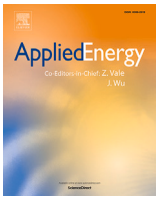
Applied Energy 389 (2025) 125734

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apen



Reinforcement learning for data center energy efficiency optimization: A systematic literature review and research roadmap

Hussain Kahil [a,*], Shiva Sharma [b], Petri Välisuo [a], Mohammed Elmusrati [a]

[a] School of Technology and Innovation, University of Vaasa, Wolffintie 32, Vaasa, 65200, Finland
[b] School of Technology, Vaasa University of Applied Sciences, Wolffintie 30, Vaasa, 65200, Finland

HIGHLIGHTS

- Discusses using Reinforcement Learning (RL) for data center cooling system.
- Discusses using RL for data center information and communication (ICT) system.
- Provides a deep critical analysis for the energy optimization results.
- Presents a comprehensive data extraction about the experimental setup and benchmarks.
- Explores future direction in RL for optimizing energy in data center environments.

ARTICLE INFO

Keywords:
Data center
Energy efficiency optimization
Cooling system
ICT system
Reinforcement learning (RL)
Deep reinforcement learning (DRL)

ABSTRACT

With today's challenges posed by climate change, global attention is increasingly focused on reducing energy consumption within sustainable communities. As significant energy consumers, data centers represent a crucial area for research in energy efficiency optimization. To address this issue, various algorithms have been employed to develop sophisticated solutions for data center systems. Recently, Reinforcement Learning (RL) and its advanced counterpart, Deep Reinforcement Learning (DRL), have demonstrated promising potential in improving data center energy efficiency. However, a comprehensive review of the deployment of these algorithms remains limited. In this systematic review, we explore the application of RL/DRL algorithms for optimizing data center energy efficiency, with a focus on optimizing the operation of cooling systems and Information and Communication Technology (ICT) processes, including task scheduling, resource allocation, virtual machine (VM) consolidation/placement, and network traffic control. Following the Preferred Reporting Items for Systematic review and Meta-Analysis (PRISMA) protocol, we provide a detailed overview of the methodologies and objectives of 65 identified studies, along with an in-depth analysis of their energy-related results. We also summarize key aspects of these studies, including benchmark comparisons, experimental setups, datasets, and implementation platforms. Additionally, we present a structured qualitative comparison of the Markov Decision Process (MDP) elements for joint optimization studies. Our findings highlight vital research gaps, including the lack of real-time validation for developed algorithms and the absence of multi-scale standardized metrics for reporting energy efficiency improvements. Furthermore, we propose joint optimization of multi-system objectives as a promising direction for future research.

* Corresponding author.
Email addresses: hussain.kahil@uwasa.fi (H. Kahil), shiva.sharma@vamk.fi (S. Sharma), petri.valisuo@uwasa.fi (P. Välisuo), mohammed.elmusrati@uwasa.fi (M. Elmusrati).

https://doi.org/10.1016/j.apenergy.2025.125734
Received 10 January 2025; Received in revised form 25 February 2025; Accepted 14 March 2025
Available online 25 March 2025

Nomenclature

A3C Asynchronous advantage actor-critic
AC Actor-critic
ACO Ant Colony Optimization
ACS Ant Colony System
ADVMC Adaptive DRL based VM Consolidation
AFED-EF Adaptive Four-threshold Energy-aware VM Deployment
ARLCA Advanced RL Consolidation Agent
ATES Aquifer Thermal Energy Storage
AVMC Autonomous VM Consolidation
AVT Active Ventilation Tile
BDQ Branching Dueling Q-Network
BF Best Fit
BFD Best Fit Decreasing
CARPO Correlation-AwaRe Power Optimization
CCO Cooling Control Optimization
CDRL Constrained DRL
CFD Computational Fluid Dynamics
CFWS Cost and carbon Footprint through Workload Shifting
CNN Convolutional Neural Network
CSLB Crow Search-based Load Balancing
CVP Chemical reaction optimization-VMP-Permutation
CW Chilled Water
D3QN Dueling Deep Q Network
DAG Directed Acyclic Graph
DBC Deadline and Budget Constrained
DCI Dynamic Control Interval
DCN Data Center Network
DDPG Deep Deterministic Policy Gradient
DL Deep Learning
DPPE Data Center Performance Per Energy
DPSO Discrete Particle Swarm Optimization
DQN Deep Q-Network
DRL Deep Reinforcement Learning
DSTS Dynamic Stochastic Task Scheduling
DTA DRL-based Task Migration
DTH-MF Dynamic Threshold Maximum Fit
DTM Dynamic Thermal Management
DUE De-underestimation Validation Mechanism
DX Direct Expansion
ECA Enclosed Cold Aisle
EDF Earliest Deadline First
EMVO Enhanced Multi-Verse Optimizer
EOM Energy Optimization Module
EQBFD Energy-efficient and QoS-aware BFD
ERE Energy Reuse Effectiveness
ERLFC Eco-friendly RL in Federated Cloud
ETAS Energy and Thermal-Aware Scheduling
ETF Earliest Time First
ETHC Elastic Task Handler over hybrid Cloud
EVCT Energy-efficient VM minimum Cut Theory
EVMM Energy-aware VM Migration
FCT Flow Completion Time
FERPTS Fast and Energy-aware Resource Provisioning and Task Scheduling
FF First Fit
FFD First Fit Decreasing
FFO FireFly Optimization
FIFO First-In-First-Out
GA Genetic Algorithm
GCD Google Cluster Dataset
GEC Green Energy Coefficient
GJO Golden Jackal Optimization
GMPR Greedy Minimizing Power consumption and Resource wastage
GRF Generalized Resource-Fair
GRR Generalized Round Robin
GRVMP Greedy Randomized VM Placement
HDDL Heterogeneous Distributed Deep Learning
HDRL Hierarchical DRL
HEFT Heterogeneous Earliest Time First
HGP Heteroscedastic Gaussian Processes
HM Host Machine
HVAC Heating, Ventilation, and Air Conditioning
ICA Imperialist Competitive Algorithm
ICO IT Control Optimization
ICT Information and communication Technology
IGGA Improved Grouping Genetic Algorithm
IQR Inter-Quartile Range
ITEE IT Equipment Energy
ITEU IT Equipment Utilization
JCO Joint IT and Cooling Control Optimization Algorithm
KMI-MRCU K-Means clustering algorithm-Midrange-Interquartile range
LECC Location, Energy, Carbon and Cost-aware vm placement
LR Logistic Regression
LRR Local regression robust
LSTM Long Short-Term Memory
MAD Median Absolute Deviation
MAGNETIC Multi-AGent machine learNing-based approach for Energy efficienT dynamIc Consolidation
MBAC Model-Based Actor-Critic
MBHC MBRL-based HVAC control
MBRL Model-Based RL
MCP Modified Critical Path
MCTS Monte Carlo Tree Search
MDP Markov Decision Process
MFFD Modified First Fit Decreasing
MGGA Multi-objective Genetic Algorithm
MILP Mixed Integer linear programming
MIMT Minimization of Migration based on Tesa
MLF Minimum Load First
MMT Minimum Migration Time
MOACO Multi-Objective Ant Colony Optimization
MOPSO Multi-Objective Particle Swarm Optimization
MPC Model Predictive Control
MSP Multi-Set Point
MVO Multi-Verse Optimizer
NFV Network Function Virtualization
NPA Non-Power-Aware
NSGA-II Non-dominated Sorting Genetic Algorithm II
OCA Open Cold Aisle
OEMACS Order Exchange and Migration Ant Colony System
PABFD Power-aware Best Fit Decreasing
PADQN PArametrized Deep Q-Network
PETS Probabilistic Ensembles with Trajectory Sampling
PID Proportional-Integral-Derivative
PM Physical Machine
PPO Proximal Policy Optimization
PRISMA Preferred Reporting Items for Systematic review and Meta-analysis
PSO Particle Swarm Optimization
PUE Power Usage Effectiveness
QEEC Q-learning Energy-Efficient Cloud computing
QL Q-learning
RAC Resource Allocation in container-based Clouds

RDHX Rear Door Heat Exchangers
RES Renewable Energy Systems
RH Relative Humidity
RLR Robust Logistic Regression
RP Residual Physics
RR Round Robin
RTP Real-Time Pricing
SAC Soft Actor Critic
SARSA State-Action-Reward-State-Action
SDAEM Stacked De-noising Auto-encoders with Multilayer Perception
SDN Software-Defined Networking
SFC Service Function Chaining
SLA Service Level Agreement
SO Snake optimizer
SSP Single-Set Point
TDBS Task Duplication-Based Scheduling
TPM Traffic Prediction Module
TRPO Trust Region Policy Optimization
UP Utilization Prediction-aware
UPS Uninterruptible Power Supply
VDN Value Decomposition Network
VDT-UMC VM-based Dynamic Threshold and Minimum Correlation of Host Utilization
VM Virtual Machine
VMC VM Consolidation
VMP VM placement
VMPMBBO Multi-objective Biogeography-Based Optimization
VMTA VM Traffic burst
VPBAR VM scheduling Based on Poisson Arrival Rate
VPME VM Placement with Maximizing Energy efficiency
WUE Water Usage Effectiveness

1. Introduction

The digitalization of society and the emergence of new AI technologies have increased the overall demand for computing power. This growth has made data centers a critical infrastructure that supports our modern digital ecosystems. The rise in the use of technologies such as the Internet of Things (IoT), cloud computing, big data, and artificial intelligence (AI) has increased the workload of data centers, which now require even more computing resources to meet demand. Data centers form the backbone of modern digital infrastructure, and their high energy consumption has substantial financial and environmental implications. According to International Energy Agency [1], an estimated 460 terawatt hours (TWh) of electricity, with projections indicating that this could exceed 1000 TWh by 2026. In the European Union (EU), data centers consumed approximately 45–65 TWh of electricity in 2022, representing 1.8 % to 2.6 % of the total electricity consumption of the EU for that year [2].

This substantial energy consumption contributes to increased operational costs and has significant environmental consequences, including large amounts of greenhouse gas emissions [3], and increased strain on power grids [4]. Therefore, improving energy efficiency in data centers has become a critical issue, requiring intelligent and automated solutions capable of dynamically adapting to real-time demands.

Among the many emerging technologies, Reinforcement Learning (RL) and its subset, Deep Reinforcement Learning (DRL), have gained attention as promising techniques for optimizing energy efficiency within complex environments like data centers. These algorithms enable systems to learn optimal policies by interacting with dynamic environments, making them suitable for resource allocation, task scheduling, and heating and cooling management. A study conducted by Jayanetti et al. [5] demonstrates the significant potential of RL/DRL for minimizing energy consumption and reducing operational costs.

The data center architecture comprises three main systems: information and communication technology (ICT), cooling, and power supply systems. Today's data centers are vast, complex, and highly sophisticated, powered by a diverse ecosystem of ICT devices. These range from high-performance servers equipped with heterogeneous computing processors, such as CPUs, GPUs, and specialized accelerators, to arrays of memory units and storage solutions. In addition to computational infrastructure, the cooling system is critical in sustaining data center functionality. Its complexity arises from integrating multiple subsystems designed to regulate thermal conditions and protect highly sensitive ICT equipment from overheating. Efficient cooling is a fundamental aspect of data center operations, directly impacting energy consumption, operational costs, and system reliability. Due to the high heat dissipation of modern ICT equipment, data center cooling systems are designed to maintain optimal temperatures, prevent hardware failures, and enhance overall performance.

A typical data center cooling system consists of multiple components, including chillers, pumps, fans, heat exchangers, and cooling towers, which work together to regulate temperature and ensure efficient heat dissipation. These systems can generally be classified into air-based and liquid-based cooling solutions. Air-based cooling relies on Computer Room Air Conditioning (CRAC) units [6] and Computer Room Air Handlers (CRAH) [7]. Liquid-based cooling, in contrast to air-based methods, incorporates technologies such as direct-to-chip cooling [8], spray/immersion cooling [9]. These approaches significantly enhance thermal management by efficiently dissipating heat and directly cooling critical components. Recently, localized heat exchanger solutions, such as in-row, rear-door cooling and in-rack, have gained popularity due to their efficiency in high-density environments. In-row cooling places cooling units between server racks, reducing airflow distance and improving cooling efficiency [10]. Rear Door Heat Exchangers (RDHX), on the other hand, attach cooling units directly to the back of racks, capturing and dissipating heat immediately as it exits the servers. These strategies enhance cooling performance while minimizing energy waste by targeting heat removal close to the source [11].

Free cooling is an energy-efficient heat rejection method that uses low ambient air or water temperature with a dry cooler or heat exchanger. Depending on the ambient media, the free cooling is also known as water-side or air-side economizer [12]. Heat pumps [13] and thermal energy storage [14] are increasingly being adopted to enhance energy efficiency and overall performance of heat reuse. Fig. 1 provides a schematic diagram of the data center cooling and heat rejection and reuse systems.

Solutions based on RL/DRL techniques enable adaptive, real-time decision making which has significant potential for enhancing energy efficiency through optimization in the complex data center environments. Despite all these promising developments, the adoption of RL/DRL for minimizing energy consumption in data centers faces various challenges, including the complexity of modeling data center environments, managing computational costs, and ensuring scalability [15]. To address these challenges, innovative and intelligent solutions are required that can adapt to complex and dynamic environments in real-time. Several previous reviews on the use of RL/DRL have been conducted for general applications [16] rather than analyzing a holistically integrated RL/DRL framework with a specific system, which this paper aims to examine. Additionally, few studies have provided systematic evaluations of RL/DRL across data center functions, leaving a gap in understanding

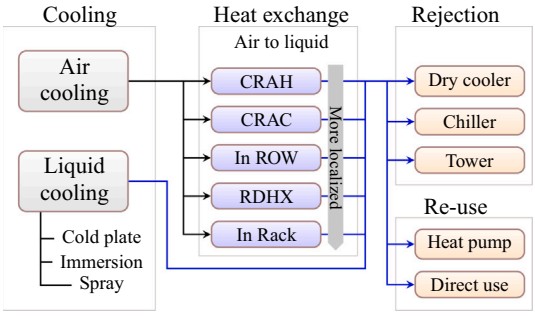


Fig. 1. Schematic diagram of data center cooling, heat rejection, and heat reuse system options. Black and blue arrows show heat flows in air, and liquid respectively. The grey arrow shows that the heat exchangers at the bottom of the middle box are localized closer to the heat source, whereas those at the top are far from it. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

these algorithms' capabilities in real-world environments. In this systematic literature review, we aim to investigate the recent advancements and applications of RL/DRL for enhancing energy efficiency in data centers by analyzing the literature using the PRISMA framework. The main objective of this work is to explore and evaluate the diverse potential applications of RL/DRL as a tool for optimizing energy efficiency in data centers while also synthesizing and consolidating existing research knowledge on their implementation in such facilities. Furthermore, this study aims to achieve the following specific objectives:

- Investigate and assess key applications of RL/DRL in data centers: This review aims to provide a comprehensive analysis of how RL/DRL algorithms have been applied to solve various energy efficiency challenges in data centers. To achieve this, we categorize RL/DRL applications by data center subsystems, giving readers insights into their roles and effectiveness.
- Evaluate and summarize each identified study in terms of algorithm type, the specific research problem addressed, primary objectives, and energy-efficiency outcomes, along with the benchmarks employed for performance evaluation, enabling a deeper understanding of the current state of research.
- Summarize details about the execution aspects of the identified studies: the implementation environment, dataset source, dataset type, and the platforms or frameworks utilized, offering insights into the practical considerations and resources required for implementing future studies.
- Utilize the identified joint optimization studies to present comprehensive guidelines for formulating the Markov Decision Process (MDP) elements, providing readers with a clear overview and foundational knowledge to construct such frameworks in future research.
- Identify technical and practical challenges in the current research direction. By investigating the essential issues related to RL/DRL usage in data centers, we aim to provide an in-depth view of the barriers that limit the broader use of these techniques in the data center industry.
- Highlight other objectives integrated with the energy efficiency problem in the identified studies, to address multi-objective optimization, thereby comprehensively ensuring sustainable and cost-effective operations in modern data centers.
- Explore research gaps, open issues, and future directions to propose a strategic roadmap for advancing the practical deployment of RL/DRL techniques in optimizing data center energy efficiency.

Through the above-mentioned objectives, this review aims to contribute a structured synthesis of RL/DRL applications for data center energy efficiency, identify persistent challenges, and chart a course for future research to address existing limitations and enhance the practical utility of RL/DRL techniques in data centers.

The remainder of this paper is organized as follows. Section 2 compares previous related reviews and this study. Section 3 provides a comprehensive background on RL/DRL algorithms. Section 4 outlines the research methodology. Section 5 explores the relevant literature in detail. Section 6 offers an overview of additional objectives combined with energy efficiency. Section 7 discusses the identified research gaps, open challenges, and suggests future directions. Finally, Section 8 concludes this review.

2. Related reviews

Several existing reviews focus on the energy efficiency of the data center cooling system as a key objective. Chang et al. [17] explore the cooling system optimization strategies in data centers by utilizing bibliometric methods. Their review examines the utilization of RL as a cooling control strategy for energy efficiency applications. Additionally, Shaqour et al. [18] investigate the literature on using DRL algorithms for HVAC energy management in data centers, which are considered as a subgroup of smart buildings.

In contrast, other reviews target the energy efficiency of ICT systems in data centers. Gari et al. [19] evaluate the effectiveness of RL algorithms for data center scaling and scheduling purposes in the literature, while partially addressing energy consumption as an optimization objective. Magotra et al. [20] provide a comprehensive overview of using VM consolidation to enhance the data center energy efficiency. This review surveys the research problem based on architecture and VM consolidation steps. Zhou et al. [21] present DRL-based approaches for resource scheduling in the cloud, highlighting their advantages, challenges, and future directions. Recently, Hou et al. [22] provided a specialized review on leveraging DRL algorithms for energy-efficient task scheduling in cloud computing. This study conducts an in-depth investigation of the Markov Decision Process (MDP) model components. Singh et al. [23] summarize previous empirical studies on multiple objectives in ICT systems, such as task scheduling and VM consolidation, to enhance energy efficiency while maintaining system performance.

Furthermore, other reviews combine cooling and ICT systems as the core topic of their review. Lin et al. [24] explore previous efforts to achieve green-aware data centers from five different perspectives: workload management, virtual resource management, energy management, thermal management, and waste heat recovery. Long et al. [25] outline performance evaluation metrics for data center energy efficiency through ICT systems and infrastructure, including cooling and power supply systems. Conversely, Zhang et al. [26] address the joint optimization of cooling and ICT systems to achieve effective data center management under a set of evaluation metrics, including thermal conditions, energy consumption, and response delay.

Although these reviews address energy efficiency objectives in data centers based on RL/DRL algorithms from different perspectives, there remains a gap in the existing literature due to the absence of a systematic overview of RL/DRL applications for improving the energy efficiency of data center systems. Additionally, there appears to be a lack of research addressing joint optimization using RL/DRL for energy efficiency objectives. Moreover, previous reviews do not sufficiently discuss experimental setups, including data sources and types used, and the implementation platforms. Our research introduces a systematic literature review that examines the use of RL/DRL for energy efficiency objectives across the main data center systems: cooling and ICT systems. We aim to explore recent advancements in this field to gain deeper insights, identify research gaps, and suggest future directions. Table 1 summarizes and compares related reviews and our work, emphasizing how our study differs from previous research.

Table 1
Related reviews on DC energy efficiency, and comparison with our review.

Reference	General focus			System specific			Review outcomes			
	Data center	Energy efficiency	RL/DRL approaches	Cooling system	ICT system	Joint optimization	Energy reporting	Algorithm comparisons	Benchmark comparisons	Experimental setup
[17]	●	●	●	●	✖	✖	●	◗	✖	✖
[18]	●	●	●	●	✖	✖	●	◗	✖	✖
[19]	●	◗	●	✖	◗	✖	◗	●	✖	✖
[20]	●	●	◗	✖	◗	✖	◗	◗	◗	●
[21]	●	◗	●	✖	◗	✖	◗	●	●	●
[22]	●	●	●	✖	◗	✖	●	●	●	●
[23]	●	◗	✖	✖	◗	✖	◗	✖	✖	✖
[24]	●	◗	◗	◗	◗	✖	◗	●	✖	✖
[25]	●	●	✖	◗	◗	✖	◗	●	✖	✖
[26]	●	◗	◗	✖	✖	●	●	●	✖	◗
Current review	●	●	●	●	●	●	●	●	●	●

● – Topic addressed in detail/self-contained, ◗ – Topic partially addressed (i.e., not self contained, requires additional readings for deep understanding), ✖ – Topic not addressed.

