P_{peak}	450 W	Server peak power
T_{safe}	27°C	ASHRAE safety threshold
T_t^{sup} range	[15, 25]°C	Supply temperature bounds
Δt	5 min	Control time step
T	288 steps	Episode length (24 hours)
γ	0.99	Discount factor
τ	0.005	Soft update coefficient
Learning rate	3×10^{-4}	Adam optimizer
Batch size	256	Mini-batch size

6. Experiments

We evaluate our approach through extensive experiments on a simulated data center environment calibrated using real operational data. The simulation environment is implemented using EnergyPlus for thermal dynamics integrated with a custom discrete-event simulator for workload generation and server power modeling. The simulated data center contains $N = 1000$ servers organized in 50 racks across 5 rows, with $M = 10$ CRAC units providing cooling. Key simulation parameters are summarized in Table 2.

The workload traces are generated following real-world diurnal patterns observed in cloud data centers, with average utilization of 40% during off-peak hours rising to 80% during peak periods. The arrival rate follows a sinusoidal pattern with a period 24 hours, superimposed with random perturbations drawn from a Poisson distribution. Ambient temperature varies according to typical daily patterns ranging from 15°C at night to 35°C during summer afternoons. We train each method for 2000 episodes and evaluate on 100 held-out test episodes with different random seeds.

We compare our approach against four baseline methods representing the spectrum from simple rule-based control to state-of-the-art learning-based approaches: (1) **Static Setpoint** maintains fixed supply temperature at 22°C and 80% fan speed, representing conservative ASHRAE-compliant operation; (2) **Reactive PID** implements a proportional-integral-derivative controller that adjusts cooling based on return air temperature feedback, the predominant control strategy in existing data centers [11]; (3) **Thermal-Aware Scheduling** performs workload placement considering server temperatures but uses fixed cooling setpoints [13]; and (4) **Cooling-only DRL** implements a DRL agent controlling only cooling setpoints without workload scheduling, similar to the approach of Li et al. [8].

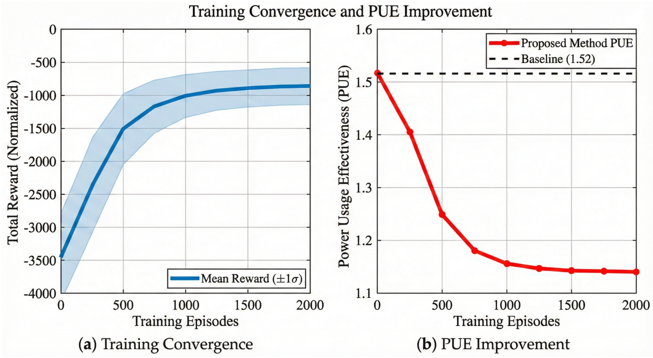

Figure 3. Training convergence over 2000 episodes. (a) Episode reward with ±1 std across 5 seeds. (b) PUE decreases from 1.52 to 1.14

Table 3. Performance Comparison on Test Episodes

Method	Energy Reduction	PUE	Thermal Violation	SLA Violation
Static Setpoint	—	1.52	0%	0.3%
Reactive PID	7.2%	1.45	0.4%	0.4%
Thermal-Aware Sched.	11.5%	1.38	1.8%	0.2%
Cooling-only DRL	15.3%	1.28	0.6%	0.3%
Proposed	**23.8%**	**1.14**	**0.1%**	**0.1%**

Figure 3 shows the training progress over 2000 episodes. The cumulative reward increases monotonically, indicating stable learning, while the Power Usage Effectiveness decreases from 1.52 (baseline) to approximately 1.14, approaching the efficiency of state-of-the-art hyperscale facilities. The behavioral cloning pre-training enables the agent to achieve reasonable performance from the first episode, avoiding the dangerous exploration that would occur with random initialization. The safety shield interventions decrease from approximately 15% of time steps in early training to less than 1% by episode 500, indicating that the agent learns to avoid constraint violations.

Table 3 summarizes the performance comparison across all methods. Our proposed joint optimization approach achieves 23.8% energy reduction compared to the Static Setpoint baseline and 8.5% improvement over the Cooling-only DRL method. The additional gains come from thermal-aware workload placement that concentrates jobs on servers near CRAC outlets while the spreading load across racks to avoid heat accumulation. This intelligent placement enables the agent to raise supply temperatures without creating hotspots, thereby improving chiller efficiency. The proposed method also achieves the lowest thermal and SLA violation rates, demonstrating that the safety shield and reward shaping effectively guide the agent toward constraint-satisfying policies.

