International Journal of Applied Science; Vol. 8, No. 3; 2025
ISSN 2576-7240 E-ISSN 2576-7259
https://doi.org/10.30560/ijas.v8n3p108

A Review on AI-Driven Optimization of Data Center Energy Efficiency and Thermal Management

Xiyong LI[1], Zhiming ZHAO[1], Xiangjun JIANG[2], Yingmei CHEN[2] & Ruxuan HE[2]

[1] Collage of Mechanical and Electrical Engineering, Shaanxi University of Science & Technology, Xi'an 710021, China

[2] State Key Laboratory of Electromechanical Integrated Manufacturing of High-Performance Electronic Equipments, Xidian University, Xi'an, Shaanxi, China

Correspondence: Xiangjun JIANG, State Key Laboratory of Electromechanical Integrated Manufacturing of High-Performance Electronic Equipments, Xidian University, Xi'an, Shaanxi, China

Received: September 14, 2025; Accepted: September 26, 2025; Published: September 27, 2025

Abstract

As data centers grow and face energy challenges, traditional thermal management struggles with dynamic loads, multi-scale coupling, and heterogeneous control. This review examines AI-driven solutions for energy efficiency, focusing on integrating deep learning and reinforcement learning. Key innovations include physics-data hybrid models and constrained RL controllers, achieving PUE<1.2, a 55.7% reduction in fan energy, and enhanced thermal stability. Challenges remain in explainable decision-making, hardware compatibility, and the complexity of multi-physics simulation. Our evaluation framework emphasizes PUE and energy savings, advocating for future advancements in digital twins, edge AI deployment, and renewable cooling integration. Policy-supported AI implementation could increase annual energy savings to 8-12%, promoting sustainable digital infrastructure. Future research should explore multi-scale optimization, reliable AI mechanisms, and renewable-cooling coordination to meet dynamic demand and support carbon neutrality goals.

Keywords: Artificial Intelligence, Data Center, Energy Consumption, Thermal Management, Deep Learning, Reinforcement Learning

1. Introduction

Since the Third Industrial Revolution, rapid advances in information technology (IT) have profoundly transformed modern society. This growth is embodied by the continuous expansion of data centers (DCs), which aggregate large-scale IT infrastructure[1]. By 2019, China hosted approximately 74,000 data centers of various sizes, as reported by CCID Consulting [2]. Globally, electricity consumption by data centers increased sharply from 152 billion kWh in 2005 to about 238 billion kWh in 2010 [3]. During 2000–2010, the average annual growth rate of energy use in data centers reached 11%, far exceeding the 3% growth observed across the general industrial sector [4].

In China alone, data center energy consumption reached 150.7 billion kWh in 2020, representing nearly 2% of the country's total electricity use and resulting in approximately 90 million metric tons of carbon dioxide emissions [5]. With the rising demands of artificial intelligence (AI) and big data, power densities per rack are also increasing—roughly one-third of data centers worldwide now use racks rated at 30 kW or higher. According to the International Energy Agency (IEA), energy consumption from data centers could double by 2026 due to AI and cryptocurrency operations [6,7]. These trends underscore the urgent economic and environmental challenges posed by data centers and highlight the need for more efficient energy management practices.

Current performance metrics reveal significant opportunities for improvement. Many facilities exhibit Power Usage Effectiveness (PUE) values between 2.00 and 2.49, indicating substantial energy waste [8]. As a result, enhancing the efficiency of both IT and cooling systems has become a major focus for operators. Cooling equipment, in particular, is the largest energy-consuming auxiliary system, accounting for 30–50% of a data center's total energy use [9,10]. Effective thermal management is essential to maintain reliable operating conditions and prevent hardware failures [11].

Research efforts have explored various strategies to reduce energy consumption in cooling systems, including improving equipment efficiency [12], optimizing room layout [13], enhancing airflow organization [14,15],

transitioning from room-level to rack-level cooling [16], integrating natural cooling sources [17], and adopting liquid cooling technologies [18]. As gains from hardware improvements diminish, the role of intelligent control systems in achieving precise and efficient cooling has become an increasingly critical area of research [19].

Air cooling remains the dominant solution, valued for its simplicity, reliability, and low cost. Nevertheless, it suffers from inherent limitations due to the low density and heat capacity of air [20]. Furthermore, the design and control of air-cooling systems have a decisive impact on overall energy performance [21]. Current thermal management strategies face several challenges. First, control systems are typically divided into chip-level, rack-level, and system-level, yet research has tended to optimise these in isolation, with limited progress on coordinated control across levels [22,23]. Second, the thermal dynamics within servers and rooms are highly complex, shaped by interactions among CPUs, memory, drives, and other components. These interactions complicate both thermal modelling and monitoring [24]. Third, model accuracy is often constrained by the complexity of thermal states and the high dimensionality of influencing variables. Finally, thermal response is subject to significant latency: while high-fidelity Computational Fluid Dynamics (CFD) models offer precision, they are computationally expensive, whereas simplified models may lack sufficient accuracy [25].

Addressing these challenges requires a comprehensive review of existing control methods and strategies. This study focuses on conventional air-cooling systems, with particular attention to improving the performance of server fans and Computer Room Air Conditioning (CRAC) units through advanced control techniques. Such optimisation provides multiple benefits: reducing energy consumption, improving thermal load management, and meeting operational requirements [26]. More precise fan control, for instance, can reduce hotspots, achieve more uniform cooling, and enhance overall system reliability [27]. In turn, effective cooling strategies help to safeguard servers from failure and minimise downtime [28].

This paper systematically reviews the application of artificial intelligence in data centre thermal management, with a focus on server fan regulation, CRAC optimisation, and coordinated multi-equipment control. The objective is to build an intelligent management framework that is dynamically adaptive, highly efficient, and energy-saving. Such a framework promises improved thermal regulation, closer integration between cooling systems and IT operations, and enhanced overall reliability. The paper is organised as follows: Section 1 outlines the importance of energy efficiency and the challenges in data centre cooling. Section 2 reviews key AI technologies, including the evolution of deep learning and advances in reinforcement learning. Section 3 introduces AI-driven optimisation methods, focusing on environmental modelling, closed-loop control, and prediction–optimisation frameworks. Section 4 presents a comparative analysis of modelling approaches and evaluation metrics. Section 5 concludes with the main contributions and future directions for thermal management in data centres.