3. Overview of RL/DRL algorithms

Reinforcement learning (RL) stands out as a machine learning technique developed by the computational intelligence community. It is inspired by natural learning mechanisms, in which organisms adjust their future behavior based on feedback from interactions with the environment. Fundamentally, RL is a closed-loop approach aimed at maximizing the cumulative reward, allowing the decision-maker or agent to learn and adapt over time. However, the actions taken by the learning agent influence its future inputs. The RL algorithm establishes an interactive relationship with the dynamic environment, allowing the agent to perform actions, observe the states of the environment, and receive feedback in the form of rewards and punishments. In most practical cases, the agent's actions may not only influence the immediate reward but also shape the ultimate reward. In this closed-loop learning approach, the absence of explicit instructions for taking actions and the uncertainty of future consequences are the key features of RL. These characteristics position RL algorithms as an integration of adaptive and optimal control techniques [27,28]. Fig. 2 illustrates the general framework of RL algorithms.

Let us consider a typical reinforcement learning scenario within a fully observable, stationary, stochastic environment, where the agent interacts with the environment by fully and accurately observing the current state. At each discrete time step, the agent selects an action based only on the current state to maximize the cumulative reward over time. The representation of this scenario is given by:

- **States (S):** The set of all possible states of the environment that the agent can observe.

$$S = \{s_1, s_2, \dots, s_n\} \tag{1}$$

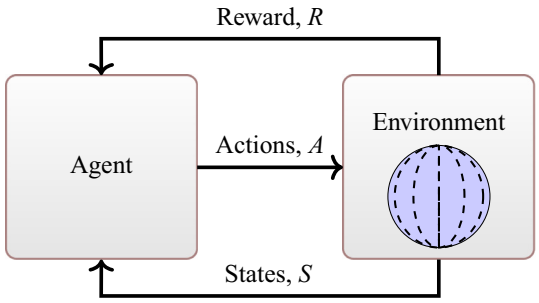


Fig. 2. RL framework.

- **Actions (A):** The set of all available actions that the agent can take in a given state.

$$A = \{a_1, a_2, \dots, a_n\} \tag{2}$$

- **Transition probabilities (P):** The probability of moving to a future state s' given the current s state and action a, which may differ over time due to dynamic changes.

$$P_t(s' \mid s, a) = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) \tag{3}$$

- **Reward function (R):** The immediate reward that the agent receives when taking action a in the state s at time t, which may differ over time due to dynamic changes.

$$R_t(s, a) = \mathbb{E}(\text{reward} \mid S_t = s, A_t = a) \tag{4}$$

- **Policy function (π):** This function determines the agent's future behavior by defining the probability of taking action a in the state s at time t, which may differ over time due to dynamic changes.

$$\pi_t(s, a) = \mathbb{P}(\text{reward} \mid S_t = s, A_t = a) \tag{5}$$

- **Discount factor (γ):** It determines the weight of future rewards compared to immediate rewards.

$$0 \le \gamma \le 1 \tag{6}$$

where the value of the discount factor is close to 0, it makes the RL agent focus on immediate reward, while a value close to 1 makes the RL agent focus on the future reward.

- **Objective (cumulative reward):** This is the ultimate goal of the RL agent to identify the trajectories that can maximize the expected discounted reward:

$$G_t = R_{t+1} + \gamma\ R_{t+2} + \gamma^2\ R_{t+3} + \cdots = \sum_{k=0}^{n} \gamma^k\ R_{t+k+1} \tag{7}$$

The tuple $\{S, A, P, R, \gamma\}$ formulates the Markov decision processes (MDP) representation for the proposed stationary stochastic environment. In the MDP framework, at each time step t, the agent interacts with the environment by observing the current state $s_t \in S$, choosing the action $a_t \in A$ according to the policy function $\pi_t(s_t, a_t)$, while estimating the probability of transitioning to a specific next state or taking a specific action using the transition probability model $P_t(s' \mid s, a)$. After taking the action, the agent obtains a reward $r_t \in R$ and transitions to the next state. The aim of reinforcement learning is to design the agent's

learning process to find the optimal policy that maximizes the expected cumulative reward over time G_t , considering the environment dynamics defined by the MDP [29–31].

However, the aforementioned process is not trivial. This challenge can be addressed recursively by introducing the state value function (V-function):

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_{s_t,a_t\sim\tau}\left[\sum_{k=0}^{n}\gamma^k R_{t+k+1}\right] \\ &= \sum_{(s_t,a_t,\dots)\sim\tau}\pi(a_t|s_t)P(s_{t+1}|s_t,a_t)\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \\ &= \mathbb{E}_\pi\left[R_{t+1}+\gamma V(S_{t+1})\mid S_t=s\right] \end{aligned} \tag{8}$$

where τ: $(s_0, a_0, s_1, a_1, \dots, a_{t-1}, s_t)$ represents the interaction trajectory of the RL agent.

Similarly, the expected return of taking a specific action a in a given state s while following the policy π can be given by the state-action value function (Q-function):

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1}+\gamma Q(S_{t+1},A_{t+1})\mid S_t=s, A_t=a\right] \tag{9}$$

Eqs. (8) and (9) are referred to as the Bellman equations [32], which are considered the fundamental formulas for tackling the decision-making process of an RL agent. The optimal V-function and Q-function are indicated by the maximum value across all states: $V^* = \max_{v_\pi(s)}$, or in all state-actions: $Q^*(s,a) = \max Q(s,a)$. In all MDP cases, at least one optimal policy always exists, and the value functions $V(s)$ and $Q(s,a)$ of all optimal policies are the same. As a result, optimizing the Q-function yields the optimal policy of the MDP:

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a\in A} Q^*(s,a) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

To obtain a solution to the MDP problem using RL techniques, two main categories of methods are used. Model-free RL algorithms allow an agent to learn a policy purely from interactions with the environment, without explicitly constructing a model of the environment's dynamics. The other category is called model-based RL algorithms and leverages a model of the environment, which can be given or learned. This model typically includes the transition probability function (3) and the reward function (4), allowing the agent to plan actions before execution [33]. Value-based algorithms are among the most popular model-free RL methods, where the agent estimates state-action values and represents them as a table (referred to as a Q-table or policy table), to optimize its decision-making. The most well-known value-based algorithms used for smaller MDP problems are tabular methods: Q-learning, in which the agent updates the table based on the maximum possible future reward (off-policy learning), making it more exploratory [34], and state-action-reward-state-action (SARSA) [35], where the agent updates the Q-table according to the actual action taken (on-policy learning), leading to more conservative behavior.

On the other hand, model-based RL leverages a model of the environment to update the Q-table of state-action pairs. This approach can be classified into two main categories based on how the environment model is acquired. In the first category, the agent learns the model through its interactions with the environment, as in the Dynamic Q-learning (Dyna-Q) algorithm [36]. In the second category, the model is provided to the agent, as seen in Monte Carlo Tree Search (MCTS) [37]. However, RL algorithms face scalability limitations when applied in large-scale learning environments. They often struggle with an extensive state space and continuous action space, leading to inefficiencies in the exploration–exploitation trade-off, slow convergence, and difficulties in learning optimal policies.

To address the limitations of traditional Reinforcement Learning (RL) methods, the computational intelligence community has developed Deep Reinforcement Learning (DRL), which integrates advancements in deep neural networks. In DRL algorithms, deep learning techniques are employed to construct at least one of the following agent components: value functions (8), (9), policy function (5), transition model(3), and the reward function (4). Such representations are essential when the RL agent interacts with environments characterized by a high-dimensional state space and a continuous action space. DRL is a powerful tool for achieving an end-to-end goal-directed learning process [38,39]. Figs. 3 and 4 present a comprehensive classification of the most popular RL/DRL algorithms based on their respective model types.

Another crucial aspect of RL/DRL algorithms is the type of policy used during the training process. The focus here is to determine whether

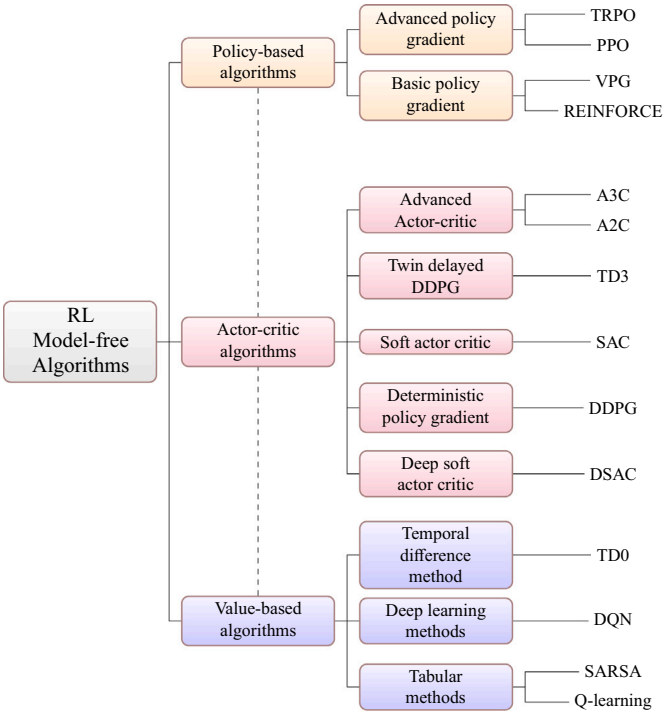


Fig. 3. RL/DRL model-free algorithms.

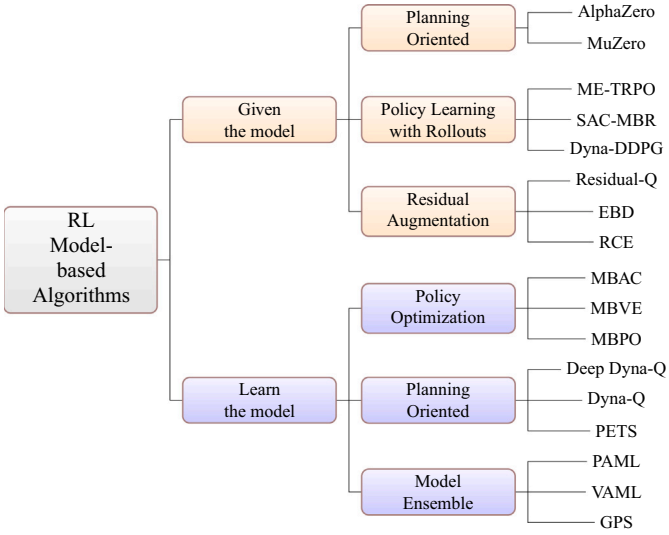


Fig. 4. RL/DRL model-based algorithms.

the behavior policy – defined as the policy interacting with the environment to collect training data – and the target policy – which represents the final policy that the agent is aiming to learn – are identical. On-policy methods utilize the collected data directly for the next round of policy optimization, meaning that the behavior and target policies are the same. However, in off-policy methods, the generated training data is stored in a buffer during the interaction with the environment. Then, during training, this stored data – which may be gathered from previous policies – is used for the target policy. In this case, the behavior policy is not the same as the target policy. The advantages of on-policy methods include greater stability and faster convergence, balanced exploration–exploitation rates, and ease of implementation, while off-policy methods offer better performance in complex environments and greater adaptability to changing policies.

Finally, RL/DRL are used to solve a wide range of optimization problems, from playing simple computer games to controlling highly complex large-scale configurations such as transportation networks and energy systems [40–42] . Both RL and DRL offer the advantage of boasting real-time adaptability and dynamic responsiveness compared to traditional control methods. However, without prior knowledge about the studied environment, they may encounter slow convergence and failures during the initial phases of operation [43,44].

4. Materials and methods

The methodology of this review was structured following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to ensure transparency, rigor, and reproducibility [45].

4.1. Research questions

The main aim of this review is to synthesize recent advancements in RL/DRL techniques for improving energy efficiency in data centers. To provide a comprehensive understanding of this topic, this study focuses on answering the following research questions based on the identified papers.

- **RQ1:** What data center subsystems (e.g., cooling, ICT equipment, power supply) are targeted by the RL/DRL algorithms?
- **RQ2:** Which RL/DRL algorithms are utilized for energy optimization in data centers?
- **RQ3:** What experimental setups and dataset sources (e.g., real-world deployments or simulations) are commonly used?
- **RQ4:** What specific research problems are addressed using RL/DRL algorithms?
- **RQ5:** What are the primary objectives addressed in the identified studies?
- **RQ6:** What benchmarks are used to evaluate the achieved results in terms of energy efficiency?
- **RQ7:** What tools, frameworks, or platforms are employed to implement RL/DRL algorithms in this context?
- **RQ8:** What metrics are used to measure and report the effectiveness of RL/DRL algorithms in improving energy efficiency?

4.2. Search strategy

4.2.1. Literature resources

To ensure that all recent and relevant studies are covered in the literature, the search was carried out in five major and well-established academic databases, known for their extensive repositories of peer-reviewed studies in computer science, engineering, and energy systems. Given that the scope of this review is relatively new, the covered time frame is limited to publications from 2019 to August 2024. To maintain high quality and credibility, only peer-reviewed journal articles from the databases mentioned below were selected.

- IEEE Xplore
- Scopus
- ScienceDirect
- Web of Science
- ACM Digital Library

4.2.2. Search terms (key words)

To ensure the high quality of this study, search queries were systematically designed using Boolean operators and keywords relevant to RL/DRL and energy efficiency in data centers. A representative search string was: (“data center” OR “data centers”) AND (“energy-aware” OR “energy utilization” OR “energy saving” OR “energy efficiency”) AND (“reinforcement learning” OR “RL”). Fig. 5 shows the search strategy used in this study.

4.3. Search process and selection criteria

To ensure the relevance and quality of the included studies, the PRISMA framework guided the article identification process, which involved four distinct stages:

1. **Identification**: Studies were retrieved using search queries across the selected databases.
2. **Screening**: The titles and abstracts were screened to eliminate irrelevant studies and duplicates.
3. **Eligibility**: Full-text articles were reviewed against the inclusion and exclusion criteria.
4. **Inclusion**: The final set of studies that met all quality assessment criteria was selected for detailed analysis.

A PRISMA flow diagram (Fig. 6) illustrates the selection process, documenting the number of studies identified, screened, excluded, and included.

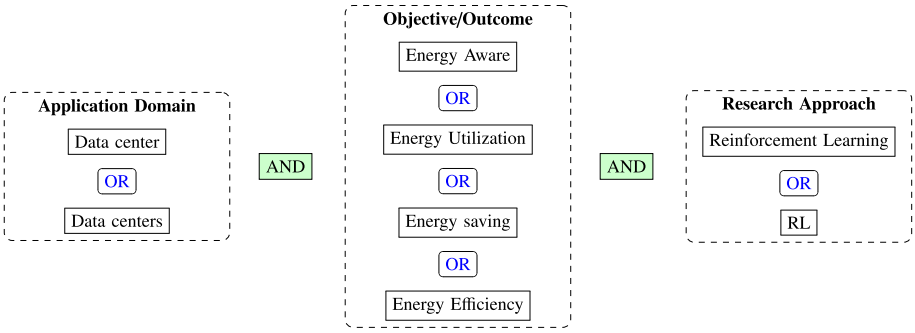


Fig. 5. Search strategy to get relevant papers.

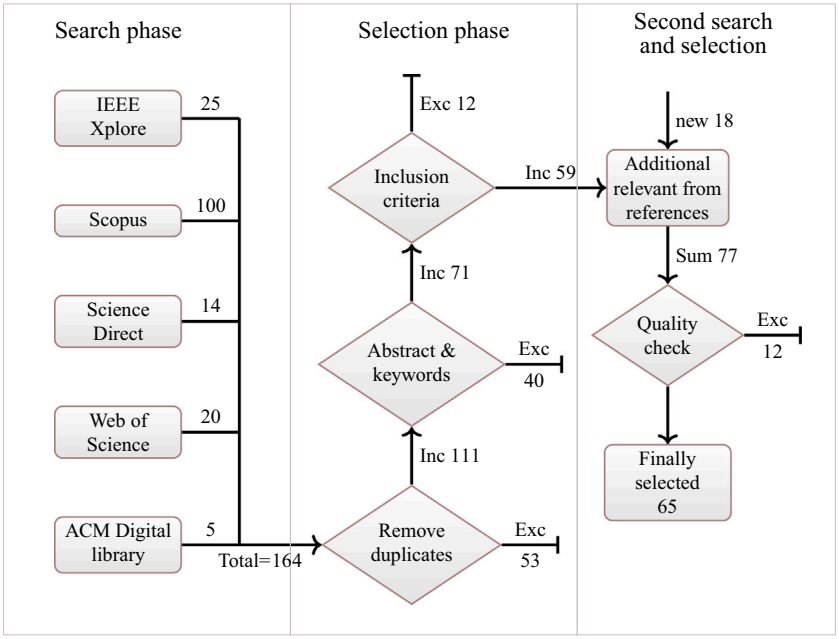


Fig. 6. Systematic literature review process stages: Removals of duplicates, removal based on abstract and keywords, removal based on inclusion and exclusion criterias, adding new articles which were found from the references, and finally removing those which did not match the quality criteria.

4.3.1. Inclusion and exclusion criteria

- **Inclusion criteria**: To ensure the inclusion of high-quality and relevant studies, the following criteria were applied:
  - Only peer-reviewed journal articles published between 2019 and August 2024.
  - Studies explicitly applying RL/DRL algorithms for energy efficiency in data center environment.
  - Studies presenting measurable outcomes, such as increased energy savings or improved Power Usage Effectiveness (PUE).
  - Studies focusing on specific or joint subsystems (e.g., cooling systems, ICT equipment, and/or power supply).
  - Only the most recent version of a study was included when duplicate publications were identified.

- **Exclusion criteria**: To facilitate the filtering of irrelevant studies, the following criteria were used:
  - Non-peer-reviewed studies, including conference papers, review articles, and opinion pieces.
  - Studies not addressing RL/DRL-based methods for energy optimization in data center environment.
  - Studies lacking empirical evidence or quantitative metrics.
  - Studies without full-text availability, making it impossible to assess the study's relevance and quality.
  - Studies focused on very small-scale experimental setups, as they lack applicability to real-world data center environments.

4.3.2. Quality assessment criteria and rating system

To ensure the final selection of identified articles are robust and reliable, a rigorous and systematic quality assessment process was implemented, based on the clearly defined criteria listed below:

- Clear and comprehensive documentation of the RL/DRL methods utilized, ensuring transparency in their implementation.
- Explicit definition and justification of the targeted subsystem's relevance within the study.
- Logical coherence in identifying the research problem and aligning it with the stated objectives.
- Methodological rigor in the design of experimental setups, including appropriate baseline comparisons and validation techniques.
- Implementation of well-defined metrics to assess energy efficiency, such as increased energy savings or improvements in Power Usage Effectiveness (PUE).
- Thorough comparative analysis of RL/DRL techniques against alternative benchmark methods to highlight their effectiveness and advantages.

Only studies that achieved a perfect score of 6 out of 6 on these criteria were included in the final synthesis.

4.4. Data extraction and synthesis

A comprehensive data extraction and synthesis template was completed for each identified study to ensure that all selected studies addressed the review's research questions. The extracted data were organized into a synthesis card and stored in an Excel file for further use throughout the systematic review stages. Table 2 summarizes the data extraction and synthesis card used to gather the necessary information from the identified studies.

To present the findings of this review, visual representations, such as pie charts, bar charts, and Venn diagrams, were created. Additionally, tables were utilized to systematically summarize and provide a detailed analysis of each identified study. This systematic approach provides a clear and structured framework for synthesizing and interpreting the collected data, while also highlighting research gaps, addressing challenges, and identifying future directions [46].

4.5. Threats to validity

The following threats to validity were acknowledged:

1. **Publication bias**: The focus on peer-reviewed journals may exclude innovative but unpublished studies.
2. **Database coverage**: Relevant articles from less-accessible databases or gray literature might have been missed.
3. **Variability in reporting**: Differences in methodologies and reporting standards across studies could limit comparability.

Table 2
Data extraction template.

Category
Unique Identifier (ID)
Study Title
Authors Names
Publication Venue
Publication Year
DC Subsystem Applications (RQ1)
RL/DRL Algorithm Type (RQ2)
Experimental Setup (RQ3)
Research Problems (RQ4)
Main Objectives (RQ5)
Benchmark Algorithms (RQ6)
Platforms and Frameworks (RQ7)
Energy Efficiency Outcomes (RQ8)
MDP Elements in Joint Optimization Studies
Abstract
Keywords
Other Performance Metrics

To mitigate these threats, standardized inclusion criteria were applied, and article selection and data extraction were independently verified by multiple reviewers.