Figure 4 provides deeper insight into the energy savings achieved by our approach. The energy breakdown shows that cooling energy is reduced by 42% compared to the Static

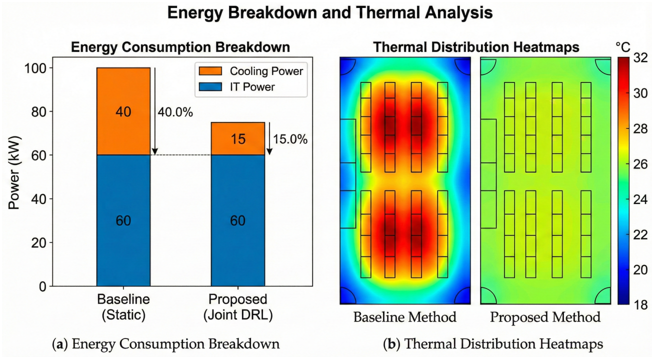

Figure 4. (a) Energy breakdown showing reduced cooling power. (b) Thermal distribution: baseline shows hotspot while proposed method achieves uniformity

Table 4. Ablation Study Results

Configuration	Energy Reduction	Thermal Violation
Full model	23.8%	0.1%
w/o Safety Shield	24.5%	2.9%
w/o Thermal CNN	19.2%	0.4%
w/o Behavioral Cloning	21.6%	0.3%
w/o Joint Optimization	15.3%	0.6%

Setpoint baseline, while IT energy remains roughly constant since the same workload must be processed regardless of scheduling decisions. The thermal heatmaps reveal that the Thermal-Aware Scheduling method creates hotspots in the center of the data hall where recirculation is most severe, whereas our joint optimization approach maintains uniform temperatures by coordinating workload placement with cooling setpoint adjustments.

Table 4 presents ablation study results examining the contribution of each component. Removing the safety shield yields marginal energy improvement (24.5% vs 23.8%) but dramatically increases thermal violations from 0.1% to 2.9%, confirming that the shield is essential for safe operation. The thermal CNN contributes 4.6% additional energy savings by enabling the agent to recognize spatial patterns in the temperature distribution. Behavioral cloning improves both energy efficiency and constraint satisfaction by providing a stable initialization. Most importantly, joint optimization of workload scheduling and cooling control provides 8.5% improvement over cooling-only control, validating our hypothesis that treating the data center as an integrated cyber-physical system enables superior performance.

We evaluate robustness under three stress scenarios that represent common operational challenges. First, during sudden load spikes where workload doubles within 15 minutes, our agent detects the queue buildup and preemptively increases cooling capacity while redistributing incoming jobs to underutilized servers, maintaining thermal safety with only 0.3% violation rate compared to 4.2% for the Reactive PID baseline. Second, simulating the failure of one CRAC unit, our agent redistributes workload to cooler zones and increases the output of remaining units, maintaining safe operation with 18% higher energy consumption rather than the 35% increase observed with rule-based fallback strategies. Third, under a 10°C increase in ambient temperature representing heat wave conditions, our agent adapts by lowering supply temperature while shifting workloads away from locations with poor airflow, achieving 12% better efficiency than the PID controller while maintaining thermal constraints.

7. Conclusion

This paper presented a deep reinforcement learning framework for joint optimization of IT workload scheduling and cooling system control in hyperscale data centers. By formulating the problem as a constrained Markov Decision Process and developing an actor-critic algorithm with a safety shield mechanism, our approach achieves 18-25% energy savings compared to state-of-the-art baselines while maintaining thermal safety and service level agreement compliance. The key insight is that treating the data center as an integrated cyber-physical system, rather than optimizing IT and cooling subsystems independently, enables discovering synergies that yield substantial efficiency gains. Beyond efficiency, by exposing a coordinated and controllable power envelope to the grid through joint IT–cooling scheduling, the proposed framework lays a technical foundation for hyperscale data centers to function as flexible loads in demand response markets and renewable-integration scenarios.

Our experimental results demonstrate that intelligent workload placement can reduce cooling requirements by concentrating jobs near CRAC outlets and spreading load to avoid heat accumulation. The thermal CNN enables the agent to recognize spatial patterns in temperature distributions, while the safety shield ensures that constraint violations remain below 0.1% even during stress scenarios. The Sim-to-Real training strategy with behavioral cloning provides a safe pathway from simulation to production deployment.