2. Data Center Energy Management Challenges

2.1 Energy Consumption Indicator System

To evaluate data centre performance, several energy efficiency indicators have been introduced. Among these, Power Usage Effectiveness (PUE), proposed by The Green Grid Association in 2006 [29], remains the most widely used [30]. PUE is defined as the ratio between total facility power (TFP) and total IT power (TITP) [31,32]. As shown in Equation (1).It reflects the proportion of energy consumed directly by IT equipment versus that required for auxiliary operations such as cooling, power distribution, and other building services [33].

$$PUE = \frac{\sum TFP}{\sum TITP} \tag{1}$$

By definition, PUE values are always greater than 1. A theoretical value of 1 would imply that all energy is consumed exclusively by IT equipment, which is unattainable in practice due to unavoidable overhead. However, lower values approaching 1 indicate higher operational efficiency [34]. Thus, PUE serves as a practical benchmark for comparing data centre energy performance.

Although PUE is the dominant metric, it has limitations, as it does not directly account for workload characteristics or energy proportionality across systems. To complement PUE, additional metrics—such as Data Centre Infrastructure Efficiency, Carbon Usage Effectiveness (CUE), and Water Usage Effectiveness (WUE)—have been developed [35-37]. These provide broader perspectives on sustainability by including carbon emissions, water consumption, and resource utilisation, offering a more holistic view of data centre efficiency.

2.2 Key Thermal Management Challenges

Thermal gradients within data centres significantly affect equipment reliability, energy efficiency, and operating costs. These gradients arise from uneven heat transfer across the facility, leading to inter-rack temperature differences, fluctuations at server surfaces, and vertical stratification in rooms. According to ASHRAE standards (2010) [38], hardware failure rates increase sharply once local temperatures exceed recommended thresholds (e.g., 38 °C for Class A1 environments). As such, managing thermal gradients is a central challenge in data centre design.

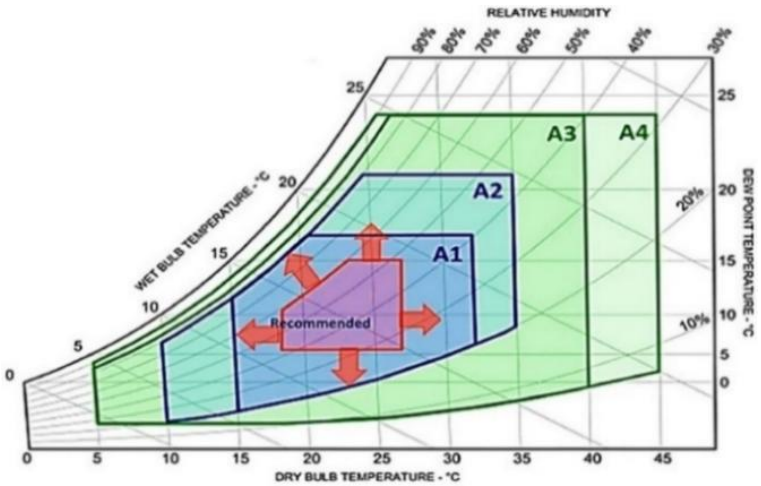

Figure 1. Allowable thermal conditions for A1~A4

At the macro scale, mismatches between cooling supply and workload distribution create non-linear thermal fields. Uniform airflow delivery, typical in conventional room-level cooling, struggles to address heterogeneous loads [39]. High-density racks (>15 kW/rack) often experience inadequate cooling due to airflow imbalances such as the Venturi effect [40]. At the micro scale, chip-level power variations and 3D packaging intensify intra-server heat accumulation [41].

Airflow management deficiencies further exacerbate gradients. Hot air recirculation can raise inlet temperatures by 5–8 °C, while bypass airflow exceeding 40% reduces cooling efficiency by more than 30% [42]. Moreover, transient events—such as virtual machine migrations—can trigger sudden thermal oscillations, complicating stable control [43].

Workload volatility is another major challenge for thermal management. These fluctuations, driven by elastic service demands, virtualisation, and AI workloads, cause rapid changes in server power consumption [44]. Examples include second-scale traffic surges in online services, abrupt local power shifts during VM migration [45,46], and pulsed loads in GPU clusters during training [47].

Such dynamics create three main problems. First, cooling systems like CRAC units, with minute-scale response times, cannot match the speed of workload spikes, which may exceed 200% within seconds [48]. Second, over-provisioned cooling infrastructure, sized for peak demand, often runs inefficiently under typical loads, with utilisation rates reported below 40% [49]. Third, while techniques such as dynamic voltage and frequency scaling (DVFS) can lower chip temperatures, they may increase overall energy use due to task delays [50].

2.3 Analysis of Existing Technical Bottlenecks

Current thermal management technologies in data centers face multiple constraints that impede the co-optimization of energy efficiency and reliability.

Firstly, multi-scale thermal coupling modeling suffers from an accuracy-realtime trade-off: While traditional CFD simulations enable high-fidelity temperature field prediction, full-scale room modeling requires hours to days[51], rendering it impractical for minute-scale response to dynamic workloads. Conversely, simplified lumped parameter models neglect local turbulent effects, incurring >15% prediction errors in high-density rack scenarios (>30kW/rack) [52], which compromises transient thermal shock anticipation (e.g., second-scale GPU temperature spikes).Secondly, control strategies exhibit problematic static and experience-dependent characteristics: Fixed-threshold PID control and rule-based scheduling frequently cause overshoot during load surges, with Google case studies showing 40% increased temperature fluctuation standard deviation and 12-18% overcooling energy

waste[53]. The strong multivariable coupling in air-liquid hybrid cooling systems further exacerbates coordinated control complexity. Thirdly, spatiotemporal resolution limitations in monitoring networks are pronounced: Existing sensor networks deployed at 5-10m intervals with sub-1Hz sampling rates [54] cannot capture vertical thermal stratification within racks or chip-level micro-hotspots. Rack sensor density reductions of 50% increase hotspot miss probabilities from 8% to 35%[55]. Fourthly, the Pareto dilemma between energy efficiency and reliability remains unresolved: Raising cold-aisle temperatures improves PUE but reduces CPU lifespan to 32% when junction temperatures approach the 85°C threshold [56]. Conversely, excessive temperature control degrades PUE beyond 1.5 - a particularly acute conflict in constrained environments like edge data centers. Finally, sustainable cooling technologies face scalability barriers: Practical adoption is hindered by phase change materials' (PCM) limited latent heat capacity (<200 kJ/kg) (Wang et al., 2021)[57], thermoelectric coolers' (TEC) efficiency collapse under high loads[58], and >20% performance degradation in low-GWP refrigerants [59].