5. Results and discussions

In this section, we discuss and present the findings of this review. First, we summarize the fundamental details of each identified study, including the study title, authors names, publication venue, and publication year. These details facilitated the systematic organization of this review, with each study assigned a unique identifier (ID) for easy reference during the data analysis and extraction process. Next, we provide a comprehensive analysis, highlighting key perspectives such as the studied subsystems, the RL/DRL algorithms applied, and the types of models utilized, offering valuable insights into the state-of-the-art. Then, we conduct a deeper synthesis, classifying the studies based on the subsystems they targeted. This categorization helped obtain quantitative and qualitative data to address the research questions for each subsystem. We focus our discussion on more detailed and specific information regarding the research problems, study objectives, and experimental setup, benchmark comparisons, the platforms used, and energy-related outcomes. Finally, we summarize the construction of Markov Decision Process (MDP) elements in joint optimization studies. Additionally, we reference related works to further support and contextualize the purpose and findings of this review.

5.1. Overview of the final identified studies

In this review, we identify 65 journal articles that apply RL/DRL algorithms to improve the energy efficiency of at least one major data center system. The publication venues and years of these articles are summarized in Table 3. Given that the research topic of this review is relatively new, all selected studies were published between 2020 and 2024, as shown in Fig. 7.

Taking a broader look at the selected studies reveals that over 60 % focus entirely on the ICT system, exploring opportunities to enhance energy efficiency by leveraging RL/DRL algorithms from various perspectives. In contrast, approximately 21 % of the papers focus exclusively on the data center cooling system. Furthermore, the remaining studies examine combinations of multiple data center systems. Fig. 8 provides a detailed overview of the specific systems addressed in each selected paper.

In the following paragraphs, we will explore the RL/DRL algorithms used in the selected studies of this review.

For the cooling system: Since the cooling system of data centers is characterized by a high-dimensional state space and a continuous action space MDP, all selected studies employed DRL methods, primarily focusing on model-free algorithms, including:

- Soft Actor-Critic Algorithm (SAC) [66,87]
- Deep Deterministic Policy Gradient (DDPG) [58,70]
- Twin Delayed DDPG (TD3) [93]
- Proximal Policy Optimization (PPO) [60]
- Trust Region Policy Optimization (TRPO) [93]
- Deep Q-Network (DQN) [69,75,101,104]

However, two studies used model-based algorithms: Model-Based Actor-Critic (MBAC) [49] to propose a safe cooling mode adhering to strict thermal constraints, and Probabilistic Ensembles with Trajectory Sampling (PETS) [102], in which the study makes a comparison between four different algorithms: two model-free off-policy algorithms: A DQN variant called Branching Dueling Q-Network (BDQ) and SAC, one model-free on-policy algorithm (PPO), and one model-based algorithm (PETS).

For ICT system: Due to the discrete nature of certain ICT processes, such as task scheduling and resource allocation, the Q-learning algorithm has been employed in multiple studies to handle the ICT MDP environment [62,79,105]. This approach allows Q-values to be updated independently from the action selection and execution, enabling the algorithm to capture delayed feedback more accurately. As a result, this method enhances the learning rate and accelerates the convergence process. Alternatively, DQN is commonly proposed for handling more complex ICT systems, as reported in [50,54,81]. However, other DRL algorithms are also used, such as:

- Actor-Critic (AC) [72,107]
- Soft Actor-Critic Algorithm (SAC) [55,65]
- Proximal Policy Optimization (PPO) [67,103]
- Asynchronous Actor-Critic Agents (A3C) [76]
- Deep Deterministic Policy Gradient (DDPG) [86]

For combining systems studies: As the complexity of the MDP problem increases when multiple systems are present, with a combination of discrete and continuous state spaces, along with high-dimensional action spaces, traditional RL approaches become less effective. In response to these challenges, all selected studies addressing the integration of multiple data center systems employed DRL algorithms. Notable DRL algorithms used in these studies include:

- Actor-Critic Algorithm (AC) [47]
- Soft Actor-Critic Algorithm (SAC) [48]
- Deep Q-Network (DQN) and its extensions [52,61,80,91,98]
- Deep Deterministic Policy Gradient (DDPG) [91,92]

Fig. 9 illustrates the distribution of various RL/DRL algorithms in the selected studies. Q-learning and DQN were the most frequently cited algorithms, appearing in 60 % of studies, followed by SAC (eight studies), PPO (four studies), DDPG (four studies), and AC/A3C (four studies). About 9 % of studies employed other algorithms.

Table 4 categorizes the algorithms implemented in the selected studies based on the utilized model type. According to Figs. 3 and 4, nearly 98 % of the algorithms employed are model-free, divided into three main groups: value-based algorithms, policy-based algorithms, and actor-critic algorithms. Only two studies utilized model-based algorithms, likely due to the complexity involved in accurately modeling a data center system. Some studies used more than one RL/DRL method, causing them to appear in multiple categories in the table.

The following sections will provide a detailed analysis of these algorithms and their applications.

Table 3
The selected studies.

ID	Authors	Publication venue	DC application (RQ1)	Year	Ref
S1	Jayanetti et al.	IEEE Transactions on Parallel and Distributed Systems	Integrating power supply and ICT systems	2024	[47]
S2	Biemann et al.	IEEE Internet of Things Journal	Integrating cooling and power supply systems	2023	[48]
S3	Wan et al.	IEEE Transactions on Emerging Topics in Computational Intelligence	Cooling system	2023	[49]
S4	Lou et al.	IEEE Transactions on Network and Service Management	ICT system	2023	[50]
S5	Ran et al.	IEEE Transactions on Services Computing	Integrating cooling and ICT systems	2023	[51]
S6	Ran et al.	IEEE Transactions on Services Computing	Integrating cooling and ICT systems	2023	[52]
S7	Zeng et al.	IEEE Transactions on Parallel and Distributed Systems	ICT system	2022	[53]
S8	Kang et al.	IEEE Transactions on Network and Service Management	ICT system	2022	[54]
S9	Pham et al.	IEEE Access	ICT system	2021	[55]
S10	Yi et al.	IEEE Transactions on Parallel and Distributed Systems	ICT system	2020	[56]
S11	Ding et al.	IEEE Access	ICT system	2020	[57]
S12	Li et al.	IEEE Transactions on Cybernetics	Cooling system	2020	[58]
S13	Cheng et al.	IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems	ICT system	2020	[59]
S14	Leindals et al.	Energy and AI	Cooling system	2024	[60]
S15	Zhao et al.	IEEE Transactions on Sustainable Computing	Integrating power supply and ICT systems	2024	[61]
S16	Ghasemi et al.	Cluster Computing	ICT system	2024	[62]
S17	Ghasemi et al.	Computing	ICT system	2024	[63]
S18	Bhatt et al.	International Journal of Advanced Computer Science and Applications	ICT system	2024	[64]
S19	Zhang et al.	IEEE Transactions on Network and Service Management	ICT system	2024	[65]
S20	Guo et al.	Applied Energy	Cooling system	2024	[66]
S21	Yang et al.	Journal of Supercomputing	ICT system	2024	[67]
S22	Bouaouda et al.	Sustainability	ICT system	2024	[68]
S23	Chen et al.	Measurement and Control	Cooling system	2024	[69]
S24	Wang et al.	ACM Transactions on Cyber-Physical Systems	Cooling system	2024	[70]
S25	Aghasi et al.	Computer Networks	Integrating cooling and ICT systems	2023	[71]
S26	Wang et al.	Journal of Cloud Computing	ICT system	2023	[72]
S27	Wang et al.	Computer Networks	ICT system	2023	[73]
S28	Ghasemi et al.	Cluster Computing	ICT system	2023	[74]
S29	Huang et al.	Energies	Cooling system	2023	[75]
S30	Wei et al.	Journal of King Saud University – Computer and Information Sciences	ICT system	2023	[76]
S31	Liu et al.	Applied Energy	ICT system	2023	[77]
S32	Ahamed et al.	Sensors	ICT system	2023	[78]
S33	Ma et al.	IEEE Transactions on Industrial Informatics	ICT system	2023	[79]
S34	Simin et al.	Journal of Intelligent and Fuzzy Systems	Integrating cooling and ICT systems	2023	[80]
S35	Nagarajan et al.	Expert Systems	ICT system	2023	[81]
S36	Yang et al.	KSII Transactions on Internet and Information Systems	ICT system	2022	[82]
S37	Pandey et al.	Mobile Information Systems	ICT system	2022	[83]
S38	Shaw et al.	Information Systems	ICT system	2022	[84]
S39	Yan et al.	Computers and Electrical Engineering	ICT system	2022	[85]
S40	Wang et al.	Computer Networks	ICT system	2022	[86]
S41	Mahbod et al.	Applied Energy	Cooling system	2022	[87]
S42	Abbas et al.	Physical Communication	ICT system	2022	[88]
S43	Uma et al.	Transactions on Emerging Telecommunications Technologies	ICT system	2022	[89]
S44	Wang et al.	Future Generation Computer Systems	ICT system	2021	[90]
S45	Zhou et al.	IEEE Network	Integrating cooling and ICT systems	2021	[91]
S46	Chi et al.	Energies	Integrating cooling and ICT systems	2021	[92]
S47	Biemann et al.	Applied Energy	Cooling system	2021	[93]
S48	Ding et al.	Future Generation Computer Systems	ICT system	2020	[94]
S49	Peng et al.	Cluster Computing	ICT system	2020	[95]
S50	Hu et al.	Electronics	Integrating power supply and ICT systems	2020	[96]
S51	Qin et al.	Applied Intelligence	ICT system	2020	[97]
S52	Yang et al.	Journal of Building Engineering	Integrating cooling, ICT, and power supply systems	2024	[98]
S53	Lin et al.	IEEE Access	ICT system	2020	[99]
S54	Caviglione et al.	Soft Computing	ICT system	2021	[100]
S55	Le et al.	ACM Transactions on Sensor Networks	Cooling system	2021	[101]
S56	Zhang et al.	Applied Energy	Cooling system	2023	[102]
S57	Li et al.	CCF Transactions on High Performance Computing	ICT system	2021	[103]
S58	Wan et al.	IEEE Intelligent Systems	Cooling system	2021	[104]
S59	Haghshenas et al.	IEEE Transactions on Services Computing	ICT system	2022	[105]
S60	Zhang et al.	IEEE Transactions on Cybernetics	Cooling system	2024	[106]
S61	Sun et al.	Computer Networks	ICT system	2020	[107]
S62	Asghari et al.	Computer Networks	ICT system	2020	[108]
S63	Siddesha et al.	Cluster Computing	ICT system	2022	[109]
S64	Asghari et al.	Soft Computing	ICT system	2020	[110]
S65	Zhang et al.	Expert Systems with Applications	Cooling system	2023	[111]

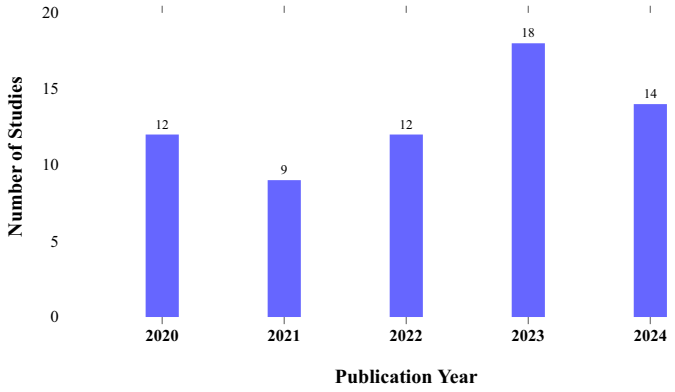


Fig. 7. Publication year distribution of selected studies.

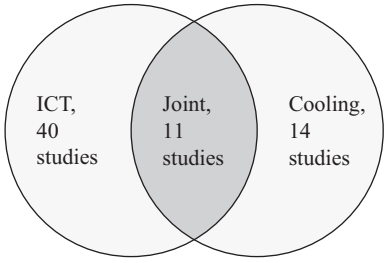


Fig. 8. The sub-systems focused in selected studies: 40 studies are focused on ICT optimization, 14 on cooling optimization and 11 are joint studies integrating multiple systems, including the power supply system.

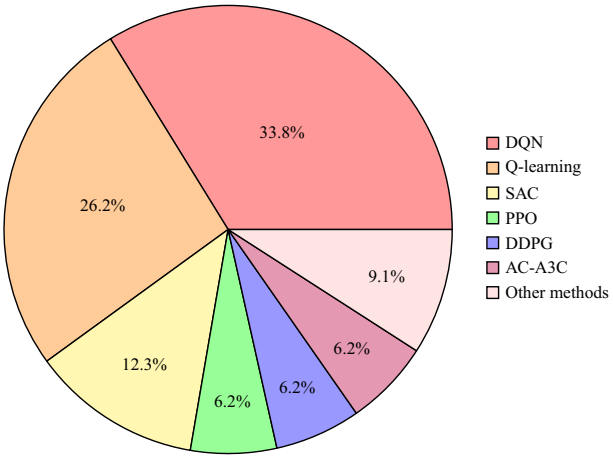


Fig. 9. Distribution of algorithms utilized in this review.

5.2. Comparison of RL/DRL algorithms applied to cooling system

Cooling systems account for approximately 40 % of energy consumption in data centers [112]. Reducing the energy consumption of this non-ICT support system will improve the power utilization efficiency (PUE) of the data center. Furthermore, optimizing the operation of the cooling systems can significantly influence the thermal conditions and cooling flow of ICT devices, leading to future reductions in the total energy consumption of the entire data center [113]. In this section, we will analyze the selected articles that use RL/DRL techniques to optimize the operation of the cooling system in the data center with the aim of reducing energy consumption. The following analysis not only provides an overview of how RL/DRL methods are applied to data center cooling system, but also investigates the specific aspects of each selected study in detail. This includes the formulation of the research problem and objectives, the energy-related outcomes, the benchmark comparisons, and the experimental setup.

5.2.1. The research problem and objective formulation

As illustrated in Fig. 8, 14 papers discussing cooling systems were identified. These studies focused on two main research problems (RQ4), each with different objectives (RQ5). The dominant category of the research problem focuses on optimizing cooling system operations in various scenarios to improve energy efficiency by utilizing various RL/DRL approaches. An interesting configuration involves using DRL to optimize the data center cooling system integrated with an active thermal management framework. For example, [60] explores the balance of aquifer thermal energy storage (ATES) while minimizing the total cost and maintaining the temperature range of the servers using a DRL agent. Similarly, [104] introduces Active Ventilation Tiles (AVTs) controllers to enhance the operation of the rack cooling system, achieving a trade-off between energy consumption and rack supply temperature distribution. An alternative scenario is integrating RL/DRL algorithms with prior physical knowledge to enhance the cooling system's energy efficiency. The study [75] integrates this knowledge by using big data, IoT sensor networks, and a digital twin model with the DRL algorithm. By leveraging historical and real-time data, this approach employs a Long Short-Term Memory (LSTM) network to predict temperatures, enabling the utilization of the DQN algorithm to effectively reduce the energy consumption of the cooling systems. Due to the strong relationship between the energy efficiency of data center cooling systems and the ambient temperatures at their locations, several studies have extensively investigated the efficiency of DRL algorithms in reducing cooling system energy consumption in tropical climates. Specifically, the study [101] focuses on optimizing the supply air temperature and relative humidity in a free-cooled tropical data center under defined boundaries, while [87] explores a single-agent DRL strategy with a floating set point approach to reduce the temperature threshold for tropical data centers based on a whole-building evaluation method. The study [69] proposes a multi-set point approach based on the DQN algorithm (DQN-MSP) to enable precise cooling control of the CRAC unit's air temperature, offering significant improvements in data center cooling energy consumption. Another key research direction in this category of literature emphasizes designing and comparing multiple state-of-the-art DRL algorithms for optimizing energy consumption while maintaining thermal conditions, as demonstrated in studies [66,93,102].

Meanwhile, the second research category shifts attention to the reliability of safety-aware DRL strategies, with the core aim of minimizing the energy consumption of the data center cooling system. These strategies are designed to ensure strict adherence to both soft and hard constraints during the learning and operational phases.

In [58], the study develops an end-to-end off-policy DDPG agent to optimize the cooling system using unprocessed and high-dimensional input data directly. Additionally, the study introduces the de-underestimation (DUE) validation mechanism for the critic network to address underestimation of overheating risks. In [106], the study focuses on incorporating residual physics using thermodynamic principles to guide the DRL agent's exploration process by estimating the desirable range of actions, ensuring future action safety. In addition, the study [49] develops safe cooling system operation by utilizing a model-based actor-critic DRL (MBAC) algorithm using two different models: a system transition model to predict the future system state, and a risk model to estimate the negative effects of executing an action. Furthermore, the paper [70] utilizes offline imitation learning and online post-hoc rectification techniques to develop three different versions of a safety-aware DDPG controller for the data center cooling system. Alternatively,

Table 4
Classification of algorithms by model type and study IDs.

Category	Algorithm type (RQ2)	Study IDs
Value-Based Model-Free	DQN	S4, S6, S7, S8, S10, S13, S15, S23, S29, S32, S34, S35, S37, S39, S42, S45, S49, S50, S53, S54, S55, S58, S22
	Q-learning	S11, S16, S17, S18, S22, S25, S28, S33, S38, S43, S44, S48, S51, S59, S62, S63, S64
	B3QN	S56
	PADQN	S5 , S27, S45
	SARSA	S38
	BDQ	S56
Policy-Based Model-Free	PPO	S14, S21, S36, S57, S60, S65, S47
	TRPO	S47
	Monte Carlo (REINFORCE)	S31
Actor-Critic (AC) Model-Free	SAC	S2, S9, S19, S20, S41, S56, S60, S65, S47
	A3C	S30
	AC	S1, S26, S61, S35
	DDPG	S12, S24, S40, S46, S35, S45, S60
	TD3	S47, S56
Learned Model-Based	PETS	S56
	MBAC	S3
Given Model-Based	None identified	

the study [111] leverages techniques like Lagrangian-based constrained DRL (CDRL) and reward shaping to satisfy soft constraints through extensive online learning. Also, within the same study, hard constraints are addressed by a parameterized shielding DRL algorithm (DRL-S), which projects unsafe actions onto safe action spaces. The ultimate goal of these studies in the second category is to design a safe cooling system for data centers, reducing energy consumption while effectively maintaining thermal constraints. The insights from this section are summarized in Table A.8.

5.2.2. The energy related outcomes

The primary motivation of this study is to address how the proposed RL/DRL algorithms enhance the energy efficiency of data centers (RQ8). The results related to energy efficiency have been carefully and thoroughly analyzed. Given the diversity of research problems and objectives addressed in the identified cooling system studies, the reporting methods for energy-related outcomes vary significantly. Some studies express the improvements in energy consumption when implementing the RL/DRL algorithm as a percentage reduction in energy consumption, compared to the baseline controller (e.g., DefaultE+) [93,102,111].

In addition, other studies compare the energy saving percentage of their proposed RL/DRL strategies to some other benchmark controllers, including DRL and non-DRL algorithms [49,70,87]. Energy efficiency is also reported in terms of improvements in key data center performance metrics, such as power usage effectiveness (PUE), compared to baseline controllers (e.g., DefaultE+) [58] or state-of-the-art controllers [66], while other studies use the PUE to evaluate the differences in energy consumption before and after applying the proposed RL/DRL algorithms [75]. Other studies focus on energy cost reductions rather than energy consumption savings [60].

Moreover, combining RL/DRL strategies with advanced setups, such as AVT systems [104] and physics-guided DRL with shielding [106], highlights the potential of RL/DRL in performing a trade-off analysis between energy efficiency and system performance. Furthermore, some studies demonstrate energy savings while maintaining thermal constraints, either by increasing the average supply air temperature of the CRAC units [69] or by raising the temperature and relative humidity thresholds [101]. A more detailed analysis of additional objectives combined with the energy efficiency will be provided in Section 6. A detailed summary of this section's findings is presented in Table A.8.

5.2.3. The benchmark comparisons

The distribution of benchmark algorithms used in the cooling systems studies for energy-related results comparison is illustrated in Fig. 10.