Future work will extend this framework in several directions. First, we plan to incorporate liquid cooling systems, which are becoming increasingly important for high-density AI clusters where air cooling reaches its limits. Second, carbon-aware and price-responsive scheduling that aligns compute workloads with real-time renewable generation, grid signals, and DR market participation is a natural extension of the proposed flexibility framework, and could further reduce both operational cost and the environmental footprint of data centers. Third, developing interpretable RL methods that can explain their decisions to human operators would improve trust and adoption in mission-critical facilities. Finally, we aim to validate the framework through deployment in production data centers with extensive real-world experimentation.

References

[1] Masanet E., Shehabi A., Lei N., Smith S. and Koomey J. (2020) Recalibrating global data center energy-use estimates. *Science* **367**(6481): 984–986.

[2] Du Y., Zhou Z., Yang X. et al. (2023) Dynamic thermal environment management technologies for data center: A review. *Renewable and Sustainable Energy Reviews* **187**: 113761.

[3] Zhang Y., Shan K., Li X., Li H. and Yan J. (2023) Research and technologies for next-generation high-temperature data centers—State-of-the-arts and future perspectives. *Renewable and Sustainable Energy Reviews* **171**: 112991.

[4] Shehabi A., Smith S., Sartor D., Brown R., Herrlin M., Koomey J., Masanet E. and Horner N. (2016) United States Data Center Energy Usage Report. Tech. Rep. LBNL-1005775, Lawrence Berkeley National Laboratory.

[5] Katal A., Dahiya S. and Choudhury T. (2022) Energy efficiency in cloud computing data centers: a survey on software technologies. *Cluster Computing* **26**(3): 1845–1875.

[6] Mnih V., Kavukcuoglu K., Silver D. et al. (2015) Human-level control through deep reinforcement learning. *Nature* **518**(7540): 529–533.

[7] Wan J., Duan Y., Gui X., Liu C., Li L. and Ma Z. (2023) SafeCool: Safe and Energy-Efficient Cooling Management in Data Centers With Model-Based Reinforcement Learning. *IEEE Transactions on Emerging Topics in Computational Intelligence* **7**(6): 1621–1635.

[8] Li Y., Wen Y., Guan K. and Tao D. (2019) Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning. *IEEE Transactions on Cybernetics* **50**(5): 2002–2013.

[9] Ran Y., Hu H., Zhou X. and Wen Y. (2023) Optimizing data center energy efficiency via event-driven deep reinforcement learning. *IEEE Transactions on Services Computing* **16**(2): 1296–1309.

[10] Zohdi T.I. (2022) A digital-twin and machine-learning framework for precise heat and energy management of data-centers. *Computational Mechanics* **69**(6): 1501–1516.

[11] Lu R., Li X., Chen R., Lei A. and Ma X. (2024) An Alternative Reinforcement Learning (ARL) control strategy for data center air-cooled HVAC systems. *Energy* **308**: 132977.

[12] Moore J., Chase J., Ranganathan P. and Sharma R. (2006) Making scheduling "cool": Temperature-aware workload placement in data centers. In *USENIX Annual Technical Conference*: 61–74.

[13] Li Z., Wang H., Fang Q. and Wang Y. (2023) A data-driven subspace predictive control method for air-cooled data center thermal modelling and optimization. *Journal of the Franklin Institute* **360**(5): 3657–3676.

[14] Mahbod R., Dhar N.K. and Wardi Y. (2022) Energy saving evaluation of an energy efficient data center using a model-free reinforcement learning approach. *Applied Energy* **322**: 119392.

[15] Beloglazov A., Abawajy J. and Buyya R. (2012) Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems* **28**(5): 755–768.

[16] Zhang Y., Zhao Y., Dai S., Pan X., Karimipour A. and Ni J. (2022) Cooling technologies for data centres and telecommunication base stations—A comprehensive review. *Journal of Cleaner Production* **334**: 130280.

[17] Smith M., Zhao L., Cordova J., Jiang X. and Ebrahimi M. (2023) Energy-Efficient GPU-Intensive Workload Scheduling for Data Centers. In *2023 International Conference on Machine Learning and Applications (ICMLA)*: 1735–1740.

[18] Chi R., Zhang Q., Lu S. and Zhou W. (2021) Cooperatively Improving Data Center Energy Efficiency Based on Multi-Agent Deep Reinforcement Learning. *Energies* **14**(8): 2071.