These bottlenecks collectively reveal fundamental limitations in dynamic responsiveness, multi-objective optimization, and cross-scale coordination within contemporary thermal management paradigms, necessitating innovative approaches that transcend conventional theoretical frameworks.

3. Foundations of Next-Generation AI Technologies

As data centers grow in scale and face rising energy challenges, the convergence of Deep Learning (DL) and Reinforcement Learning (RL) has become a critical driver of efficiency. This section reviews research progress from two perspectives: the evolution of system architectures and advances in algorithmic design. Together, these developments illustrate the shift from static, single-objective optimization toward dynamic coordination across multiple subsystems.

3.1 The Evolution of Deep Learning Technology Architecture

Deep Learning has enhanced the ability to characterize, predict, and control complex energy systems through innovations in model design, training efficiency, and hardware integration. In terms of model architectures, researchers have moved beyond single-modal designs by integrating heterogeneous data sources. For example, Ran et al. [60,61] proposed a Parameterized Deep Q-Network (PADQN) that jointly encodes discrete scheduling and continuous cooling into one action space. By introducing a temporal factor mechanism (PADQN-D), they coordinated IT scheduling on a minute scale with cooling control on a second scale, reducing overall energy use by 15%. Building on this, Zhou et al. [62] coupled computational fluid dynamics with deep reinforcement learning to develop a digital twin platform (Figure. 2(a)). This system achieved a 9% reduction in IT energy and a 15% reduction in cooling energy by aligning thermal-aware scheduling with load-aware cooling. Similarly, Wang and Yi [63,64] applied Long Short-Term Memory (LSTM) networks to predict network traffic and server thermal inertia, enabling optimized routing and job allocation that cut network energy by 10% and lowered server temperatures by 3°C.

Progress has also been made in training mechanisms. Li et al. [65] improved Deterministic Policy Gradient (DPG) by adding a dual-critic undervaluation elimination mechanism, which reduced prediction bias in cooling control and saved 15% energy. Chi et al. [66] introduced a Multi-Agent Distributed DDPG (MAD3C) framework (Figure. 2(b)) that uses adaptive scoring for coordinated decision-making, achieving a 16.42% reduction in energy use. Yang et al. [67] combined Dueling Double Deep Q-Networks (D3QN) with Value Decomposition Networks (VDN) to coordinate decisions among battery, workload, and waste-heat management agents, reducing renewable energy waste by 18.37%.

At the hardware level, deep reinforcement learning (DRL) has been linked with physical models to improve control precision. Chu et al. [68] embedded thermo-hydraulic models (with less than 5% Nusselt number error) into DRL to dynamically adjust fan duty cycles, cutting fan energy by 55.7%. Wan et al. [69] developed a distributed control algorithm based on the Dyna architecture (Figure. 2(c)), which improved rack-level cooling uniformity through shared reward and state recognition mechanisms. Collectively, these advances demonstrate how innovations in modeling, learning algorithms, and hardware integration are driving more efficient energy management.