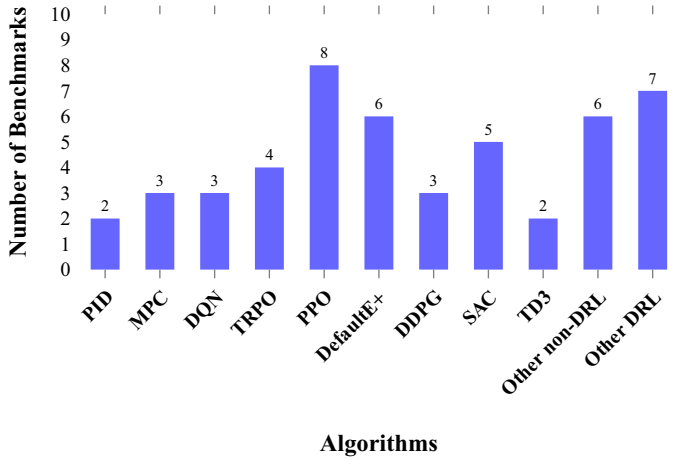


Fig. 10. Number of benchmarks in the literature for cooling system.

Analyzing the statistical data reveals two distinct groups. The first group involves the use of DRL algorithms due to their adaptability as benchmarks for comparison, with PPO being the most widely used, appearing in eight studies. Other prominent DRL algorithms include SAC (used five times), TRPO (used four times), DDPG (used three times), DQN (used three times), and TD3 (used twice). The second group consists of non-DRL algorithms, where the built-in EnergyPlus baseline controller (DefaultE+) was used in five studies, the classical PID controller was used twice, and the optimal model predictive controller (MPC) was used three times. Other DRL and non-DRL algorithms, including those used as benchmarks only once, are also considered. Table 5 outlines the benchmark algorithm comparisons (RQ6) for each selected cooling system study, including both DRL and non-DRL algorithms.

5.2.4. The experimental setup

Among the 14 selected cooling system studies, only one study directly implemented the proposed DRL strategy on a real-world data center [104]. In contrast, the remaining studies tested the designed DRL algorithms in simulated environments, highlighting a gap in direct real-world application and validation. These simulations utilized either real-world datasets, synthetic datasets, or a hybrid approach combining both. The EnergyPlus building energy simulation program [117] emerged as a primary tool for simulating energy consumption in data

Table 5
Selected cooling system studies experimental setup.

ID	Environment (RQ3)	Data source (RQ3)	Data type (RQ3)	Benchmarks (RQ6)	Platform (RQ7)
S3	Simulation	Simulated a typical data center room with Alibaba's 2018 cluster data	Real-world	MBRL-MPC, MBHC	Unspecified CFD simulator, Python (PyTorch)
S12	Simulation	National Super Computing Centre (NSCC) of Singapore	Real-world	DefaultE+, Two-stage (TS), A3C, TRPO	EnergyPlus, Python (Scipy)
S14	Simulation	Naviair data center (the Danish airspace control company)	Real-world	No reward PPO, Delayed reward PPO, Uniform future PPO, Trend-based future policy to estimate the return	Python (OpenAI Gym)
S20	Simulation	Simulated liquid-cooled data center with unspecified real-world data set	Real-world	PID, MPC, DQN, TRPO, PPO	Matlab (Simscape)
S23	Simulation	Simulated a small data center with a real-world dataset from the PlanetLab system	Real-world	DQN-SSP, PPO-MSP, DDPG-MSP	6SigmaRoom, CloudsimPy, Python
S24	Simulation	Four simulated configurations of CW- and DX-cooled data centers under two climate conditions	Synthetic	For the first three proposed controllers: DefaultE+, Reward shaping DDPG, Simplex DDPG, Projection post-hoc rectification DDPG For the fourth controller: PID, Vanilla DDPG, Reward shaping DDPG	EnergyPlus, OpenFOAM, Python (OpenAI Gym and PyTorch)
S29	Simulation	Simulation for real-world data center room located in Shenzhen	Real-world	Comparison of DC energy efficiency metrics before and after the DRL strategy	6SigmaRoom, Autodesk Revit, Python
S41	Simulation	Simulated mid-tier stand-alone data center located in a tropical climate region	Synthetic	DefaultE+, Load Aware, Temperature Aware, Joint-IT, Multi-Agent DRL, TD3, PPO, TRP, various versions of SAC	EnergyPlus, Python
S47	Simulation	Simulated medium-sized DC with two zones, a direct expansion cooling coil, and a chilled water cooling coil	Synthetic and real-world	DefaultE+, TD3, PPO, TRPO, SAC	EnergyPlus, Python (OpenAI Gym)
S55	Simulation	A real free-cooled data center located in a tropical zone	Real-world	Hysteresis-based controller, MPC	Matlab, Python (Keras and TensorFlow)
S56	Simulation	Simulated data center test bed developed in [114]	Synthetic	DefaultE+, PETS, BDQ, PPO, SAC	EnergyPlus, Python (PyTorch)
S58	Real-time	Inner Mongolia Meteorological Information Center (IMMIC)	Real-world	DL, DN, DQN	Python (TensorFlow), Real-time
S60	Simulation	Simulated data center test bed developed in [115]	Synthetic	SAC, RP-SAC, DDPG, RP-DDPG, PPO, RP-PPO, Lagrangian-based safe DRL, Physics	EnergyPlus, Python (PyTorch, TensorFlow)
S65	Simulation	Simulated data center test bed developed in [116]	Synthetic	DefaultE+, PPO, SAC, PPO-Lag	EnergyPlus

center cooling systems, often integrated with various Python libraries to implement DRL agents. Other simulation environments utilized include Computational Fluid Dynamics (CFD) simulators such as OpenFOAM [118] and 6SigmaRoom [119], which offer detailed modeling of airflow and thermal dynamics. Furthermore, MATLAB, along with its advanced toolboxes like Simulink and Simscape, was frequently employed to simulate the operational processes of data center cooling systems, providing a robust platform for evaluating control strategies and optimizing system performance. Table 5 presents a comprehensive overview of the experimental setup, including the environment, dataset source and type (RQ3), and platform (RQ7) for all identified studies on cooling systems.

5.3. Comparison of RL/DRL algorithms applied to ICT systems

Over the past few years, data centers have grown significantly in size and complexity driven by the rapid advancements in ICT systems. The advancements involve a wide range of devices, including high-performance servers, processing units such as CPUs and GPUs, advanced memory units, and storage arrays [120]. This technological progress has enabled data centers to support more complex operations, such as training large language models (LLMs) and real-time data processing. As a result, improving the energy efficiency of ICT systems has become a critical priority, not only to enhance the performance and scalability of data centers but also to minimize energy consumption and operational costs. In this section, we will comprehensively examine the role of RL/DRL algorithms in tackling energy efficiency challenges within ICT systems as identified in the literature.

5.3.1. The research problem and objective formulation

The majority of the identified papers in this review focus on ICT systems, specifically 40 studies. The research problems (RQ4) and objectives (RQ5) of these studies can be categorized into the following areas:

Scheduling optimization: A considerable number of existing studies discuss the scheduling optimization challenge in a DC environment using RL/DRL approaches; however, few studies have explored the energy efficiency aspects of applying these algorithms. The three main types of RL/DRL algorithms applied to the scheduling optimization problem in the identified studies are: jobs scheduling, tasks scheduling, and resources scheduling.

Jobs scheduling: Job scheduling refers to the process of assigning and allocating the entire arriving job which may consist of one or multiple tasks to the DC resources, aiming to manage workloads with a high-level approach. Traditional job scheduling mechanisms often struggle to cope with extensive, heterogeneous DC environments, especially in cases involving long-lasting jobs. This limitation leads to inefficiencies in energy consumption and resource management. Three studies [56,77,85] have addressed this challenge by proposing RL/DRL algorithms. The primary approach to handling this challenge dynamically involves considering real-world constraints, such as job dependencies and QoS levels, to minimize energy consumption and carbon emissions in data centers.

Tasks scheduling: Tasks are the components of the jobs that typically need to be performed in a specific order due to their interdependence. Task scheduling refers to the process of managing the execution

of individual tasks within a job in a low-level approach. The main objective of the task scheduling studies is to select the optimal DC resource for task execution, ensuring compliance with time and QoS constraints. Ten studies were identified that discussed the task scheduling problem highlighting three main approaches:

- Dependency- and workflow-oriented RL/DRL task scheduling approaches [72,82,94,110].
- Heterogeneous cloud DC online RL/DRL task scheduling approaches [67,103,109].
- Adaptive and hybrid RL/DRL task scheduling approaches [50,54,59].

Resources scheduling: While task and job scheduling focus on the DC workload, resource scheduling concentrates on the physical (e.g., servers) or virtual (e.g., VM) infrastructure level of the DC. The main aim of the resource scheduling process is to maximize resource utilization, and it does not directly consider job and task dependencies. Two studies specifically focused on addressing the resource scheduling problem [89,95].

Virtual machines and containers management: The virtualization of physical resources in data centers to meet the growing demands of workloads has received significant attention from researchers in recent years. Two primary technologies are commonly employed for virtualization: hardware-level virtualization (VM) in which each virtual machine (VM) utilizes a hypervisor to run its own operating system and applications. In contrast, operating system (OS)-level virtualization leverages the host system's kernel to create containers which share the host's resources [121]. In this review, we selected 14 studies focused on managing VMs and containers using RL/DRL algorithms and present the energy efficiency results. These studies address three key areas: VM consolidation, VM and container placement, and VM replacement.

VM consolidation: This refers to reducing the number of physical machines (PMs) required to operate the data center workload. This process includes three stages: workload detection (overutilization and underutilization), VM selection, and VM placement. By running multiple VMs on fewer PMs, several objectives can be achieved, including optimizing ICT resources, reducing operational costs, and minimizing energy consumption. Five studies in this review collection discuss the VM consolidation problem in data centers using RL/DRL algorithms with two main approaches:

- Centralized adaptive RL/DRL strategies [53,57,84,88]
- Multi-agent RL strategies [105]

VM and container placement: This is a sub-process of consolidation, where the objective is solely to decide the optimal location (PM) for a VM. It is applied at the PM (host) level rather than at the DC system level. Eight studies have been identified on this topic: seven for VM placement and only one for container placement [68].

VM replacement: This refers to the process of reassigning an already placed VM to a new physical machine (PM). This process is triggered by changes in the current state (e.g., overloading, failures). It is also considered a sub-process of VM consolidation, enabling VM migration. Among the selected studies, only one specifically addressed this issue, proposing a novel approach that combines fuzzy logic with an RL algorithm to enhance decision-making and adaptability in this process [74].

Two studies combine the two aforementioned categories as a research problem, focusing on VM scheduling by allocating tasks or jobs to VMs assigned to hosts, leveraging RL/DRL algorithms to optimize the scheduling process [79,90].

DCN traffic control: Data Center Networks (DCNs) play a critical role in ensuring the smooth operation of ICT systems. However, they often suffer from bandwidth surges, which degrade data center performance and significantly increase energy consumption. Traditional methods to address these issues are limited in their adaptability and fail to dynamically handle sudden network traffic fluctuations, leading to substantial energy waste. RL/DRL algorithms offer effective approaches to tackle these challenges. Four studies have been identified that explore solutions to this problem, each employing a unique structural RL/DRL approach:

- Combining LSTM networks for traffic prediction and proactive RL/DRL agents to optimize traffic control and energy efficiency [73,86].
- Formulating the problem as an MILP model to define the optimal solution space and integrating RL/DRL algorithms to find near-optimal solutions dynamically [55].
- Employing Software-Defined Networking (SDN) and RL/DRL to dynamically schedule traffic flows, aiming to reduce energy consumption while maintaining an optimal Flow Completion Time (FCT) [107].

Multi-objective framework: Five studies are identified here that address job/task scheduling, task offloading, and resource allocation as multi-objective research problems. The resources considered in these studies include containers [65,81], multi-user, multi-data center resources [99], and general data center resources [83,108].

A detailed summary of the identified ICT studies' research problems (RQ4) and objectives (RQ5) is presented in Table A.9.

5.3.2. The energy related outcomes

As energy efficiency is the primary focus of this review, a comprehensive analysis of the energy efficiency outcomes of using RL/DRL algorithms in ICT systems in the identified studies is presented in Table A.9. This table answers this review's RQ8 and demonstrates that the proposed RL/DRL algorithms consistently outperform baseline and benchmark non-RL/DRL methods in terms of energy efficiency. The reported energy efficiency improvements range from small percentages (1 %–3 %) to significant enhancements (over 60 %), depending on the specified scenario and context, such as varying VM/task loads, DCN traffic sizes, or the use of real-world or synthetic datasets. The majority of the studies reported achieving energy efficiency as a percentage improvement when compared to benchmark algorithms.

Additionally, some studies highlighted energy efficiency enhancements in terms of scalability and dataset-based performance. For instance, studies [62,108] focus on performance across diverse datasets and scalability metrics. Other studies compare the achieved energy savings in multiple experimental setups or configurations. [67] investigated task scheduling across three distinct scenarios with 10, 50, and 100 servers, examining the impact of server configurations on energy efficiency. [88] explored VM consolidation under different workloads, assessing its impact on resource utilization and energy consumption. [86] analyzed DCN traffic control with both more than 70 nodes and fewer than 70 nodes, assessing performance across different network sizes. [109] conducted task scheduling across two different task counts and varying numbers of VMs, evaluating the performance under diverse configurations.

In addition, a few studies presented a generalized approach without explicitly referencing benchmark algorithms. For instance, [74] reported energy savings in a generalized context, providing insights into the potential applicability of the proposed RL algorithm.

5.3.3. The benchmark comparisons

Each research problem discussed in the identified studies of the ICT system was compared to other baseline or state-of-the-art benchmark methods commonly used in the respective problem domain. As presented in Fig. 11, the most commonly used baseline method for scheduling optimization studies was the RANDOM method. In this method, jobs/tasks/VMs were assigned to resources without considering optimization criteria. This approach is simple and achieves unbiased

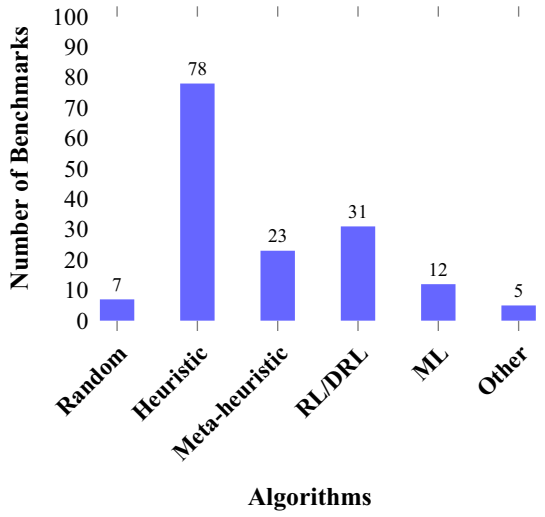


Fig. 11. Number of benchmarks in the literature for ICT system.

scheduling; however, it is inefficient as it overlooks critical DC metrics such as energy efficiency, quality of service (QoS), and workload balancing. This method was used in seven studies as a baseline for comparison with the proposed RL/DRL algorithms in the scheduling optimization identified studies. Additionally, heuristic-based algorithms were widely used as benchmarks to evaluate the proposed RL/DRL algorithms for various ICT research problems. For scheduling research problems, the Round-Robin (RR) method was highlighted as the primary heuristic-based method for performance comparison. Greedy algorithms, including First-Fit (FF), Best-Fit (BF), and their variants, were the main benchmarks for VM management research problems. Elastic-Tree was a common benchmark for DC network traffic control problems.

Approximately 78 additional heuristic-based algorithms were employed as comparison methods across all the research problems discussed in ICT systems. Meta-heuristic methods were also used 23 times as evaluation methods. These included Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Genetic Algorithms (GA), and their variants, applied to various ICT system research problems. Machine learning algorithms were occasionally utilized in a limited number of identified studies as benchmark methods, particularly for VM management.

Other RL/DRL algorithms developed in previous studies were used 31 times for comparison with newly proposed algorithms, demonstrating internal comparisons within RL/DRL approaches in the identified studies. Finally, some specially designed algorithms were also employed. Table 6 outlines the benchmark algorithm (RQ6) comparisons for each selected study on ICT systems.

5.3.4. The experimental setup

CloudSim [122] and its variant WorkflowSim [123], an extended and optimized version of CloudSim designed for dependent task workflows, were used as simulation environments in approximately 50 % of the identified studies focusing on scheduling optimization and VM management research problems in DC ICT systems. In addition to these tools, programming languages such as Java and Python were frequently employed for simulation experiments in multiple studies.

MATLAB was used as the simulation environment in four studies. However, six studies did not specify the simulation environment used. On the other hand, several real-world datasets from large-scale data centers like Google, Wikipedia, and Alibaba, as well as smaller data centers such as the National Supercomputing Centre (NSCC) of Singapore and the Nottingham University Data Center, were utilized as data sources in the identified studies. Moreover, well-known datasets such as PlanetLab and the CoMon project were also employed for simulation experiments. Synthetic datasets were another key data source, enabling controlled and customized testing scenarios. Table 6 provides a comprehensive overview of the experimental setup, encompassing the simulation environment, the sources and types of datasets (RQ3), and the platforms used (RQ7) in all the identified studies on ICT systems.

5.4. Comparison of RL/DRL algorithms applied to optimizing integrated data center systems

Developing an accurate, intelligent, and real-time DC environment requires seamless integration of all systems, including the cooling, ICT, and power supply systems. The joint optimization of these systems has become a promising research direction, aiming to achieve multiple objectives across multiple systems using advanced optimization strategies.

Among these, RL/DRL algorithms have emerged as powerful approaches, demonstrating significant potential in addressing the complexities of integrated DC systems. This section delves into a detailed analysis of 11 identified studies that leverage RL/DRL algorithms for the joint optimization of DC systems. A vital aspect of the identified studies in this section lies in formulating these studies as a multi-objective research problem across multiple systems. To further enrich this discussion and align with the growing interest in this field, we define the key elements of the Markov Decision Process (MDP) models employed in these studies, highlighting their critical role in achieving efficient and effective system integration.

Various identified studies explored the integration of the ICT operation optimization with energy-efficient cooling system controlling as a research problem (RQ4) with different objectives (RQ5). [71] investigates the implementation of a decentralized strategy to simultaneously optimize the cooling system and the VM placement. Additionally, scheduling optimization combined with cooling system control is another prominent research focus. Task scheduling was discussed in [52,91,92], whereas job scheduling was examined in [52,80]. In both cases, the scheduling process is integrated with the optimization of the cooling system. On the other hand, three studies [47,61,96] examined the workflow scheduling of DC powered by renewable energy systems (RES). The primary objective in these studies is to optimize energy consumption from RES during the execution of DC workloads. In study [48], a DRL strategy was applied to optimize the cooling system by integrating it with the power supply system using real-time electricity pricing (RTP). Finally, global optimization using a multi-agent approach to enhance energy efficiency across more than two DC systems was addressed in a recent study [98].

The majority of the identified papers report results related to energy efficiency (RQ8) of the developed RL/DRL algorithms as a percentage of energy savings compared to baseline algorithms. For example, [48] reported a slight improvement in energy savings compared to a PID controller, while [91] compared the energy efficiency results of the proposed algorithm with a controller designed based on domain expert knowledge, achieving up to 30 % energy savings. Another method of reporting energy efficiency results involves using data center efficiency metrics, such as PUE. This approach was demonstrated in [52,80], where the proposed RL/DRL algorithms enhanced energy efficiency compared to benchmark algorithms. Table A.10 provides an overview of the research problem (RQ4), related objectives (RQ5), and energy-related outcomes (RQ8) of the identified joint optimization studies.

The proposed joint RL/DRL algorithms were compared against various benchmark algorithms. Multiple studies evaluated the performance of the developed RL/DRL strategies against state-of-the-art individual optimization techniques, such as ICT algorithms (e.g., random or

Table 6
Selected ICT system experimental setup.