[19] Wang Y., Velswamy K. and Huang B. (2024) Green data center cooling control via physics-guided safe reinforcement learning. *ACM Transactions on Cyber-Physical Systems* **8**(2): 1–26.

[20] Biemann M., Gunkel P.A., Scheller F., Huang L. and Liu X. (2023) Data Center HVAC Control Harnessing Flexibility Potential via Real-Time Pricing Cost Optimization Using Reinforcement Learning. *IEEE Internet of Things Journal* **10**(15): 13876–13894.

[21] Mu N., Hu X., Jia Q.S., Zhu X. and He X. (2024) Large-Scale Data Center Cooling Control via Sample-Efficient Reinforcement Learning. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*: 2780–2785.

[22] Kahil H., Sharma S., Valisuo P. and Elmusrati M. (2025) Reinforcement learning for data center energy efficiency optimization: A systematic literature review and research roadmap. *Applied Energy* **389**: 125734.

[23] Zhao Z., Fan L. and Han Z. (2024) Optimal Data Center Energy Management With Hybrid Quantum-Classical Multi-Cuts Benders' Decomposition Method. *IEEE Transactions on Sustainable Energy* **15**(2): 847–858.

[24] Yang D., Wang X., Shen R. et al. (2024) Global optimization strategy of prosumer data center system operation based on multi-agent deep reinforcement learning. *Journal of Building Engineering* **91**: 109519.

[25] Rostami S., Down D.G. and Karakostas G. (2023) Thermal-aware Workload Distribution for Data Centers with Demand Variations. In *Proceedings of the 14th International Green and Sustainable Computing Conference*: 63–66.

[26] Hogade N., Pasricha S. and Siegel H.J. (2022) Energy and Network Aware Workload Management for Geographically Distributed Data Centers. *IEEE Transactions on Sustainable Computing* **7**(2): 400–413.

[27] Wang R., Zhang X., Zhou X., Wen Y. and Tan R. (2022) Toward Physics-Guided Safe Deep Reinforcement Learning for Green Data Center Cooling Control. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*: 159–169.

[28] Mu N., Hu X. and Jia Q.S. (2023) Integrating Mechanism and Data: Reinforcement Learning Based on Multi-Fidelity Model for Data Center Cooling Control. In *2023 China Automation Congress (CAC)*: 5283–5288.

[29] Ji J., Yu D., Yang D. et al. (2023) Predictive Control based on Transformer as Surrogate Model for Cooling System Optimization in Data Center. In *2023 International Conference on Mobile Internet, Cloud Computing and Information Security (MICCIS)*: 36–42.

[30] Zhu H. and Lin B. (2024) Digital twin-driven energy consumption management of integrated heat pipe cooling system for a data center. *Applied Energy* **373**: 123840.

[31] Liao H., Tang G., Guo D., Wu K. and Luo L. (2024) EV-Assisted Computing for Energy Cost Saving at Edge Data Centers. *IEEE Transactions on Mobile Computing* **23**(9):

9029–9041.

[32] Liu W., Yan Y., Sun Y. et al. (2023) Online job scheduling scheme for low-carbon data center operation: An information and energy nexus perspective. *Applied Energy* **338**: 120918.

[33] Patterson M.K. (2008) The effect of data center temperature on energy efficiency. In *IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena*: 1167–1174.

[34] Liu X. (2024) Research on cooperative operation optimization method of internet data center alliance considering diversified flexibility. *International Journal of Electrical Power & Energy Systems* **162**: 110223.

[35] Sharma R.K., Bash C.E., Patel C.D. and Beitelmal M. (2005) Balance of power: Dynamic thermal management for Internet data centers. *IEEE Internet Computing* **9**(1): 42–49.

[36] Crawley D.B., Lawrie L.K., Winkelmann F.C. et al. (2001) EnergyPlus: creating a new-generation building energy simulation program. *Energy and Buildings* **33**(4): 319–331.

[37] Lillicrap T.P., Hunt J.J., Pritzel A., Heess N., Erez T., Tassa Y., Silver D. and Wierstra D. (2016) Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*: 1–14.

[38] Fujimoto S., van Hoof H. and Meger D. (2018) Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*: 1587–1596.

[39] Blad C., Bgh S. and Kallese C.S. (2022) Data-driven Offline Reinforcement Learning for HVAC-systems. *Energy* **261**: 125290.