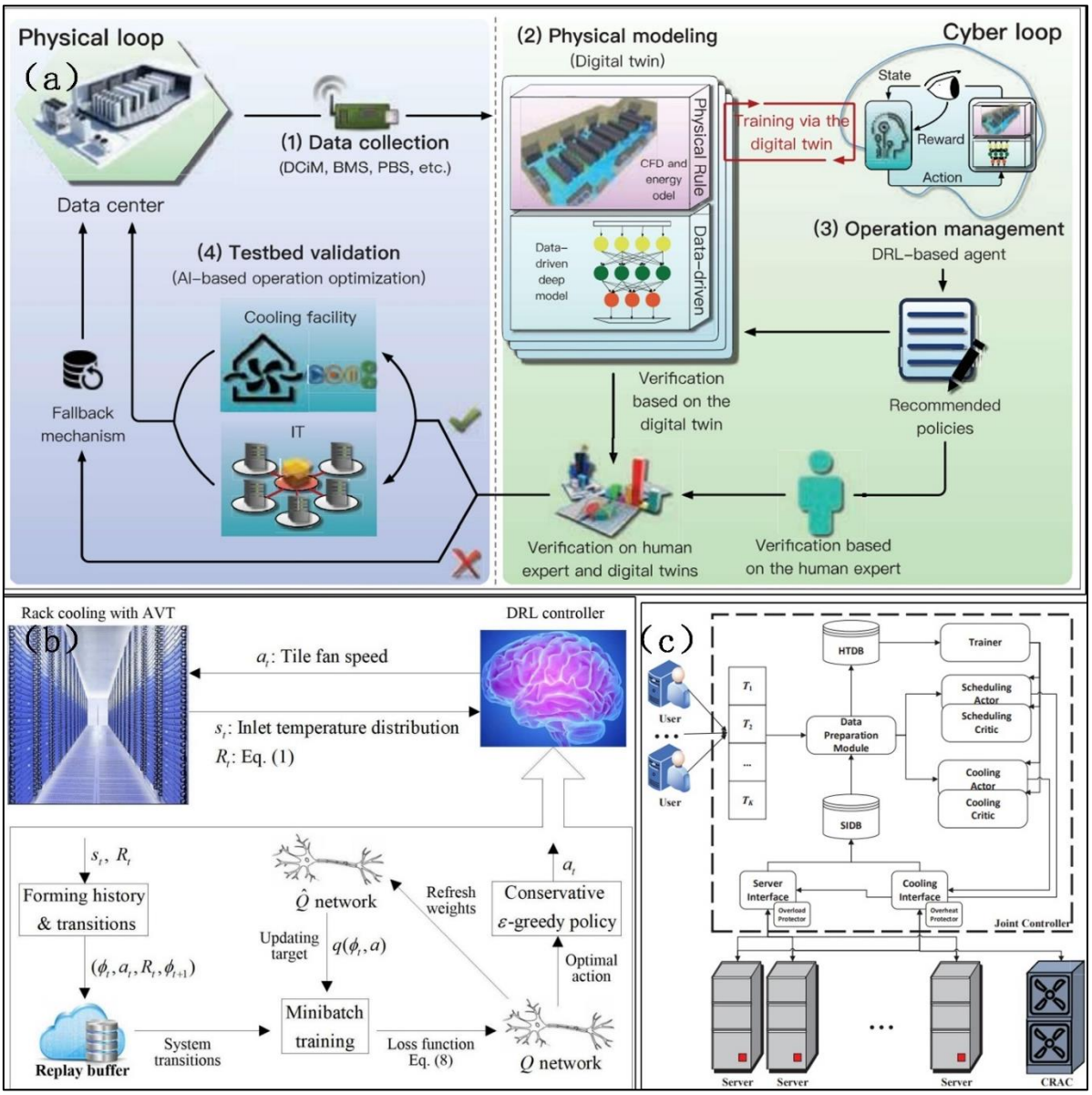

Figure 2. (a) DRL-based workflow for heterogeneous integration solutions; (b) DRL-based joint control system design for training mechanism optimization; (c) Distributed control algorithm based on Dyna architecture at the hardware level

3.2 Scenario-Adaptive Innovations in Reinforcement Learning Paradigms

RL has made notable progress in addressing the optimization challenges of data centers, which operate under strict constraints and rapidly changing conditions. Recent research has focused on three key aspects: ensuring operational safety, improving learning efficiency, and balancing multiple dynamic objectives.

In terms of safe and efficient exploration, Wan et al. [70] proposed SafeCool, a risk-aware model-based RL approach that builds transition models to predict future system states. By integrating risk evaluation with Model Predictive Control (MPC) and rolling optimization, the method dynamically adapts cooling strategies and reduced safety violations to as low as 0.03%.

Progress has also been achieved in how state and action spaces are represented. Shaw et al. [71] developed a continuous percentage space model that translates complex parameters—such as VM utilization and migration costs—into a unified 0–100% scale. When combined with Sarsa and potential-based reward shaping, this method reduced SLA violations by 63%. Similarly, Farahnakian et al. [72] proposed the RL-DC framework, which applies

hierarchical action decomposition: an upper-level Q-learning agent manages host power mode transitions, while a lower-level module selects VM migration targets. This structure significantly reduced decision complexity.

Multi-objective optimization has been another focus. Yan et al. [73] incorporated battery degradation into charge–discharge control by applying prioritized experience replay with DQNs. Their model dynamically balanced cost savings against battery wear, increasing the net revenue of energy storage systems by 55%.

Together, these studies highlight how reinforcement learning can deliver systematic optimization in data centers—achieving safer operation, more efficient learning, and improved balancing of competing objectives.

4. AI-Driven Energy Efficiency Optimization Technologies

4.1 Dynamic Environment Modeling Techniques

Dynamic environment modeling forms the foundation of thermal management optimization in data centers. These techniques establish relationships between environmental parameters and equipment responses, providing the theoretical basis for control strategies. Current research mainly follows two complementary directions: physics-based approaches and hybrid data–physics approaches, which differ in accuracy and computational efficiency but can also reinforce each other.

Physics-based approaches rely on fluid dynamics and thermodynamic equations to describe environmental behavior. One early example is the Temperature Change Index (TCI) model proposed by Bash et al. [74], which used distributed sensors to collect temperature data and adjusted CRAC supply air setpoints with PID control. Building on this, Baxendale et al. [75] introduced regression-based temperature prediction to regulate return air setpoints, updating them every 30 minutes. More recently, Sami Alkharabshe et al. [76] applied CFD simulations to study airflow resistance inside servers, showing that internal resistance in modular data centers can reduce airflow by up to 60%. This finding provides a physical basis for optimizing fan design and layout. While such methods offer clear physical interpretability, their high computational cost limits their use in real-time control.

To overcome this limitation, hybrid data–physics approaches combine physical modeling with data-driven techniques. Simon et al. [77] integrated CFD simulations with artificial neural networks (ANNs), using Latin Hypercube Sampling to generate diverse CFD datasets for ANN training. This enabled multi-unit cooling control with better energy efficiency than traditional PID methods. Similarly, Lorenzi et al. [78] replaced computationally intensive CFD with neural networks that predict server inlet and CRAC return temperatures based on airflow and power inputs. When embedded in cooling system models, this approach supported closed-loop control and achieved 30% annual energy savings. These methods use physical models to generate training data while employing data-driven models to capture nonlinear dynamics, striking a balance between accuracy and real-time performance.

Further progress has been made by embedding hybrid models into broader control frameworks. Choi et al. [79] developed a Cyber-Physical System (CPS) that integrates physical modeling with adaptive algorithms and Bayesian hyperparameter tuning, enabling accurate prediction and stable cooling control. Huang et al. [80] proposed a hierarchical approach in which air conditioning systems are modeled using both low-order linear physical models and high-order nonlinear Random Vector Functional Link (RVFL) neural networks [81]. This dual-structure method efficiently captures complex nonlinear behaviors while maintaining computational efficiency.

4.2 Closed-Loop Control Strategies

The workflow of closed-loop control strategies in data centers operates in the following manner. Sensor data are first transmitted to control servers, where they are aggregated and processed to extract key control parameters. These parameters are subsequently fed into evaluation algorithms, which incorporate both classical control methods and agent-based strategies. The algorithms generate updated setpoints, which are then dispatched to actuators—such as compressors, fans, and dampers in smart ventilation systems—to adjust their operational states. The resulting changes in system conditions are monitored and fed back to the control servers, forming a continuous feedback loop that enables dynamic regulation and optimisation of environmental factors.