ID	Environment (RQ3)	Data source (RQ3)	Data type (RQ3)	Benchmarks (RQ6)	Platform (RQ7)
S4	Simulation	Google cluster	Real-world	RR, B, MAD, DRL-DTM, DRL-DTA	NA
S7	Simulation	Google cluster	Real-world	FF, MFFD, PABFD, RL-DC, UP-VMC	EnergyPlus, CloudSim
S8	Simulation	Google cluster	Real-world	RR, HDRL, DRL-Cloud, MO-DQN	Python (TensorFlow)
S9	Simulation	Abilene, Geant, and Synthetic topology datasets	Synthetic and Real-world	TEDO, TEDI	Java, Python (TensorFlow)
S10	Simulation	National Supercomputing Center (NSCC) of Singapore	Real-world	RR, Job consolidator, Online optimizer with two different reward functions	NA
S11	Simulation	CoMon Project	Real-world	LR-MMT, VDT-UMC, DTH-MF	CloudSim
S13	Simulation	Google cluster	Real-world	RR, HDRL, DRL-Cloud	NA
S16	Simulation	Amazon EC2 and Simulated dataset	Synthetic and Real-world	FFD, BFD, GRVMP, GMPR, NSGA-II, RLVMP	CloudSim
S17	Simulation	Simulated dataset	Synthetic	VMPMORL, EVCT, VPME, AFED-EF	CloudSim
S18	Simulation	GWA-T-12 Bitbrains	Real-world	MOPSO, MOACO, VMPORL	MATLAB
S19	Simulation	Simulated tasks following an exponential workload distribution	Synthetic	Cloud, PREM, RANDOM, REQ	Python (PyTorch and Gym)
S21	Simulation	Open-source: BitBrains, Scientific workflows: Ligo, Montage, Cybershake	Real-world	RR, RF, GRR, GRF, Tetris, RLScheduler, ACS	NA
S22	Simulation	Simulated dataset	Synthetic	GA, ACO, SA, FFD	Java, CloudSim
S26	Simulation	Google cluster	Real-world	RR, RANDOM, SO, GJO	NA
S27	Simulation	Simulated dataset using a K-port FatTree topology	Synthetic	Greedy-ElasticTree, LSTM + DRL, DDPG	Python (TensorFlow, Keras)
S28	Simulation	Nottingham University, Gaussian distribution Synthetic datasets	Synthetic and Real-world	MOVMrB, RLVMrB, VMPMORL	CloudSim
S30	Simulation	PlanetLab dataset, Amazon EC2 instance configurations	Synthetic and Real-world	MOVMrB, RLVMrB, ADVMC	MATLAB
S31	Simulation	Alibaba Cloud	Real-world	FIFO, Ideal MPC, Tetris	Python (TensorFlow)
S32	Simulation	Azure 2017 workload	Real-world	HGP, IQR-MMT, MAD-MMT, RLR-MMT, GA	Python (PyTorch, Gymnasium, Scikit-learn)
S33	Simulation	Ligo, Genome, Cybershake, Montage, and Sipht datasets	Real-world	EcoCloud, KMI-MRCU, AFED-EF	Java
S35	Simulation	Simulated two common datasets	Synthetic	DSTS, LSTM, RF, CNN	CloudSim
S36	Simulation	Alibaba Cluster	Real-world	EINFORCE, FF, RANDOM, Tetris	Python (TensorFlow, NumPy, Matplotlib)
S37	Simulation	Simulated dataset	Synthetic	Small task sizes: Load Aware, FFO-EVMM, MIMT, DQN. Medium task sizes: FFO-EVMM, MIMT, L-No-Deaf, Worn-Dear, DQN. Larger task sizes: FFO-EVMM, MIMT, multiple PSO variants, DBC, EDF	CloudSim
S38	Simulation	PlanetLab dataset	Real-world	PowerAware VM consolidation	CloudSim
S39	Simulation	Simulated dataset	Synthetic	RANDOM, RR, EDF	Python (PyTorch)
S40	Simulation	Packet trace files from three data centers, generated using Wireshark	Real-world	Shortest-path-based routing, Gurobi optimizer	Python (Keras, TensorFlow)
S42	Simulation	PlanetLab Monitoring	Real-world	IQR, MAD, THR, LR, PABFD	CloudSim
S43	Simulation	Simulated dataset	Synthetic	RoFFR, CSLB, TDBS	WorkflowSim, Python
S44	Simulation	1998 FIFA World Cup Dataset, UNSW-17 Network Traffic Dataset	Real-world	VPBAR, LRR-MMT, DTH-MF, VMTA, Megh, EQBFD-0.1, EQBFD-0.3	CloudSim
S48	Simulation	Simulated dataset	Synthetic	MMS-RANDOM, MMS-FAIR, MMS-GREEDY	CloudSim
S49	Simulation	Google cluster	Real-world	RANDOM, Round Robin (RR), MoPSO	Python (TensorFlow)
S51	Simulation	Simulated dataset	Synthetic	Multi-objective optimization algorithms: MGGA, VMPACS, VMPMBBO, ICA-VMPLC, CVP. Single-objective optimization algorithms: FFD, OEMACS	MATLAB
S53	Simulation	Simulated dataset	Synthetic	Job scheduling: RANDOM, RR, Greedy, MoPSO Resource allocation: RANDOM, RR, MLF, FERPTS	Python (TensorFlow)
S54	Simulation	Production-quality cloud DC, simulated dataset	Synthetic and Real-world	FF, Dot Product, Norm2 heuristics	Python (NumPy, PyTorch)
S57	Simulation	Google cluster	Real-world	Tetris, H2O-Cloud	NA
S59	Simulation	CoMon Project (PlanetLab data)	Real-world	NPA, PABFD, IGGA, E-Eco	CloudSim
S61	Simulation	Wikipedia trace files	Real-world	ElasticTree, CARPO, FCTcon, Optimal (it is not practical in use)	Python (Keras)
S62	Simulation	Montage, Cybershake, Sipht, Inspiral datasets generated using the Pegasus Workflow Generator	Real-world	MPC, ETF, Lr-RL, Q-SCH, QL-HEFT	CloudSim
S63	Simulation	Google Cloud Jobs dataset (GoCJ)	Real-world	PSO, MVO, EMVO	MATLAB, Python (PyTorch)
S64	Simulation	Sipht, Inspiral, Cybershake datasets generated using the Pegasus Workflow Generator	Real-world	MPC, ETF	WorkflowSim

Table 7
Selected integrated studies experimental setup.

ID	Environment (RQ3)	Data source (RQ3)	Data type (RQ3)	Benchmarks (RQ6)	Platform (RQ7)
S1	Simulation	Pegasus workflow framework	Synthetic	Random, Green-Opt (Greedy), Common-Actor	CloudSim, Python (Keras)
S2	Simulation	Weather: Collected from Denmark Electricity pricing: Danish electricity spot market	Real-world	Other RL Controllers (For SAC and PPO), PID controller	EnergyPlus
S5	Simulation	LLNL Thunder	Real-world	ICO, MPC, Joint optimization (JCO), Original-DQN	Matlab, 6SigmaDCX, TensorFlow
S6	Simulation	LLNL Thunder	Real-world	PADQN, E-QL	Matlab, 6SigmaDCX, TensorFlow
S15	Simulation	Workload: Google Cluster dataset (GCD). Renewable energy: National Renewable Energy Laboratory (NREL)/NE-3000 wind turbines. Electricity Price: The US EIA. Carbon Footprint: The US Department of Energy Electricity Emission Factors	Real-world	Greenpacker, LECC, ADVMC, ADVMC-RES	Python
S25	Simulation	PlanetLab, Google Cluster	Real-world	DeepEE, Deep-Q with LSTM, ETAS, Improved Genetic, Hierarchical Deep-Q, MPC	CloudSim integrated with four CRAC units, and perforated floor tiles to simulate realistic cooling dynamics
S34	Simulation	A simulation-based data set	Synthetic	Schedule: Single-agent method, Hybrid DQN, Independent DQN, Original DQN	6SigmaDC, CloudSimPy
S45	Combining real-world and simulation	Operational data from Singapore's National Supercomputing Center	Real-world	Based on expert domain knowledge algorithm. Heuristic Algorithms: For independent IT or cooling optimization. Thermal-Unaware Scheduling: Traditional task scheduling without considering thermal dynamics	6SigmaRoom, EnergyPlus
S46	Simulation	Google Cluster data	Real-world	Random, RR, PowerTrade, DeepEE	Python (OpenAI Gym and TensorFlow), Matlab
S50	Simulation	Wiki data center	Real-world	Static, Random, K-means	Python (PyTorch)
S52	Simulation	Simulated dataset	Synthetic	Non-optimization: No algorithm-based control, Non-algorithm optimization: Logic-based manual controls	NA

heuristic approaches) and traditional cooling control methods, including PID and Model Predictive Control (MPC). Additionally, other studies compared the results with joint optimization approaches. Furthermore, several studies benchmarked the outcomes against other RL/DRL algorithms proposed in previous research.

The tools discussed in Sections 5.2.4 and 5.3.4 were similarly employed in the joint optimization studies to create simulation environments. These include the EnergyPlus building energy simulation program [117] and the Computational Fluid Dynamics (CFD) simulators, 6SigmaRoom [119], which were utilized for cooling systems. CloudSim [122] served as a simulation environment for the ICT system. Furthermore, Python, along with its extensive libraries, served as the main programming language for implementing RL/DRL algorithms, while MATLAB was also employed in several studies for simulation and analytical tasks. Table 7 summarizes details of experimental setups in joint optimization literature: simulation environments (RQ3), platforms (RQ7), and benchmarks (RQ6).

5.5. The MDP elements

As detailed in Section 3, the Markov Decision Process (MDP) provides the foundational structure for modeling the RL/DRL environment. The key components of the MDP are: the state space $\{S\}$, the action space $\{A\}$, and the reward function $\{R\}$. In the context of the identified joint optimization problem, the MDP features a large and complex state space, as well as a mixed action space encompassing both discrete and continuous actions. Furthermore, the reward function guiding the RL/DRL agent in these studies consists of multiple terms to capture the various systems within the DC environment. This highlights that the MDP for joint optimization studies is considerably more complex than in studies addressing only one system. Table A.11 provides a comprehensive summary of the MDP components in the joint optimization studies.

6. Other objectives combined with energy efficiency in the identified studies

Besides energy efficiency objectives in the identified studies, other objectives have been investigated. It is essential to highlight these objectives which will shape the direction of future efforts in the field of multi-objective optimization. The RL/DRL algorithms have proven their effectiveness in resolving the conflicts between objectives in several identified works. For instance, in [95], where multi-objective optimization aims to balance the energy consumption of various numbers of tasks (between 100 and 250 tasks) and the average task makespan. Moreover, [92] examines the classical trade-off between quality of service (QoS), resource utilization, and energy consumption. Fig. 12 outlines a taxonomy of other optimization objectives integrated with enhancing the energy efficiency of the data center systems. Although the majority of the identified studies address data center energy efficiency enhancement aspects as the core research objective, some studies combine this objective with other environmental metrics, which can directly improve the operation mode of the data center and reduce its negative impact on the surrounding ecosystems in terms of carbon footprint and RES utilization.

In contrast, the identified studies examine the proposed RL/DRL strategies for ICT and cooling in terms of system performance. In one dimension, these strategies refine time-related aspects, including

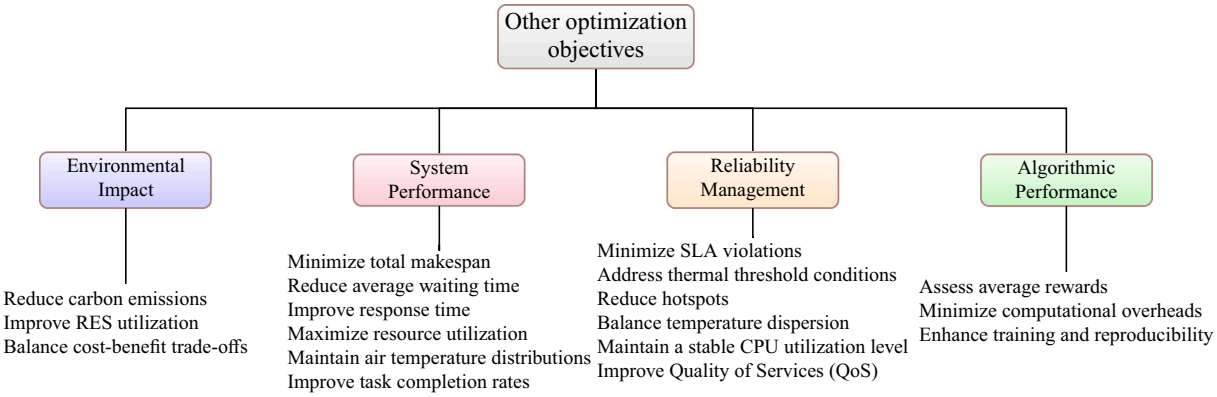


Fig. 12. Taxonomy of the other objectives than energy efficiency targeted in the included articles.

makespan (execution time), waiting time, and response time. In another dimension, they enhance resource efficiency, including resource utilization, cooling system air temperature distributions, and task completion rates. Another objectives category integrated with the energy efficiency improvement objective in many studies is resilience optimization. For ICT systems, these objectives include optimizing Service Level Agreement (SLA) violations, Quality of Service (QoS), and consistent CPU utilization. Resilience is also vital for data center cooling systems, involving aspects such as addressing thermal threshold conditions, reducing hotspots, and balancing temperature dispersion. Finally, the objectives related to the performance of the proposed RL/DRL methods are addressed in several identified works, with a primary focus on key aspects such as reward optimization, computational efficiency, and the reproducibility of the algorithms employed.

These objectives emphasize the significant impact of choosing rewards as an indicator of effective decision-making, boosting the scalability and real-time applicability of these algorithms by minimizing computational requirements, and validating the accuracy and universal applicability of these algorithms.

7. Research gaps, open challenges and future directions

Many of the studies reviewed in this work have recognized the potential of utilizing RL/DRL strategies for optimizing data center systems, demonstrating their success in reducing energy consumption and enhancing overall performance. However, despite these advancements, several obstacles must be addressed and overcome to achieve more effective, robust, and stable solutions. Based on the current review, this section identifies key open challenges and explores research gaps and potential future directions across three critical dimensions: real-time validation, standardized energy efficiency reporting metrics, and advancements in DRL multi-agent approaches.

7.1. Dependence on simulated environments

Exploring the experimental setups (RQ3). tools, frameworks, and platforms (RQ7) discussed in the literature reveals a growing trend of leveraging real-world datasets from diverse sources to construct simulated environments using a variety of software tools and programming languages. These environments are then used to evaluate the developed RL/DRL algorithms in terms of achieving the research objectives. However, a critical gap remains in the implementation of real-time validation, which is essential to prove the practical capability and robustness of the proposed algorithms. Real-time validation serves as a vital benchmark to ensure that the algorithms can operate effectively under dynamic conditions. The absence of real-time validation in 95 % of the identified studies in this review represents one of the most significant limitations in the current adoption of RL/DRL strategies for cooling and ICT systems. The main barrier of adapting the real-world validation framework in data center applications is the reward-shaped behavior of the RL/DRL agent [31], which can cause a substantial safety and security problems, where failure could lead to system damage or even data center service disruptions [124]. Also, the complex and varying nature of real-time energy management scenarios, further complicates the application of RL/DRL on a larger scale, as existing models may lack the adaptability needed to respond effectively to varying load and environmental conditions. To cope with real-world validation, integration between simulation implementation and real-world testing is needed to enable more reliable and scalable deployments. This approach can mitigate safety risks and enhance the security of RL/DRL algorithms. In summary, while the implementation of RL/DRL algorithms in simulated environments has proven invaluable, advancing to real-time validation is an open research gap for unlocking the full potential of these algorithms in data center environments.

7.2. Evaluation the energy efficiency results

The energy-related results (RQ8) in the identified literature highlight the current trend of reporting these results based on the percentage reduction in energy consumption compared to baseline or benchmark results. However, more comprehensive and multi-scale reporting metrics are needed to reflect the functionality of the proposed optimization algorithms to enhance energy efficiency. Advanced data-center-specific energy efficiency metrics can be employed, including:

- **Power usage effectiveness (PUE)**: This method was the first energy efficiency data center metric proposed by [125]. It measures the ratio of the data center's total energy consumption to the ICT system energy consumption. The PUE value is defined as a dimensionless quantity between 1.0 and infinity [126]. A PUE value close to 1.0 indicates optimal energy efficiency, indicating that most of the consumed energy is dedicated to ICT operations rather than other systems. Although used in some of the studies identified in this review, this metric does not account for objectives related to other data center systems, such as integrating renewable energy sources (RES) into the power supply or implementing waste heat recovery in cooling systems.

- **Energy reuse effectiveness (ERE):** This metric, introduced by [127], evaluates energy recovery in data centers by dividing the difference between total energy and reused energy by the ICT energy consumption. In an ideal scenario where ERE = 0, all waste heat is effectively recovered within the data center.
- **Data center energy productivity (DCeP):** It integrates both the energy consumption of infrastructure systems, including cooling, and the ICT system in order to evaluate the data center's energy efficiency [128]. DCeP is calculated as the ratio of useful work performed to the total energy consumption of the data center. Since data centers in various industries perform different types of work, defining useful work cannot be generalized and depends on the specific application.
- **Data center performance per energy (DPPE):** This metric combines four data center energy efficiency metrics: IT Equipment Energy (ITEE), IT Equipment Utilization (ITEU), Green Energy Coefficient (GEC), and Data Center Infrastructure Efficiency (DCiE) [129]. This metric measures the energy consumption performance output relative to each unit inside the data center.

Comprehensive details about these metrics, along with additional related methods, can be found in specialized review articles dedicated to data center energy efficiency and performance evaluation [25,112,130].

Combining multiple metrics to report the energy efficiency results targeted by the proposed RL/DRL optimization strategies is essential. This approach ensures a comprehensive and multi-scale evaluation of performance across various dimensions.

7.3. Multi-agent DRL algorithms

The emerging developments in data centers have led to increasingly complex cooling and ICT systems. These sophisticated systems require innovative optimization approaches to achieve key operational objectives such as resource allocation, job/task scheduling, and efficient and safe thermal management. However, optimizing each system independently has proven insufficient in achieving the overarching goal of enhancing the overall operational performance of data centers. Multi-agent Deep Reinforcement Learning (MADRL) algorithms, which enable agents to interact with other agents operating within the same environment, present a promising approach for improving optimization and achieving more efficient solutions. While MADRL has previously been utilized in fields such as UAV optimization [131,132], power grid management [133,134], and games [135,136] , there is a growing interest in leveraging MADRL in other fields to address multiple objectives across multi-system environments. Several studies identified in Section 5.4 have explored the application of MADRL algorithms in data centers. However, achieving a global optimization of data center operations that accounts for all upstream and downstream energy consumption and recovery objectives remains a significant challenge. The primary complexity of employing MADRL in data center environments arises from the heterogeneous nature of the systems. While the cooling system operates as a continuous process, ICT systems are generally discrete. Combining the optimization of these systems using MADRL requires careful and accurate selection of suitable DRL algorithms tailored to these characteristics. A comprehensive review of the application of MADRL in optimizing multi-system environments can be found in [137–139].

Other common challenges in deploying more advanced algorithms to optimize data center energy efficiency include accurately modeling data center environments, managing computational costs, and achieving scalability. These challenges have restricted the application of advanced algorithms, such as RL/DRL, to small, isolated use cases in small-scale data centers, rather than facilitating broader adoption in large-scale facilities. In large-scale data centers, RL/DRL could be holistically integrated, as demonstrated in the implementation of Green Data Center Cooling Control using Physics-guided Safe RL in [91]. However, some previous studies have highlighted that traditional RL/DRL methods face considerable barriers in large-scale applications due to high computational demands and slow convergence rates, which further limit their scalability and practical implementation in real-world settings [140].

8. Conclusion

In this work, we carried out a systematic literature review following the PRISMA Protocol, focusing on optimizing the energy efficiency of data centers by leveraging RL/DRL algorithms. This review examines recent research from different perspectives in the context of the two main systems of the data center: the cooling system and the ICT system. In this review, a comprehensive analysis and synthesis of the 65 identified studies was conducted, addressing eight major research questions (RQs): the targeted system, the RL/DRL employed algorithm, the reporting metrics for energy-related outcomes, the experimental setups, including dataset source and type and the environment, the research problems, the main objectives, the benchmark algorithm comparisons, and the platforms and software used. A more in-depth discussion was conducted on the literature regarding the joint optimization, where the MDP elements were analyzed in detail. Additionally, a brief investigation of other objectives combined with energy efficiency was explored. Finally, we comprehensively analyzed the research gaps, open challenges, and future directions from three different standpoints. We hope this in-depth review will serve as a valuable roadmap for upcoming research in the field of optimizing the energy efficiency and performance of current and future data centers.