As shown in Figure. 3, closed-loop control is shifting the paradigm in data center management from reactive failure remediation toward proactive risk prevention. Through collaboration across industry, academic, and research sectors, these strategies help overcome bottlenecks in real-time performance and system reliability. This approach also supports the objectives of national initiatives such as the “East Data, West Computing” project, which aims to achieve annual energy savings of up to 12%.

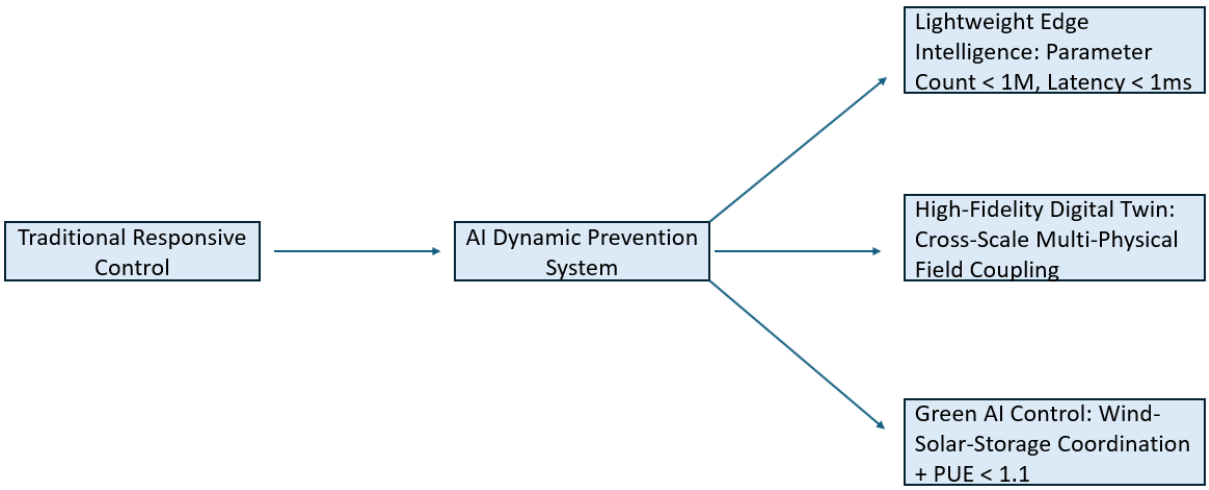

Figure 3. Evolution direction of closed-loop control

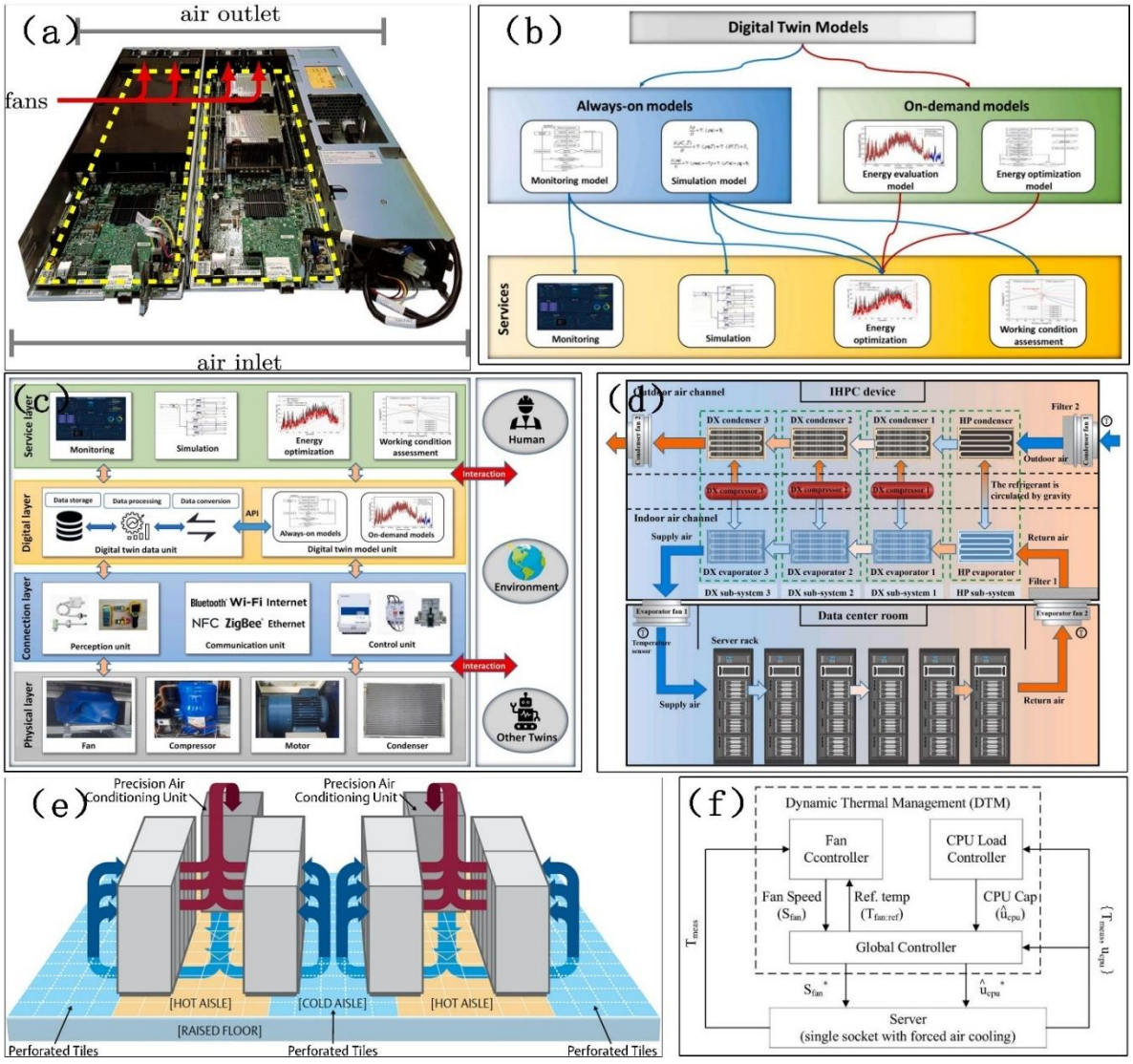

Figure 4.(a) Standardized chassis of an encapsulated server;(b) Digital twin model and its relationship with digital twin services;(c) ECM digital twin architecture of the IHPC system;(d) Schematic of air circulation via the IHPC system in a data center room;(e) A typical air conditioning system;(f) Schematic of the joint control system

Fan systems in data centers regulate indoor airflow through operational adjustments, with layout designs varying across different configurations. As shown in Figure 4(a), standardized chassis often come with predefined fan arrangements aimed at optimising airflow and improving cooling efficiency. Functionally, these fans are categorised by their installation positions: intake fans draw cold air from outside into the facility, while exhaust fans remove hot air from the server environment. Through coordinated operation, they ensure efficient delivery of cold air to server zones and prompt extraction of hot air, thereby avoiding uncontrolled mixing of air streams. This synergy enhances both fan cooling effectiveness and overall thermal management.

However, conventional control strategies—such as fixed-speed operation or basic threshold-based mechanisms—are unable to adapt effectively to dynamic changes in thermal load. This shortcoming is especially evident in multi-heat-source, high-density server clusters, where balancing fan energy use with adequate cooling presents a major challenge.

In response, researchers have begun developing intelligent, algorithm-driven control methods for fan operation. A summary of these results is provided in Table 1.

Table 1. Literature Review on Fan System Selection Design Using Model-Based Control

Authors	Model	Objective	Control variables	Solution
Ghazal Mohsenian et al.[82]	ANN	Improve energy efficiency	fan speed	Build and train the model
Berezovskaya et al.[83]	Server Fan agent	Minimize fan power consumption	fan speed	Adjust fan speed via heuristic rules
Berezovskaya et al.[84]	RL	Reduce fan energy consumption	fan speed	
Brannval et al.[85]	RNN	Low electricity bills	fan speed	Optimization problem
Wang et al.[86]	Power and temperature	Minimize fan power consumption	fan speed	Convex optimization
Zapater et al.[87]	Leak model	Energy consumption is reduced to a minimum.	fan speed	LUT, Bang Bang control strategy
Lucchese et al.[88]	Flow model	Minimize overall cooling costs	fan speed	Discrete-time RHC strategy with fixed sampling period

Beyond server-level strategies, rack-scale control has emerged as a promising direction for improving cooling efficiency. For example, Eiland et al. [89] proposed replacing individual server fans with shared rack-mounted fans, using larger units to cool multiple servers simultaneously. They also examined how fan sizing influences energy consumption. Expanding on this concept, Fernandes et al. [90] developed a control system for rack fans that interfaces via microcomputers with the servers' baseboard management controllers (BMCs). This system dynamically adjusts each rack fan's speed according to the control signals originally intended for the individual server fans.

In parallel, Li et al. [91] focused on optimizing thermal design within data center cabinets. They developed a reduced-order model using Proper Orthogonal Decomposition (POD) combined with an improved Multi-Objective Genetic Algorithm (MOGA). By integrating Kriging surrogate models, their approach reduces the number of simulation calls required to estimate Pareto fronts by 50% compared to conventional MOGA, offering an efficient solution for complex thermal design problems.

Data center cooling systems typically employ dedicated CRAC units and are often designed with segregated hot and cold aisles to improve thermal management. As shown in Figure. 4(e) [92], CRAC units supply cold air (indicated by blue arrows) from lower vents and extract hot air (red arrows) through upper sections. This physical separation guides cold airflow directly to server intakes at the front, while hot exhaust air is efficiently captured from the rear. Containment of air streams reduces unintended mixing, thereby increasing cooling efficiency and overall system performance.

To achieve further energy savings, variable frequency drives and smart control systems can be integrated. These technologies adjust cooling output dynamically according to real-time temperature and humidity conditions, enabling more precise and efficient environmental control. Moreover, CRAC operation significantly affects the server microclimate. In particular, the temperature at server inlets—which is used by infrared equipment to trigger

cooling requests—directly influences the effectiveness of fan cooling. This leads to a critical trade-off: higher inlet temperatures (e.g., above 27°C) reduce CRAC energy use but impair server fan performance, whereas lower inlet temperatures (below 24°C) improve server cooling at the expense of higher CRAC energy consumption. As indicated in Figure. 5, studies suggest an optimal server inlet temperature range of 24–27°C for minimizing total cooling energy [93]. Control strategies for CRAC systems, much like server fan control, can be categorised as model-based or model-free. Key differences between the two arise from operational scales—room-level versus device-level—and variations in thermal response times.

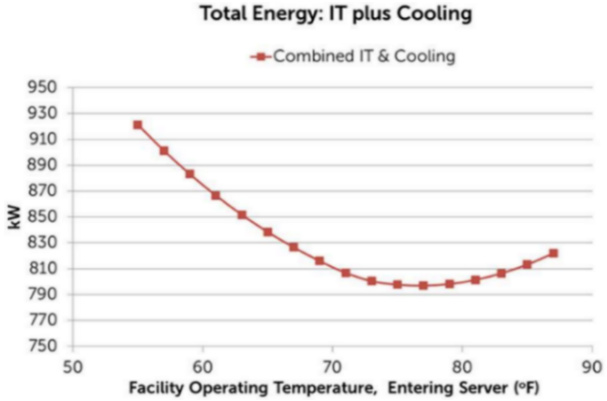

Figure 5. Total Energy -IT plus cooling vs Operating temperature

Recent research in thermal management for data centers has increasingly moved beyond single-component control toward integrated optimization across multiple systems. A prominent example is the coordinated control of server fans and CRAC units. Chen et al. [94], for instance, developed a predictive feedforward-feedback controller based on thermodynamic models that account for time-delay effects. Their approach features a multi-level monitoring system—spanning both server and room scales—a hybrid prediction model combining thermodynamics and machine learning for short-term temperature forecasting, and a dynamic programming algorithm aimed at minimizing energy use under thermal constraints. This coordinated strategy effectively reduces both overcooling and localized hot spots—common issues in decoupled control—while also leveraging IT workload scheduling to lower cooling energy consumption.