CRediT authorship contribution statement

Hussain Kahil: Writing – original draft, writing – review & editing, visualization, conceptualization, methodology, formal analysis, investigation, resources, and data curation. **Shiva Sharma:** Writing – original draft, writing – review & editing, visualization, conceptualization, and methodology. **Petri Välisuo:** Writing – review & editing, visualization, supervision, and project administration. **Mohammed Elmusratia:** Writing – review & editing and supervision.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Academy of Finland under grant number 353562.

Appendix A. Related tables

See Tables A.8–A.11.

Table A.8
Selected cooling system studies objectives and outcomes.

ID	Research problem (RQ4)	Objectives (RQ5)	Energy related results (RQ8)
S3	DC cooling system safe operations	Reduce energy while maintaining thermal constraints using transition and risk models	OCA configuration: 18.18 % energy savings compared to 11.92 % for the MBHC and 15.63 % for MBRL-MPC. ECA configuration: 10.94 %, and 13.18 % energy savings compared to the benchmark controllers, while reducing thermal violations by up to 48 %
S12	DC cooling system safe operations	Optimize cooling system via DDPG with neural networks and DUE validation to mitigate overheating risks	Simulation: Improved PUE by approximately 11 %. Reducing cooling costs. Trace-based experiment: 15 % cooling energy reduction
S14	DC efficient cooling system optimization	Minimize total energy costs while maintaining server temperature and balancing the aquifer thermal energy storage (ATES)	The Delayed algorithm increased energy cost by 1 %, while Cool and Cool 2.0 reduced it by 1.2 % and 12.5 %, respectively, compared to the baseline. Only Cool 2.0 achieved ATES balance, while the others negatively affected it
S20	DC efficient cooling system optimization	Optimize PUE and chip thermal stability while improving adaptability	In the first scenario, SAC agent reduced energy consumption 32.23 %, 9.86 %, 10.77 %, 6.95 %, 1.83 % while enhancing the PUE by: 0.051, 0.011, 0.013, 0.008, 0.002 compared to: PID, MPC, DQN, TRPO, PPO controllers, respectively. In the second scenario, SAC (300 s control interval) saved 3.45 % energy vs. SAC final value state. DCI-SAC (combined value state) achieved 4.4 % savings over final value state DCI-SAC and 9.48 % over SAC with a PUE reduction of 0.01
S23	DC efficient cooling system optimization	Optimize CRAC unit supply air temperature using a DRL multi-set point (DQN-MSP) method for thermal recognition and energy efficiency	DQN-MSP outperformed other benchmark DRL algorithms by 5.7 %, 2.4 %, and 4.2 %, in terms of reducing energy consumption. DQN-MSP increased the average supply air temperature of the CRAC units while maintaining thermal constraints
S24	DC cooling system safe operations	Develop a safety-aware DRL strategy using imitation learning and rectification to optimize energy and maintain thermal constraints	In the EnergyPlus environment, the three proposed Safari controllers achieved between 18 % and 26.6 % energy savings compared to the benchmarks while minimizing violations. In the OpenFOAM environment with non-uniform temperature distribution, the proposed Safari-4 controller achieved approximately 14 % energy savings compared to PID while maintaining thermal safety constraints
S29	DC efficient cooling system optimization	Integrating big data, IoT, and LSTM network with DRL to optimize energy consumption	Achieve improvements in the energy-saving metrics of the data center, specifically in PUE and WUE by around 2.55 %
S41	DC efficient cooling system optimization	Optimizing DC energy consumption using a floating set point and whole-building evaluation in tropical climates	In an environment setup proposed by [141], the proposed algorithm achieved 13.8 % energy savings, comparable to the 13 % savings recorded with SAC developed in that study. However, the other DRL algorithms used in [141] (TD3, PPO, and TRPO) achieved slightly better performance. It achieved energy savings by 5.5 % (full-load IT) and 3 % (part-load IT), which demonstrated that the server fan identified as the primary energy-saving component
S47	DC efficient cooling system optimization	Compare four state-of-the-art DRL algorithms for optimizing cooling in a medium-sized DC under varying weather conditions	Achieve energy savings of 13 % for SAC, 19 % for TD3, 18.3 % for PPO, and 14.6 % for TRPO compared with the baseline model-based EnergyPlus controller (DefaultE+)
S55	DC efficient cooling system optimization	Minimize energy consumption of air-free cooling system in a tropical DC while maintaining the supply air temperature and relative humidity (RH)	Reduce energy consumption across all scenarios compared to the benchmark hysteresis-based controller. As lower thresholds require more cooling demand, Energy consumption rose as temperature and RH thresholds decreased
S56	DC efficient cooling system optimization	Evaluation of four state-of-the-art DRL algorithms for dynamic thermal management in DC cooling systems, focusing on balancing energy consumption and thermal constraints across different scenarios	The evaluation highlights the DRL algorithms ability to balance energy efficiency and constraint satisfaction. Compared to the baseline EnergyPlus controller (DefaultE+), the proposed DRL strategy demonstrates energy savings influenced by the selected reward function, achieving up to 8.84 % in some cases
S58	DC efficient cooling system optimization	Optimize rack-level cooling by implementing multi-active ventilation tiles (AVTs) controllers, while integrating Dyna architecture for energy-temperature trade-off	Energy consumption was reduced by optimizing fan speeds and minimizing temperature variances. While exact percentages are not specified, a trade-off analysis between energy efficiency and AVT system performance was conducted by adjusting the weight parameter ω. Once stabilized, increasing ω prioritized energy savings by reducing the average fan speed
S60	DC cooling system safe operations	Introduce Residual Physics DRL approach (RP-SDRL) leverages thermodynamics to estimate desirable action ranges for guiding DRL exploration. It enhances learning and mitigates unsafe actions	The annual energy consumption comparison for the entire DC, including all subsystems, was conducted between the proposed SAC, RP-SAC, and the baseline physics algorithm. Results confirm that SAC and RP-SAC achieved over 10 % higher energy savings than the physics-based method
S65	DC cooling system safe operations	Propose a shielding DRL algorithm (DRL-S) that leverages techniques such as Lagrangian-based CDRL and Reward Shaping. Shielding can transform unsafe actions into safe action spaces	Compared to DefaultE+, PPO-based algorithms achieved better energy savings than the SAC-based algorithms. The best performance related to energy saving came from the proposed PPO-Lag algorithm, which can save between 8.12 % and 12.45 % of the energy compared to the baseline. PPO10 achieved the best energy saving among the reward shaping algorithms at 10.67 %. The PPO1 and SAC1 were excluded from the comparison as they cannot effectively handle the violations

Table A.9
Selected ICT system studies: objectives and outcomes.

ID	Research problem (RQ4)	Objectives (RQ5)	Energy related results (RQ8)
S4	Task scheduling	Optimize DC energy consumption while maintaining a high level of QoS using DRL hybrid task scheduling framework	Outperform the benchmark BF strategy by saving 14 % of energy. This is achieved by scheduling tasks more efficiently, thereby minimizing the number of active servers
S7	VM consolidation	Reduce energy by optimal selection and placement of VM using adaptive DRL algorithm	Achieve up to a 125.24 % improvement in energy consumption Compared to MFFD
S8	Task scheduling	Decrease task response time and increase resource utilization using an adaptive DRL task scheduling framework	The results show a 61.9 % improvement in energy consumption over the RR method when using the Google Cluster, while achieving 2.46 times better performance than RR when using the synthetic dataset, with faster response times in both cases
S9	DC network traffic control	Optimize NFV traffic for SFCs in DC, aiming to minimize energy consumption and maximize cost efficiency while meeting delay-guarantee constraints	In both scenarios: fixed and dynamic SFCs, the energy consumption of both proposed models (TEDI, and TEDO) was nearly identical. However, the TEDI outperforms the TEDO in terms of computation time in dynamic scenarios
S10	Job scheduling	Reduce energy consumption and improve thermal conditions in compute-intensive, long-lasting jobs DC environments	Effectively reduce energy consumption by more than 10 % and improved the running temperature of the processors by over 4 °C, while maintaining the same job processing throughput
S11	VM consolidation	Optimize resource allocation by achieving the balance between server energy consumption and application performance utilizing an adaptive RL framework	Reduce energy consumption compared to the LR-MMT and VDT-UMC methods. However, the DTH-MF method achieve the same level of energy consumption as the proposed PPR-RL approach. Nonetheless, the DTH-MF uses temperature as the upper limit, which can affect the thermal conditions of hardware, making it unsuitable for real-world DCs
S13	Task scheduling	Improve task scheduling while maintaining QoS using comprehensive of a hierarchical and hybrid online DRL agent in warehouse-scale DC	Achieve up to a 47.88 % improvement in energy consumption compared to baseline state-of-the-art methods
S16	VM placement	Cluster VM using the K-means algorithm and employs a multi-reward RL approach to reduce energy consumption and improve resource utilization	Achieve 14 % better results compared to the GMPR algorithm using a synthetic dataset, and up to 26 % improvement using real data (Amazon EC2), resulting in a significant reduction in energy consumption
S17	VM placement	Reduce energy via proposing: VMPMFuzzyORL (which uses a fuzzy system to create the reward signal). - MRRL (which utilizes the K-means clustering method)	The VMPMFuzzyORL algorithm achieves an improvement in energy usage of between 1 % and 3 % compared to the other benchmark methods, while MRRL achieve an energy consumption improvement between 4 % and 6 %
S18	VM placement	Minimize the number of active physical machines (PMs), and reduces energy consumption and resource fragmentation by Proposing a multi-objective RL approach	The proposed framework achieve up to 17 % improvement in energy consumption compared to other benchmark algorithms
S19	Task offloading/Resource (container) allocation	Develop a two-stage optimizer (ETHC), using DRL agent with Lyapunov optimization to minimize energy consumption at public cloud DC rental costs	Outperform other benchmark methods in maintaining energy consumption below the selected threshold (except for CLOUD, as no tasks are processed in the on-premises DC). Compared to REQ, the proposed ETHC incurs approximately a 6 % additional cost while achieving a 40 % reduction in energy consumption
S21	Task scheduling	Simultaneously optimizes energy consumption and QoS in heterogeneous cloud DC by implementing DAG-based hierarchical DRL strategy	Outperform other benchmark methods, achieving between 2.7 % and 21.6 % less energy consumption, in three different scenarios with 10, 50, and 100 servers
S22	Container placement	Reduce the number of active hosts while efficiently utilizing containers to enhance energy efficiency and curb the environmental impact in DCs	Outperform the other baseline metaheuristic and FFD heuristic algorithms in terms of energy efficiency across all eight configurations of the DC
S26	Task scheduling	Enhance energy efficiency and mitigating carbon emissions in a federated DCs cloud environment (ERLFC) by managing task dependencies in DAGs	Depending on the used dataset size: Achieve better energy efficiency compared to traditional methods (RR and random), ranging from 1 % to 9 %. Additionally, reduce energy consumption by 5 % to 26 % compared to heuristic algorithms (SO and GJO)
S27	DC network traffic control	Optimize DC energy consumption by developing SmartDCN: a PAS-DQN framework that integrates a dynamic traffic predictor (TPM) using LSTM and an intelligent optimizer (EOM) with parametrized DRL	Considering a network size of 250 servers, SmartDCN achieves energy savings of up to 11 %, 8 %, and 4 % compared to the three benchmark methods, respectively. It also demonstrated superior scalability, achieving higher energy savings under the same network load for larger network sizes
S28	VM replacement	Enhance load balancing in two dimensions: between HMs and within each HM using two fuzzy algorithms, Fuzzy-MOVMrB and Fuzzy-RLVMrB, to address non-dominance limitations	While the main goal of the proposed methods is not to directly address energy consumption, the strategy relying on fuzzy logic and RL (Fuzzy-RLVMrB) outperform the other methods in terms of reducing energy consumption
S30	VM placement	Achieve load balancing in HMs while maintaining SLA violations and reducing energy consumption by proposing the VMP-A3C strategy. Additionally, optimize the number of HMs through dynamic consolidation	Outperform the other three DRL algorithms (A2C, DQN, and PPO) by 3.1 %, 1.6 %, and 1.4 %, respectively. It also achieve superior energy efficiency with different VM number scenarios, delivering the best performance with 38.4 % more energy savings compared to the MOVMrB and 7.6 % compared to the ADVMC
S31	Job scheduling	Reduce operational costs in large-scale heterogeneous DCs under continuous job arrivals, considering job dependencies and QoS constraints	Outperform the FIFO algorithm by 37 %, the Tetris algorithm by 30 %, and the ideal MPC algorithm by 1.6 % in terms of energy consumption costs
S32	VM placement	Offer comprehensive solutions for problems related to high energy consumption, SLA violations, VM migrations, and migration duration in geo-distributed cloud DC environments	Achieve better energy efficiency compared to other benchmark algorithms, with improvements of about 5.5 %
S33	VM scheduling	Optimize multiple objectives simultaneously, including energy consumption, performance costs, and SLA violations, for Industrial IoT (IIoT) DC	achieve the lowest energy consumption across five scenarios with varying VM counts: 33.3 % lower to EcoCloud, 15.19 % lower to KMI-MRCU for VM counts over 1000, and 10.79 % lower to AFED-EF for VM counts over 1000

(*continued on next page*)

Table A.9 *(continued)*

ID	Research problem (RQ4)	Objectives (RQ5)	Energy related results (RQ8)
S35	Task scheduling/Resource allocation	Achieve energy efficiency and minimize SLA violations by demonstrating that container-based environments enhance DC energy consumption more effectively than VM-based methods	Outperforms state-of-the-art methods in energy consumption while improving the cloud DC's PUE. The proposed MADRL-RAC approach consumes only 29.637 W for 5000 nodes, compared to 49.346 W (DSTS), 43.926 W (LSTM), 38.625 W (RF), and 35.287 W (CNN)
S36	Task scheduling	Enhances DeepEnergyJSV's performance by deploying the PPO algorithm to schedule hybrid tasks (independent and interdependent) in cloud DCs, improving energy efficiency	Outperform state-of-the-art algorithms in energy efficiency, achieving maximum energy reductions of 7.81 %, 8.93 %, 13.86 %, and 8.69 % compared to EINFORCE, FF, Random, Tetris, respectively
S37	Task scheduling/Resource allocation	Enhance energy efficiency, and scalability in Dc by Proposing an LSTM model for load prediction, followed by a combination of DQN and DPSO	Outperforms all benchmark algorithms in minimizing energy consumption across all cases. For example, in Case 1, the proposed framework reduced energy consumption by 2.63 %, 5.24 %, 41.61 %, and 65.26 % compared to Load Aware, traditional DQN, FFO-EVMM, and MIMT, respectively
S38	VM consolidation	Reduce energy consumption, enhance SLA, and optimize DC resources by developing ARLCA: an autonomous RL agent interacting with the environment	Both proposed Q-learning policies consume more energy than SARSA policies within the same selection strategy. Additionally, softmax policies consumed less energy than ϵ-greedy policies, though the difference was minimal, ranging from 0.02 % and 0.2 %
S39	Job scheduling	Optimize energy consumption while maintaining the required high level of QoS for real-time received jobs in a cloud DC environment	Confirm the superiority of the proposed strategy over baseline methods in energy cost. For example, in scenarios with varying mean arrival rates, baseline methods had similar energy costs, while the proposed algorithm achieved a 60 %–70 % reduction
S40	DC network traffic control	Address the challenge of achieving energy efficiency in DCNs by introducing a DRL framework for solving bandwidth allocation and routing optimization problems, while considering dynamic and real-time flow demands	Outperform heuristic baseline algorithms by up to 7.4 % and an average of 4 % in energy consumption. For small-scale DCN, the Gurobi optimizer provides better energy-efficient solutions. However, when network nodes exceed 70, its convergence time becomes unacceptable, making results unavailable
S42	VM consolidation	Reduce energy consumption and minimize SLA violations by developing a DRL-based VM consolidation (AVMC) strategy that addresses the challenge of continuous dynamic workloads	Confirm the superior performance of the proposed strategy over other methods. AVMC consumed 161.14, 138.31, 153, and 140.78 kWh for the four workloads, achieving energy savings of 6 % to 15 % compared with benchmark algorithms
S43	Resource scheduling	Ensure scalability, adaptability, and efficient resource allocation under varying workloads to reduce computational complexity, optimize task execution, enhance energy efficiency, and improve system performance	Compares across VM numbers ranging from 5 to 20. While energy consumption increases with higher VM counts, the proposed strategy consistently achieves the lowest energy consumption, outperforming state-of-the-art algorithms in energy savings
S44	VM scheduling	Achieve energy efficiency in large-scale cloud DCs using a DRL-based VM scheduling framework, leveraging robust QoS features extracted through SDAE to enhance the scheduling algorithm	Outperform benchmark algorithms in reducing energy consumption across two different scenarios. Using the world cup dataset, the proposed SDAEM-MMQ algorithm saves 4.7 % and 22 % more energy compared to the other benchmark algorithms
S48	Task scheduling	Enhances response time, CPU utilization, and energy efficiency in DCs with QEEC, a two-step RL scheduling framework featuring centralized task dispatching and dynamic local prioritization	Reduces energy consumption compared to benchmark methods and achieves the best energy savings among state-of-the-art Q-learning methods
S49	Resource scheduling	Enhance the QoS levels and improve the energy-saving in complex cloud DC environment by proposing DQN algorithm	Achieve lower energy consumption than traditional benchmarks (Random and RR) across varying task numbers. However, it consumes slightly more energy than MoPSO for less than 200 tasks but outperforms MoPSO when tasks exceed 200
S51	VM placement	Explore the NP-hard optimization problem to balance objectives: minimizing energy consumption and reducing resource wastage in DCs, While also addressing weight selection using the Chebyshev scalarization method	Outperforms multi-objective benchmark algorithms in nearly all scenarios by consuming less energy and reducing resource waste more effectively. It also demonstrates superior performance over single-objective algorithms across various scenarios
S53	Job scheduling/Resource allocation	Achieve multi-objective optimization, including energy efficiency and job delay reduction in large-scale cloud computing environments	Shows that the energy consumption of the proposed scheduling algorithm (HDDL) is nearly equal to the Greedy algorithm among all baseline methods. In global job scheduling and resource allocation optimization, the HDDL-DQN framework outperforms other algorithms in energy efficiency by 5.7 % and 9.7 % compared to MLF and FERPTS, respectively
S54	VM placement	The two main objectives are enhancing Quality of Experience (QoE) and reducing data center energy consumption	In all four scenarios, the proposed algorithm outperforms the FF state-of-the-art algorithm in overall energy consumption. The dot product method surpasses the proposed algorithm in two scenarios but is statistically equivalent to DRL-VMP. Norm2 outperforms the proposed algorithm in only one scenario, where the workload has a constant mean. In this case, simple heuristic methods are recommended
S59	VM consolidation	Optimize energy consumption while maintaining the QoS by proposing a centralized-distributed multi-agent RL algorithm called MAGNETIC	Using three synthetic traces, the results confirm that the proposed algorithm outperforms benchmark algorithms, reducing energy consumption by 58 %, 10 %, and 15 % compared to NPA, PABFD, and E-Eco, respectively, while maintaining QoS
S57	Task scheduling	The objectives include improving energy consumption and reducing waiting time in a large-scale heterogeneous cloud DC	– Small-scale DC (80,000 tasks): The proposed DRL algorithm achieved 23.8 % energy savings, while H2O-Cloud achieved 16.7 % compared to Tetris. – Medium-scale DC (120,000 tasks): The proposed DRL algorithm achieved 27.5 % energy savings, while H2O-Cloud achieved 21.4 % compared to Tetris. – Large-scale DC (240,000 tasks): The proposed algorithm achieved 35.4 %, while H2O-Cloud achieved 24.3 % energy savings compared to Tetris. However, Tetris achieved better QoS, resulting in shorter waiting times across all scenarios. Additionally, as the scale and heterogeneity increase, the proposed algorithm effectively reduces energy consumption while maintaining QoS in task scheduling

(*continued on next page*)

Table A.9 *(continued)*

ID	Research problem (RQ4)	Objectives (RQ5)	Energy related results (RQ8)
S61	DC network traffic control	Dynamically consolidates traffic without prior knowledge to enhance DCN energy efficiency	Shows that algorithms without FCT constraints (ElasticTree and CARPO) performed better than the proposed approach (SmartFCT). However, when considering FCT, the proposed scheme outperforms the benchmark method (FCTcon) by 11.3 %, 11.7 %, and 12.2 % in traffic datasets 1 to 3, respectively. Additionally, it achieves energy savings very close to the optimal algorithm across all datasets
S62	Job scheduling/Resource allocation	Reduce energy consumption, lower operational costs, decrease makespan, and optimize resource allocation	Outperform all benchmark algorithms across datasets by achieving lower energy consumption
S63	Task scheduling	Reduces energy consumption, while balancing throughput, resource utilization, and makespan in heterogeneous cloud computing environments	The energy aspect was evaluated in two scenarios: varying task numbers and VM counts. In both cases, the proposed algorithm outperformed other methods in energy reduction. Simulations showed that as task or VM numbers increased, the algorithm improved load balancing and optimized resource utilization more effectively, minimizing energy consumption
S64	Task scheduling	The objectives include minimizing makespan, reducing energy consumption, lowering operational costs, and maximizing resource utilization	Reduce energy consumption compared to MCP and ETF benchmarks across all datasets by optimizing resource count and frequency

Table A.10
Selected integrated studies objectives and outcomes.