In response, Ayoub et al. [95] introduced the GentleCool framework, which incorporates cooling energy costs into workload scheduling decisions through thermal-aware policies. By combining steady-state modeling with thermal simulation, their method enables dynamic optimization across infrastructure layers and improves thermal distribution through task migration at the chip and virtual machine levels—reportedly reducing cooling time by 60% in experimental settings. Their subsequent JETC algorithm further integrates memory and CPU load management with server fan control via PI controllers, improving thermal coupling and energy efficiency.

Concurrently, Huang et al. [96] expanded the scope of optimization to include cooling towers and overall infrastructure. They proposed a hierarchical thermal-aware power management strategy: at the data center level, HVAC setpoints are dynamically adjusted to optimize cooling power, while at the server level, V-shaped power curves are used to guide thermal setpoint tuning. Their experiments demonstrated a 5.4% reduction in total server power.

Despite these advances, widespread model simplifications—such as fixed temperature setpoints and linearized server energy models—remain common. Future progress will likely require greater computational resources to support more complex and dynamic models, enabling more accurate system representations and facilitating the transition from experimental research to real-world implementation.

4.3 Prediction-Optimization Joint Framework

In recent years, research in data center thermal management has undergone a significant shift—from isolated prediction or control approaches toward integrated prediction–optimisation frameworks. By combining dynamic prediction models with intelligent optimisation strategies, these frameworks allow precise regulation of thermal conditions and considerably improve energy efficiency. The following sections review recent advances across three core areas: prediction modelling, optimisation methodologies, and synergistic mechanisms.

The joint framework enables dynamic and intelligent regulation through closed-loop feedback, offering advantages across three key aspects: real-time control, multi-objective coordination, and risk prevention.

In the context of real-time control, Chen et al. [97] developed the PTEC system, which uses sensor networks to predict inlet temperatures and dynamically adjusts fan speeds and cooling equipment, effectively reducing recirculation-related energy losses.

For multi-objective coordination, Xu et al. [98] combined Support Vector Regression (SVR), Logistic Regression (LR), and Light Gradient Boosting Machine (LGBM) to develop a predictive model (RMSE = 0.086). This model, integrated with multi-objective optimization, helps balance energy efficiency and thermal safety.

In the area of Model Predictive Control (MPC), Wan et al. [70] combined an MPC architecture with risk-aware exploration, achieving a 13.18% reduction in cooling energy consumption through rolling-horizon optimisation. Similarly, Yao et al. [99] developed an RLS-MPC framework that adapts model parameters online using recursive least squares, simultaneously lowering total power consumption and improving temperature tracking accuracy.

Beyond MPC, evolutionary algorithms offer efficient solutions for complex optimisation tasks. Song et al. [100] introduced a Chaotic Firefly Genetic Algorithm (CFGA) to tune PID controller parameters, enhancing air conditioning efficiency when coupled with artificial neural network predictions. In a complementary approach, Zapater et al. [101] used Grammar Evolution (GE) to automatically optimise hyperparameters of thermal prediction models, maintaining CPU temperature prediction errors within 2°C.

At the level of task scheduling and resource allocation, Wang et al. [102] designed a server scheduling algorithm based on spatiotemporal temperature forecasting. By minimising inter-rack temperature variations, they reduced the incidence of over-heated and over-cooled servers by 19–32%. Liu et al. [103] employed Long Short-Term Memory (LSTM) networks to predict power patterns of equipment, enabling dynamic load adjustment that cut overall operational costs by 20.6%. These studies highlight the benefit of tightly integrating prediction and optimisation within thermal management systems.

In intelligent cooling systems for data centers, integrated closed-loop feedback frameworks substantially improve both energy efficiency and thermal safety through multi-level dynamic regulation, as depicted in Figure. 4.

As summarised in Table 2, current research indicates that although traditional CFD simulations deliver high predictive accuracy, they require substantial computational resources, often resulting in simulation times ranging from 2 to 72 hours. In contrast, data-driven approaches significantly reduce computation time to values between seconds and several minutes (10 seconds to 5 hours), while maintaining a mean absolute error (MAE) typically within 1–2°C. Hybrid modelling techniques further improve this trade-off by integrating physical constraints with data-driven features. These methods achieve higher accuracy, with MAEs between 0.5–1°C, and simultaneously enhance extrapolation capability beyond the training data.

Table 2. Comparative analysis of data-driven methods and simulation calculations

Authors	Model	Prediction direction	Computation time	Control group under identical conditions with required time
Wan et al [70]	SafeCool	Cooling management	7.01 seconds	The MBRL-MPC algorithm involves complex computations, resulting in prohibitively long decision times
Asgari et al [74]	Grey-box model	Temperature	< 0.5 seconds	< 0.5 seconds
Saiyad et al [93]	ANN	Thermal parameters	10–30 seconds	CFD computational time significantly exceeds that of the proposed model
Song et al [100]	ANN	Thermal operating conditions	Several minutes	Full CFD simulations demand approximately 72 hours
Song et al [104]	VPM and GA	Airflow	Transient computations require only 1 minute	Each transient CFD simulation requires ≈2 hours for convergence
Jin et al [105]	POD - MARS	Temperature field	30 seconds	In contrast, CFD requires ≈2 hours

Authors	Model	Prediction direction	Computation time	Control group under identical conditions with required time
Wang et al [106]	Kalibre	Temperature	Approximately 5 hours	The Kalibre model outperforms CFD

5. Methodological Comparison and Evaluation Framework

5.1 Comparative Analysis of Modeling Paradigms

The rapid advancement of artificial intelligence and cross-disciplinary technologies has spurred diverse innovations in modelling approaches. These range from mechanism-based dynamic physical models to data-driven black-box predictions, grey-box hybrid models, and reinforcement learning-enabled autonomous optimisation. Each paradigm offers distinct advantages in specific applications, yet also faces inherent limitations in adaptability. For example, the VPM zonal model proposed by Song et al. [104] significantly improves computational efficiency but struggles to capture airflow details under off-design conditions. In contrast, Zhu et al.'s digital twin framework [107] supports dynamic optimisation—though it relies on continuous calibration via high-precision sensor networks.