ID	Research problem (RQ4)	Objectives (RQ5)	Energy related results (RQ8)
S1	Workflow Scheduling/RES-powered DC	Reduce brown energy usage and optimize workflow execution across geo-distributed cloud DCs with a multi-agent DRL: a global scheduler assigns workflows to DCs, and a local scheduler allocates tasks to physical machines	Minimize energy consumption by 47 %, while outperforming the baseline algorithms
S2	Load shifting/DC efficient cooling system optimization	Assess the flexibility of the DC cooling system by implementing the SAC-LSTM strategy to optimize costs, shifting cooling demand to low-price periods via RTP while maintaining temperature constraints	Using real-world weather data and electricity price signals over multiple years, the proposed controller reduces energy costs by 2 %–4 % compared to PID and SAC controllers while maintaining DC temperature within the desired range
S5	Task scheduling/DC efficient cooling system optimization	Propose DeepEE framework using a Parametrized Action Space Deep Q-Network (PADQN) to handle high-dimensional state space and discrete-continuous action space issues in ICT and cooling systems, enhancing PUE, preventing rack overheating, and balancing load distribution	Improve power consumption for IT and cooling systems compared to baseline algorithms using PUE metrics, saving 7 % over ICO, 15 % over CCO, 10 % over JCO, and 5 % over O-DQN
S6	Job scheduling/DC efficient cooling system optimization	Propose the E-DRL strategy to make decisions based on critical events rather than time intervals, enhancing PUE, reducing regulating decisions, stabilizing system performance, and improving ICT-cooling system coordination by addressing time constant differences	Using two types of workload traces, the designed algorithm achieved better PUE than benchmark algorithms, including PADQN with time-driven cooling intervals (10, 300, and 900 s) and E-QL
S15	Workflow Scheduling/RES-powered DCs	Develop a novel multi-objective framework, CFWS, based on DRL to balance energy costs and carbon emissions across geo-distributed cloud DCs. It enables workload shifting via VM migration while maximizing renewable energy utilization	Reduce brown energy consumption by 5.67 % to 13.22 % compared to benchmark methods while maintaining the same energy usage. Additionally, it allocates more renewable energy and lowers carbon emissions
S25	VMs Placement/DC efficient cooling system optimization (Thermal awareness)	Handle vast state-action spaces and random delayed feedback in the DC environment with a scalable, hierarchical RL approach, improving energy efficiency, maintaining thermal conditions, and satisfying SLAs	Outperform other benchmark methods by over 17 % in total energy savings
S34	Job scheduling/DC efficient cooling system optimization	Proposes a DRL strategy (MADDQN) with a two-agent structure to optimize job scheduling and cooling. Each agent has an action network (ActNet) for local data and an evaluation network (EvalNet) for global data, enabling centralized training and decentralized execution. This ensures ideal room temperature, meets workload deadlines, and reduces energy consumption	Outperform other DQN algorithms in energy efficiency, achieving the best PUE, lowest total energy consumption, reduced hot spots, and improved scalability for larger DC configurations
S45	Task scheduling/DC efficient cooling system optimization	Proposes a DRL-based framework to optimize IT and facility operations in DCs. The DRL algorithm interacts with the physical DC system by continuously collecting real-time states and applying control actions across three areas: – Thermal-Aware Task Scheduling: DDPG optimizes resource allocation with thermal considerations. – Load-Aware Target Cooling: DQN manages CRAC airflow based on IT workload. – IT-Facility Optimization: PADQN coordinates IT and facility operations for global efficiency.	Cooling Energy Savings: – Up to 15 % reduction in cooling energy for air-cooled systems. – Up to 30 % reduction for water-cooled systems Overall Energy Efficiency: – Joint IT and facility optimization achieved up to 15 % total energy savings compared to baseline manual control Task Scheduling: – Thermal-aware task scheduling reduced IT power consumption by 9 %

(*continued on next page*)

Table A.10 *(continued)*

ID	Research problem (RQ4)	Objectives (RQ5)	Energy related results (RQ8)
S46	Task scheduling/DC efficient cooling system optimization	Propose a multi-agent framework (MAC3C) to jointly optimize IT infrastructure and cooling operations, enhancing DC energy efficiency. The framework interacts with the real-time DC environment instead of sub-system models to observe dynamic state space and generate corresponding discrete-continuous actions	Compared to traditional methods (Random and RR), the proposed framework achieves significant energy savings. Additionally, MAC3C consumes 42.82 % and 18.95 % less energy than the joint optimization approaches DeepEE and PowerTrade, respectively
S50	Workflow Scheduling/RES-powered DCs	Propose a green workload framework that simultaneously optimizes two objectives: maximizing fuel cell utilization benefits and minimizing power budget fragmentation	Confirm the proposed algorithm's effectiveness in reducing energy consumption when DCs exceed two, demonstrating its efficiency in high-dimensional upper-level environments. Additionally, in low-level DC environments with varying rack numbers, the algorithm consumes less energy and minimizes power budget fragmentation. It achieves energy savings of up to 7.5 %, 5.2 %, and 4.3 % compared to benchmark methods
S52	Workflow Scheduling/DC efficient cooling system optimization/Battery charge and discharge	Address the challenge of dynamically optimizing DC operations over a year with a DRL framework. It combines D3QN and VDN in a multi-agent system to train three DRL agents that optimize battery charging/discharging, computational workload distribution, and waste heat utilization from the cooling system, achieving global integration and efficiency	Confirm that the proposed framework improves energy efficiency compared to pre-optimization, achieving: 18.37 % reduction in renewable energy waste, 9.78 % improvement in operational cost efficiency, 4.01 % reduction in electricity consumption, and 29.74 % decrease in grid electricity consumption. Additionally, results show that algorithmic optimization outperforms non-algorithmic methods in reducing renewable energy waste and operational costs

Table A.11
Overview of the main MDP elements in each joint optimization selected work.

ID	RL elements
S1	**state-space** $\{S\}$**:** For global agent: green energy surplus or deficit levels of the i-th datacenter, average processing speed of a server in the i-th datacenter, the current utilization level of the i-th datacenter, CPU requirement of j-th task, and Memory requirement of j-th task. For Local Scheduler: the processing speed of λi-th server in datacenter dci, the current utilization level of the λi-th server in datacenter dci, CPU requirement of j-th task, and Memory requirement of j-th task **Action-Space** $\{A\}$**:** For global agent: the selected data center (the local scheduler) where the task will be executed. For Local Scheduler: the selection of a server, mi in which the task will be executed **Reward-Function** $\{R\}$**:** For global agent: the current green energy deficit or surplus of the selected data center, and the local scheduler that utilize the physical machine to execute the task. For Local Scheduler: Energy efficiency and execution time **Constraints:** NA **Discount factor:** Unspecified
S2	**state-space** $\{S\}$: Outdoor (drybulb) air temperature, the indoor air temperatures of both data center zones, the IT equipment power demand, the HVAC equipment power demand, the real-time electricity price, and the hourly electricity cost **Action-Space** $\{A\}$: The common setpoint of the outlet air temperature of the evaporative coolers and cooling coils, as well as the mass flow rates of the fans in the air handler (i.e., fan mass flow rate) for both zones of the data center **Reward-Function** $\{R\}$: The temperatures in a recommended range in each data center zone and electricity cost of operating the data center through RTP **Constraints:** NA **Discount factor:** 0.99 for 1 day, 0.997 for 2 weeks
S5	**state-space** $\{S\}$: The number of required CPU cores of the candidate task i, the airflow rates, the IT workload states, and the thermal states **Action-Space** $\{A\}$**:** The action for task scheduling is to select a proper server. The action for regulating cooling facilities is to adjust the airflow rate **Reward-Function** $\{R\}$**:** The PUE of the data center. Penalties for overheating of each rack n and overloading of each server k **Constraints:** The number of available CPUs must be greater than the required number of CPUs to perform the dispatcher task. The airflow rate for each J ACUs must be greater than 0 and less than or equal to the maximum airflow rate that the ACUs can supply **Discount factor:** 0.99
S6	**state-space** $\{S\}$: Instead of using state space, this study used event space, including: Job Dispatching Events: Activated when the length of the job queue is greater than zero. State Transition Events: Represent the common state transitions of the following elements: CPU cores, server utilization, power consumption, two inlet temperatures, and the outlet temperature. Change Ratio Events: Aim to sense the change ratio of utilization and outlet temperature for each rack. Boundary Events: Handle critical scenarios, including hotspot events and over-provisioning events **Action-Space** $\{A\}$**:** The action for task scheduling is to select a proper server. The action for regulating cooling facilities is to adjust the airflow rate **Reward-Function** $\{R\}$: The PUE of the data center. Penalty for overheating of each rack n. Penalty for overloading of each server k **Constraints:** The number of available CPUs must be greater than the required number of CPUs to perform the dispatcher task. The airflow rate for each J ACUs must be greater than 0 and less than or equal to the maximum airflow rate that the ACUs can supply **Discount factor:** 0.99
S15	**state-space** $\{S\}$: The CPU utilization of the j-th host of DC k **Action-Space** $\{A\}$**:** Selecting VM migration schemes (migrated VM, destination DC, and PM) to balance overloaded or underloaded PMs **Reward-Function** $\{R\}$**:** The energy cost. The carbon footprint **Constraints:** The CPU capacity **Discount factor:** Unspecified
S25	**state-space** $\{S\}$: A tree structure to distribute VM requests among the servers. The inner nodes of the tree serve double roles: VM requests distributor and state of the Markov model **Action-Space** $\{A\}$**:** Picking a route from a particular node **Reward-Function** $\{R\}$**:** Thermal : server inlet temperature and CPU temperature. Power: power consumption relative to utilization rate **Constraints:** Maximum VM numbers statement: A constraint can be set on the maximum number of VMs that can be placed on each PM to prevent overloading. While putting other PMs into a sleep state **Discount factor:** 0.99

(continued on next page)

Table A.11 *(continued)*

ID	RL elements
S34	**state-space** {S}: For ICT system: The current server resource state and job request resources in the queue. For the cooling system: server inlet air temperature, CRAC return air temperature, and CRAC set temperature **Action-Space** {A}**:** For the ICT system: assigning job j in the queue to a server. For cooling system: the temperature regulation action **Reward-Function** {R}**:** For ICT system: minimize the energy consumption of the server on the premise of avoiding hot spots. For cooling system: minimize the cooling energy consumption while ensuring the safe operation of the server **Constraints:** NA **Discount factor:** Unspecified
S45	**state-space** {S}: For load-aware target cooling: power consumption and workload of the ICT subsystem, along with cooling subsystem power consumption and ambient temperature. For thermal-aware task scheduling: power consumption of the ICT and cooling systems, and IT subsystem temperatures. For iterative IT-cooling optimization: observations from the discrete-space ICT subsystem and continuous-space cooling subsystem **Action-Space** {A}**:** For load-aware target cooling case: Airflow rate adjustment, and Pump flow rate adjustment. For Thermal-aware task scheduling: Assigning a task to a specific server in a thermal-aware manner. For Iterative IT-cooling optimization: two kinds of control actions simultaneously **Reward-Function** {R}**:** For load-aware target cooling: optimize the trade-off between IT workload, ambient temperature, and facility energy cost. For thermal-aware task scheduling: minimize IT and facility power consumption while maintaining server temperature and computing throughput. For iterative IT-cooling optimization: jointly control IT and facility subsystems to balance energy consumption and efficiency **Constraints:** NA **Discount factor:** Unspecified
S46	**state-space** {S}: The available resources of each server. The power consumption of each server. The adaptability scores of each server to the current task. The outlet air temperature of each rack. The power consumption of each rack. The supply air temperature and flow rate. The requested resources of the current task **Action-Space** {A}**:** For ICT system: scheduling task to server. For cooling system: supply air temperature, and the flow rate **Reward-Function** {R}**:** The direct power influence of actions, measured by the change in IT and cooling power before and after execution. The waiting time of the current task. The available resource that the target server holds. The outlet air temperature of servers **Constraints:** The supply air temperature range and flow rate of CRACs should respect upper and lower thresholds. Optimization should consider both factors alongside task scheduling to maximize data center power savings **Discount factor:** 0.99
S50	**state-space** {S}: Global Scheduler: Green energy surplus/deficit levels of the i-th datacenter. –Average server processing speed in the i-th datacenter. –Current utilization level of the i-th datacenter. –CPU and memory requirements of the j-th task. –Memory requirement of j-th task. Local Scheduler: Processing speed of the λi-th server in datacenter. –Current utilization level of the λi-th server in datacenter. –CPU requirement of j-th task. –Memory requirement of j-th task **Action-Space** {A}: Global Scheduler: Action corresponds to the selection of a datacenter (hence a local scheduler) to which the task will be submitted for execution. Local Scheduler: The selection of a server in which the task will be executed **Reward-Function** {R}: Global Scheduler: Maximize green energy utilization by balancing surplus/deficit in the selected datacenter. –The reward from the local scheduler for successful task allocation. Local Scheduler: Weighted function of task execution time. –The corresponding energy consumption during the execution of task **Constraints:** NA **Discount factor:** Unspecified
S52	**state-space** {S}: For the ICT system (computational workload): ambient temperature, municipal electricity price, average occupancy rate, battery capacity, wind power, and photovoltaic generation. For the cooling system (heating temperature of surrounding buildings): ambient temperature, municipal electricity price, battery capacity, wind power, and photovoltaic generation. For the power system (battery charge/discharge): ambient temperature, municipal electricity price, battery capacity, wind power, photovoltaic generation, and total electricity consumption of ICT and cooling **Action-Space** {A}**:** - Adjustment parameter of computational workload scheduling. - Adjustment parameter of heating temperature of the surrounding buildings. - Adjustment parameter of battery charge and discharge, i.e., the change of state of charge (SOC) **Reward-Function** {R}**:** The global reward (reducing renewable energy waste and operational cost) is used as a common reward for evaluating the optimization situation. For ICT and cooling systems: Total electricity consumption, and grid electricity consumption **Constraints:** NA **Discount factor:** Unspecified

Data availability

Data will be made available on request.

References

[1] International Energy Agency. Analysis and forecast to 2026. IEA Report; 2024. https://www.iea.org/reports/electricity-2024.

[2] Kamiya G, Bertoldi P, et al. Energy consumption in data centres and broadband communication networks in the EU. European Commission, Joint Research Centre; 2024.

[3] Andrae AS, Edler T. On global electricity usage of communication technology: trends to 2030. Challenges 2015;6(1):117–57.

[4] Zhang Y, Tang H, Li H, Wang S. Unlocking the flexibilities of data centers for smart grid services: optimal dispatch and design of energy storage systems under progressive loading. Energy 2025;316:134511.

[5] Jayanetti A, Halgamuge S, Buyya R. Deep reinforcement learning for energy and time optimized scheduling of precedence-constrained tasks in edge-cloud computing environments. Fut Gener Comput Syst 2022 Dec; 137:14–30.

[6] Iyengar M, Schmidt R, Caricari J. Reducing energy usage in data centers through control of room air conditioning units. In: 2010 12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems. IEEE; 2010. p. 1–11.

[7] Kumar R, Khatri SK, Diván MJ. Data center air handling unit fan speed optimization using machine learning techniques. In: 2021 9th International conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO). IEEE; 2021. p. 1–10.

[8] Marcinichen JB, Olivier JA, Thome JR. On-chip two-phase cooling of datacenters: cooling system and energy recovery evaluation. Appl Therm Eng 2012; 41:36–51.

[9] Wang H, Yuan X, Zhang K, Lang X, Chen H, Yu H, et al. Performance evaluation and optimization of data center servers using single-phase immersion cooling. Int J Heat Mass Transfer 2024;221:125057.

[10] Gao T, Sammakia BG, Geer J, Murray B, Tipton R, Schmidt R. Comparative analysis of different in row cooler management configurations in a hybrid cooling data center. In: International electronic packaging technical conference and exhibition; vol. 56888. American Society of Mechanical Engineers; 2015. p. V001T09A011.

[11] Shalom Simon V, Modi H, Sivaraju KB, Bansode P, Saini S, Shahi P, et al. Feasibility study of rear door heat exchanger for a high capacity data center. In: International electronic packaging technical conference and exhibition; vol. 86557. American Society of Mechanical Engineers; 2022. p. V001T01A018.

[12] Deymi-Dashtebayaz M, Namanlo SV, Arabkoohsar A. Simultaneous use of air-side and water-side economizers with the air source heat pump in a data center for cooling and heating production. Appl Therm Eng 2019;161:114133.

[13] Jang Y, Lee D, Kim J, Ham SH, Kim Y. Performance characteristics of a waste-heat recovery water-source heat pump system designed for data centers and residential area in the heating dominated region. J Build Eng 2022;62:105416.

[14] Oró E, Depoorter V, Pflugradt N, Salom J. Overview of direct air free cooling and thermal energy storage potential energy savings in data centres. Appl Therm Eng 2015;85:100–10.

[15] Bousnina D, Guerassimoff G. Deep reinforcement learning for optimal energy management of multi-energy smart grids. In: Nicosia G, Ojha V, La Malfa E, La Malfa G, Jansen G, Pardalos PM, Giuffrida G, Umeton R, editors. Machine learning, optimization, and data science. Cham: Springer International Publishing; 2022. p. 15–30.

[16] Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge: MIT Press; 1998.

[17] Chang Q, Huang Y, Liu K, Xu X, Zhao Y, Pan S. Optimization control strategies and evaluation metrics of cooling systems in data centers: a review. Sustainability 2024;16(16):7222.

[18] Shaqour A, Hagishima A. Systematic review on deep reinforcement learning-based energy management for different building types. Energies 2022;15(22):8663.
[19] Garí Y, Monge DA, Pacini E, Mateos C, Garino CG. Reinforcement learning-based application autoscaling in the cloud: a survey. Eng Appl Artif Intel 2021;102:104288.
[20] Magotra B, Malhotra D, Dogra AK. Adaptive computational solutions to energy efficiency in cloud computing environment using VM consolidation. Arch Comput Methods Eng 2023;30(3):1789–818.
[21] Zhou G, Tian W, Buyya R, Xue R, Song L. Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions. Artif Intell Rev 2024;57(5):124.
[22] Hou H, Jawaddi SNA, Ismail A. Energy efficient task scheduling based on deep reinforcement learning in cloud environment: a specialized review. Fut Gener Comput Syst 2024;151:214–31.
[23] Singh S, Kumar R, Singh D. An empirical investigation of task scheduling and VM consolidation schemes in cloud environment. Comput Sci Rev 2023;50:100583.
[24] Lin W, Lin J, Peng Z, Huang H, Lin W, Li K. A systematic review of green-aware management techniques for sustainable data center. Sustain Comput Inf Syst 2024;100989.
[25] Long S, Li Y, Huang J, Li Z, Li Y. A review of energy efficiency evaluation technologies in cloud data centers. Energy Build 2022;260:111848.
[26] Zhang W, Wen Y, Wong YW, Toh KC, Chen C-H. Towards joint optimization over ICT and cooling systems in data centre: a survey. IEEE Commun Surv Tutor 2016;18(3):1596–616.
[27] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015 Feb; 518:529–33.
[28] Frank LL, Vrabie D, VamVoudakis KG. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. IEEE Control Syst 2012 Dec; 32:76–105.
[29] Busoniu L, Babuska R, De Schutter B, Ernst D. Reinforcement learning and dynamic programming using function approximators. CRC Press; 2017 July.
[30] Zanini E. Markov decision processes. Citeseer; 2014.
[31] Ladosz P, Weng L, Kim M, Oh H. Exploration in deep reinforcement learning: a survey. Info Fusion 2022 Sep; 85:1–22.
[32] Bellman R. Dynamic programming. Science 1966;153(3731):34–7.
[33] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res 1996 May; 4:237–85.
[34] Watkins CJ, Dayan P. Q-learning. Mach Learn 1992;8:279–92.
[35] Rummery GA, Niranjan M. On-line Q-Learning using connectionist systems, vol. 37. Cambridge, UK: University of Cambridge, Department of Engineering; 1994.
[36] Sutton RS. Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bull 1991;2(4):160–3.
[37] Winands MH, Björnsson Y, Saito J-T. Monte-Carlo tree search solver. In: Computers and games: 6th international conference, CG 2008, Beijing, China, September 29–October 1, 2008. Proceedings 6; Springer; 2008. p. 25–36.
[38] Wang X, Wang S, Liang X, Zhao D, Huang J, Xu X, et al. Deep reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst 2024 Apr; 35:5064–78.
[39] Li Y. Deep reinforcement learning: an overview. arXiv:1701.07274; 2018 Nov.
[40] Shao K, Tang Z, Zhu Y, Li N, Zhao D. A survey of deep reinforcement learning in video games. arXiv:1912.10944; 2019 Dec.
[41] Parvez Farazi N, Zou B, Ahamed T, Barua L. Deep reinforcement learning in transportation research: a review. Transp Res Interdiscip Perspect 2021 Sep;11:100425.
[42] Cao D, Hu W, Zhao J, Zhang G, Zhang B, Liu Z, et al. Reinforcement learning and its applications in modern power and energy systems: a review. J Mod Power Syst Clean Energy 2020 Nov;8:1029–42.
[43] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th international conference on machine learning. PMLR; 2018 July. p. 1861–70. ISSN: 2640-3498.
[44] Kurte K, Munk J, Kotevska O, Amasyali K, Smith R, McKee E, et al. Evaluating the adaptability of reinforcement learning based HVAC control for residential houses. Sustainability 2020 Sep;12:7727.
[45] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151(4):W-65.
[46] Munn Z, Tufanaru C, Aromataris E. Jbi's systematic reviews: data extraction and synthesis. AJN Am J Nurs 2014;114(7):49–54.
[47] Jayanetti A, Halgamuge S, Buyya R. Multi-agent deep reinforcement learning framework for renewable energy-aware workflow scheduling on distributed cloud data centers. IEEE Trans Parallel Distrib Syst 2024;35(4):604–15.
[48] Biemann M, Gunkel PA, Scheller F, Huang L, Liu X. Data center HVAC control harnessing flexibility potential via real-time pricing cost optimization using reinforcement learning. IEEE Internet Things J 2023;10(15):13876–94.
[49] Wan J, Duan Y, Gui X, Liu C, Li L, Ma Z. SafeCool: safe and energy-efficient cooling management in data centers with model-based reinforcement learning. IEEE Trans Emerg Top Comput Intel 2023;7(6):1621–35.
[50] Lou J, Tang Z, Jia W. Energy-efficient joint task assignment and migration in data centers: a deep reinforcement learning approach. IEEE Trans Netw Serv Manage 2023;20(2):961–73.
[51] Ran Y, Hu H, Wen Y, Zhou X. Optimizing energy efficiency for data center via parameterized deep reinforcement learning. IEEE Trans Serv Comput 2023;16(2):1310–23.
[52] Ran Y, Zhou X, Hu H, Wen Y. Optimizing data center energy efficiency via event-driven deep reinforcement learning. IEEE Trans Serv Comput 2023;16(2):1296–309.
[53] Zeng J, Ding D, Kang XK, Xie H, Yin Q. Adaptive DRL-based virtual machine consolidation in energy-efficient cloud data center. IEEE Trans Parall Distrib Syst 2022;33(11):2991–3002.
[54] Kang K, Ding D, Xie H, Yin Q, Zeng J. Adaptive DRL-based task scheduling for energy-efficient cloud computing. IEEE Trans Netw Serv Manage 2022;19(4):4948–61.
[55] Pham T-M. Traffic engineering based on reinforcement learning for service function chaining with delay guarantee. IEEE Access 2021;9:121583–92.
[56] Yi D, Zhou X, Wen Y, Tan R. Efficient compute-intensive job allocation in data centers via deep reinforcement learning. IEEE Trans Parall Distrib Syst 2020;31(6):1474–85.
[57] Ding W, Luo F, Gu C, Lu H, Zhou Q. Performance-to-power ratio aware resource consolidation framework based on reinforcement learning in cloud data centers. IEEE Access 2020;8:15472–83.
[58] Li Y, Wen Y, Tao D, Guan K. Transforming cooling optimization for green data center via deep reinforcement learning. IEEE Trans Cybern 2020;50(5):2002–13.
[59] Cheng M, Li J, Bogdan P, Nazarian S. H2O-Cloud: a resource and quality of service-aware task scheduling framework for warehouse-scale data centers. IEEE Trans Comput Aided Des Integr Circuits Syst 2020;39(10):2925–37.
[60] Leindals L, Grønning P, Dominković DF, Junker RG. Context-aware reinforcement learning for cooling operation of data centers with an aquifer thermal energy storage. Energy AI 2024;17:100395.
[61] Zhao D, Zhou J-T, Li K. CFWS: DRL-based framework for energy cost and carbon footprint optimization in cloud data centers. IEEE Trans Sustain Comput; 2025;10(1):95–107.
[62] Ghasemi A, Keshavarzi A. Energy-efficient virtual machine placement in heterogeneous cloud data centers: a clustering-enhanced multi-objective, multi-reward reinforcement learning approach. Clust Comput 2024;27(10):14149–66.
[63] Ghasemi A, Toroghi Haghighat A, Keshavarzi A. Enhancing virtual machine placement efficiency in cloud data centers: a hybrid approach using multi-objective reinforcement learning and clustering strategies. Computing 2024;106(9): 2897–922.
[64] Bhatt C, Singhal S. Multi-objective reinforcement learning for virtual machines placement in cloud computing. Int J Adv Comput Sci Appl 2024;15(3):1051—8.
[65] Zhang J, Yu H, Fan G, Li Z. Elastic task offloading and resource allocation over hybrid cloud: a reinforcement learning approach. IEEE Trans Netw Serv Manage 2024;21(2):1983–97.
[66] Guo Y, Qu S, Wang C, Xing Z, Duan K. Optimal dynamic thermal management for data center via soft actor-critic algorithm with dynamic control interval and combined-value state space. Appl Energy 2024;373:123815.
[67] Yang W, Zhao M, Li J, Zhang X. Energy-efficient DAG scheduling with DVFS for cloud data centers. J Supercomput 2024;80(10):14799–823.
[68] Bouaouda A, Afdel K, Abounacer R. Unveiling genetic reinforcement learning (GRLA) and hybrid attention-enhanced gated recurrent unit with random forest (HAGRU-RF) for energy-efficient containerized data centers empowered by solar energy and AI. Sustainability 2024;16(11):4438.
[69] Chen Y, Guo W, Liu J, Shen S, Lin J, Cui D. A multi-setpoint cooling control approach for air-cooled data centers using the deep Q-network algorithm. Meas Control 2024;57(6):782–93.
[70] Wang R, Cao Z, Zhou X, Wen Y, Tan R. Green data center cooling control via physics-guided safe reinforcement learning. ACM Trans Cyber-Phys Syst 2024;8(2):1–26.
[71] Aghasi A, Jamshidi K, Bohlooli A, Javadi B. A decentralized adaptation of model-free Q-learning for thermal-aware energy-efficient virtual machine placement in cloud data centers. Comput Netw 2023;224:109624.
[72] Wang Z, Chen S, Bai L, Gao J, Tao J, Bond RR, et al. Reinforcement learning based task scheduling for environmentally sustainable federated cloud computing. J Cloud Comp 2023;12(1):174.
[73] Wang T, Fan X, Cheng K, Du X, Cai H, Wang Y. Parameterized deep reinforcement learning with hybrid action space for energy efficient data center networks. Comput Netw 2023;235:109989.
[74] Ghasemi A, Toroghi Haghighat A, Keshavarzi A. Enhanced multi-objective virtual machine replacement in cloud data centers: combinations of fuzzy logic with reinforcement learning and biogeography-based optimization algorithms. Clust Comput 2023;26(6):3855–68.
[75] Huang N, Li X, Xu Q, Chen R, Chen H, Chen A. Artificial intelligence-based temperature twinning and pre-control for data center airflow organization. Energies 2023;16(16):6063.
[76] Wei P, Zeng Y, Yan B, Zhou J, Nikougoftar E. VMP-A3C: virtual machines placement in cloud computing based on asynchronous advantage actor-critic algorithm. J King Saud Univ Comput Inf Sci 2023;35(5):101549.
[77] Liu W, Yan Y, Sun Y, Mao H, Cheng M, Wang P, et al. Online job scheduling scheme for low-carbon data center operation: an information and energy nexus perspective. Appl Energy 2023;338:120918.
[78] Ahamed Z, Khemakhem M, Eassa F, Alsolami F, Basuhail A, Jambi K. Deep reinforcement learning for workload prediction in federated cloud environments. Sensors 2023;23(15):6911.
[79] Ma X, Xu H, Gao H, Bian M, Hussain W. Real-time virtual machine scheduling in industry IoT network: a reinforcement learning method. IEEE Trans Ind Inf 2023;19(2):2129–39.
[80] Simin W, Lulu Q, Chunmiao M, Weiguo W. Research on overall energy consumption optimization method for data center based on deep reinforcement learning. J Intell Fuzzy Syst 2023;44(5):7333–49.
[81] Nagarajan S, Rani PS, Vinmathi MS, Subba Reddy V, Saleth ALM, Abdus Subhahan D. Multi agent deep reinforcement learning for resource allocation in container-based clouds environments. Expert Syst 2025;42(1):e13362.

[82] Yang Y, He C, Yin B, Wei Z, Hong B. Cloud task scheduling based on proximal policy optimization algorithm for lowering energy consumption of data center. KSII Trans Internet Inf Syst 2022;16(6):1877–91.
[83] Pandey NK, Diwakar M, Shankar A, Singh P, Khosravi MR, Kumar V. Energy efficiency strategy for big data in cloud environment using deep reinforcement learning. Mob Inf Syst 2022;2022:1–11.
[84] Shaw R, Howley E, Barrett E. Applying reinforcement learning towards automating energy efficient virtual machine consolidation in cloud data centers. Inf Syst 2022;107:101722.
[85] Yan J, Huang Y, Gupta A, Gupta A, Liu C, Li J, et al. Energy-aware systems for real-time job scheduling in cloud data centers: a deep reinforcement learning approach. Comp Electr Eng 2022;99:107688.
[86] Wang Y, Li Y, Wang T, Liu G. Towards an energy-efficient data center network based on deep reinforcement learning. Comput Netw 2022;210:108939.
[87] Mahbod MHB, Chng CB, Lee PS, Chui CK. Energy saving evaluation of an energy efficient data center using a model-free reinforcement learning approach. Appl Energy 2022;322:119392.
[88] Abbas K, Hong J, Tu NV, Yoo J-H, Hong JW-K. Autonomous DRL-based energy efficient VM consolidation for cloud data centers. Phys Commun 2022;55:101925.
[89] Uma J, Vivekanandan P, Shankar S. Optimized intellectual resource scheduling using deep reinforcement Q-learning in cloud computing. Trans Emerg Tel Tech 2022;33(5):e4463.
[90] Wang B, Liu F, Lin W. Energy-efficient VM scheduling based on deep reinforcement learning. Fut Gener Comput Syst 2021;125:616–28.
[91] Zhou X, Wang R, Wen Y, Tan R. Joint IT-facility optimization for green data centers via deep reinforcement learning. IEEE Netw 2021;35(6):255–62.
[92] Chi C, Ji K, Song P, Marahatta A, Zhang S, Zhang F, et al. Cooperatively improving data center energy efficiency based on multi-agent deep reinforcement learning. Energies 2021;14(8):2071.
[93] Biemann M, Scheller F, Liu X, Huang L. Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. Appl Energy 2021;298:117164.
[94] Ding D, Fan X, Zhao Y, Kang K, Yin Q, Zeng J. Q-learning based dynamic task scheduling for energy-efficient cloud computing. Fut Gener Comput Syst 2020;108:361–71.
[95] Peng Z, Lin J, Cui D, Li Q, He J. A multi-objective trade-off framework for cloud resource scheduling based on the deep Q-network algorithm. Clust Comput 2020;23(4):2753–67.
[96] Hu X, Sun Y. A deep reinforcement learning-based power resource management for fuel cell powered data centers. Electronics 2020;9(12):2054.
[97] Qin Y, Wang H, Yi S, Li X, Zhai L. Virtual machine placement based on multi-objective reinforcement learning. Appl Intell 2020;50(8):2370–83.
[98] Yang D, Wang X, Shen R, Li Y, Gu L, Zheng R, et al. Global optimization strategy of prosumer data center system operation based on multi-agent deep reinforcement learning. J Build Eng 2024;91:109519.
[99] Lin J, Cui D, Peng Z, Li Q, He J. A two-stage framework for the multi-user multi-data center job scheduling and resource allocation. IEEE Access 2020;8:197863–74.
[100] Caviglione L, Gaggero M, Paolucci M, Ronco R. Deep reinforcement learning for multi-objective placement of virtual machines in cloud datacenters. Soft Comput 2021;25(19):12569–88.
[101] Le DV, Wang R, Liu Y, Tan R, Wong Y-W, Wen Y. Deep reinforcement learning for tropical air free-cooled data center control. ACM Trans Sen Netw 2021;17(3):1–28.
[102] Zhang Q, Zeng W, Lin Q, Chng C-B, Chui C-K, Lee P-S. Deep reinforcement learning towards real-world dynamic thermal management of data centers. Appl Energy 2023;333:120561.
[103] Li J, Zhang X, Wei Z, Wei J, Ji Z. Energy-aware task scheduling optimization with deep reinforcement learning for large-scale heterogeneous systems. CCF Trans HPC 2021;3(4):383–92.
[104] Wan J, Zhou J, Gui X. Intelligent rack-level cooling management in data centers with active ventilation tiles: a deep reinforcement learning approach. IEEE Intell Syst 2021;36(6):42–52.
[105] Haghshenas K, Pahlevan A, Zapater M, Mohammadi S, Atienza D. MAGNETIC: multi-agent machine learning-based approach for energy efficient dynamic consolidation in data centers. IEEE Trans Serv Comput 2022;15(1):30–44.
[106] Zhang Q, Mahbod MHB, Chng C-B, Lee P-S, Chui C-K. Residual physics and post-posed shielding for safe deep reinforcement learning method. IEEE Trans Cybern 2024;54(2):865–76.
[107] Sun P, Guo Z, Liu S, Lan J, Wang J, Hu Y. SmartFCT: improving power-efficiency for data center networks with deep reinforcement learning. Comput Netw 2020;179:107255.
[108] Asghari A, Sohrabi MK, Yaghmaee F. A cloud resource management framework for multiple online scientific workflows using cooperative reinforcement learning agents. Comput Netw 2020;179:107340.
[109] Siddesha K, Jayaramaiah GV, Singh C. A novel deep reinforcement learning scheme for task scheduling in cloud computing. Clust Comput 2022;25(6):4171–88.
[110] Asghari A, Sohrabi MK, Yaghmaee F. Online scheduling of dependent tasks of cloud's workflows to enhance resource utilization and reduce the makespan using multiple reinforcement learning-based agents. Soft Comput 2020;24(21):16177–99.
[111] Zhang Q, Chng C-B, Chen K, Lee P-S, Chui C-K. DRL-s: toward safe real-world learning of dynamic thermal management in data center. Expert Syst Appl 2023;214:119146.
[112] Shao X, Zhang Z, Song P, Feng Y, Wang X. A review of energy efficiency evaluation metrics for data centers. Energy Build 2022;271:112308.
[113] Jin C, Bai X, Yang C, Mao W, Xu X. A review of power consumption models of servers in data centers. Appl Energy 2020;265:114806.
[114] Moriyama T, Magistris GD, Tatsubori M, Pham T-H, Munawar A, Tachibana R. Reinforcement learning testbed for power-consumption optimization. In: Proceedings of Asia Simulation conference (AsiaSim), Kyoto, Japan; 2018. p. 45–59.
[115] Phan L, Lin C-X. A multi-zone building energy simulation of a data center model with hot and cold aisles. Energy Build 2014;77:364–76.
[116] Sun K, Luo N, Luo X, Hong T. Prototype energy models for data centers. Energy Build 2021;231:110603.
[117] U.S. Department of Energy. EnergyPlus: energy simulation software; 2024. https://energyplus.net/ [Accessed 18.11.2024].
[118] OpenFOAM Foundation. OpenFOAM: the open source CFD toolbox; 2024. https://www.openfoam.com/ [Accessed 18.11.2024].
[119] Cadence Design Systems. Cadence reality digital twin platform; 2024. https://www.cadence.com/en_US/home/tools/reality-digital-twin.html [Accessed 19.11.2024].
[120] Van Geet O, Sickinger D. Best practices guide for energy-efficient data center design. Technical Report. Golden, CO (United States): National Renewable Energy Laboratory (NREL); 2024.
[121] Sharma P, Chaufournier L, Shenoy P, Tay Y. Containers and virtual machines at scale: a comparative study. In: Proceedings of the 17th international middleware conference; 2016. p. 1–13.
[122] Calheiros RN, Ranjan R, Beloglazov A, De Rose CA, Buyya R. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw Pract Exp 2011;41(1):23–50.
[123] Chen W, Deelman E. Workflowsim: a toolkit for simulating scientific workflows in distributed environments. In: 2012 IEEE 8th international conference on E-science. IEEE; 2012. p. 1–8.
[124] He H, Meng X, Wang Y, Khajepour A, An X, Wang R, et al. Deep reinforcement learning based energy management strategies for electrified vehicles: recent advances and perspectives. Renew Sustain Energy Rev 2024;192:114248.
[125] Metrics GG. Describing datacenter power efficiency. Technical Committee White Paper. The Green Grid; 2007.
[126] Horner N, Azevedo I. Power usage effectiveness in data centers: overloaded and underachieving. Electr J 2016;29(4):61–9.
[127] Patterson M, Tschudi B, Vangeet O, Cooley J, Azevedo D. ERE: a metric for measuring the benefit of reuse energy from a data center. White Paper 29; 2010.
[128] Sego LH, Marquez A, Rawson A, Cader T, Fox K, Gustafson WI Jr, et al. Implementing the data center energy productivity metric. ACM J Emerg Technol Comput Syst 2012;8(4):1–22.
[129] Green I. New data center energy efficiency evaluation index dppe (datacenter performance per energy). Measurement guidelines (ver 2.05). 2012 Mar.
[130] Reddy VD, Setz B, Rao GSV, Gangadharan G, Aiello M. Metrics for sustainable data centers. IEEE Trans Sustain Comput 2017;2(3):290–303.
[131] Pham HX, La HM, Feil-Seifer D, Nefian A. Cooperative and distributed reinforcement learning of drones for field coverage. arXiv preprint arXiv:1803.07250; 2018.
[132] Qie H, Shi D, Shen T, Xu X, Li Y, Wang L. Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning. IEEE Access 2019;7:146264–72.
[133] Biagioni D, Zhang X, Wald D, Vaidhynathan D, Chintala R, King J, et al. PowerGridworld: a framework for multi-agent reinforcement learning in power systems. In: Proceedings of the thirteenth ACM international conference on future energy systems; 2022. p. 565–70.
[134] Wang J, Xu W, Gu Y, Song W, Green TC. Multi-agent reinforcement learning for active voltage control on power distribution networks. Adv Neural Inf Process Syst 2021;34:3271–84.
[135] Terry J, Black B, Grammel N, Jayakumar M, Hari A, Sullivan R, et al. PettingZoo: gym for multi-agent reinforcement learning. Adv Neural Inf Process Syst 2021;34:15032–43.
[136] Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv preprint arXiv:2011.00583; 2020.
[137] Oroojlooy A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning. Appl Intell 2023;53(11):13677–722.
[138] Canese L, Cardarilli GC, Di Nunzio L, Fazzolari R, Giardino D, Re M, et al. Multi-agent reinforcement learning: a review of challenges and applications. Appl Sci 2021;11(11):4948.
[139] Ibrahim AM, Yau K-LA, Chong Y-W, Wu C. Applications of multi-agent deep reinforcement learning: models and algorithms. Appl Sci 2021;11(22):10870.
[140] Wang F, Wang X, Sun S. A reinforcement learning level-based particle swarm optimization algorithm for large-scale optimization. Inf Sci (NY) 2022;602:298–312.
[141] Biemann M, Scheller F, Liu X, Huang L. Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. Appl Energy 2021;298:117164.