Given these trade-offs, a systematic comparative analysis of modelling paradigms is crucial to assess their applicability boundaries, performance constraints, and potential for integration. Such evaluation is essential for developing efficient, robust, and scalable thermal management systems in data centers. A comprehensive comparison of mainstream paradigms—including dynamic physical models, data-driven modelling, hybrid approaches, and reinforcement learning—is provided in Figure.6, illustrating their architectural layouts and operational workflows.

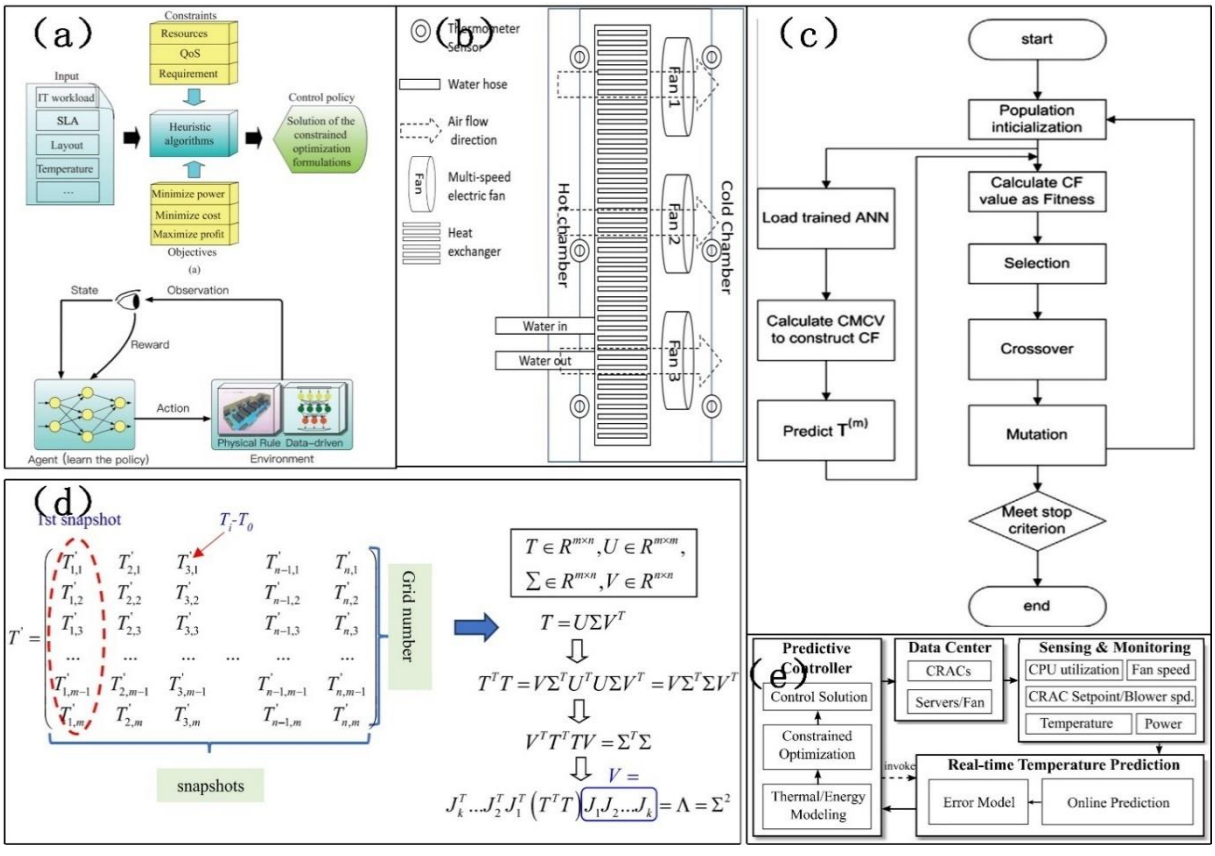

Figure 6. (a) Workflow in reinforcement learning models[105]; (b) Unit cooling architecture of the ALTM model in hybrid models ;(c) ANN-based CFGA flowchart in optimized control models; (d) Snapshot matrix and flowchart of POD modal computation in surrogate models[105]; (e) Thermal balance equation optimizing the feedback framework for fan control in dynamic physical models

5.2 Evaluation Metrics System

PUE serves as a core metric for evaluating energy efficiency in data centers, directly quantifying the ratio between the energy consumed by IT equipment and the total energy used by the facility. According to estimates from [108], the global average PUE is projected to reach approximately 1.55 by 2025 (Fig. 7), although leading facilities have already achieved values below 1.2. Notable contributions driving this progress include: a digital twin-assisted heat pipe cooling system by Zhu et al. [107], which reduced PUE from 1.325 to 1.248; a genetic algorithm-based dynamic load control strategy by He et al. [109] that reached a PUE of 1.221; and a virtualization-guided resource consolidation approach by Jia et al. [110] that consistently maintained PUE below 1.5. Together, these studies illustrate a multi-faceted approach to energy reduction, combining advanced cooling, operational optimization, and renewable integration—with some experimental systems even attaining PUE values under 1.1.

Policy measures are further accelerating efficiency gains through three key mechanisms: mandatory standards (such as the EU's requirement for PUE ≤ 1.3), fiscal incentives (supporting AI-driven optimization research), and transparency frameworks (e.g., Singapore's mandatory PUE reporting).

From an economic perspective, improving PUE offers nonlinear energy savings. Reducing PUE from 2.0 to 1.5 cuts total energy use by 25%, while a further reduction to 1.2 saves an additional 20%. For example, a 10 MW data center can save approximately $1.2 million annually for every 0.1 reduction in PUE, assuming an electricity cost of $0.1 per kWh. Moreover, facilities with lower PUE values often benefit from improved access to green financing, reducing borrowing costs by 1.5 to 2 percentage points.

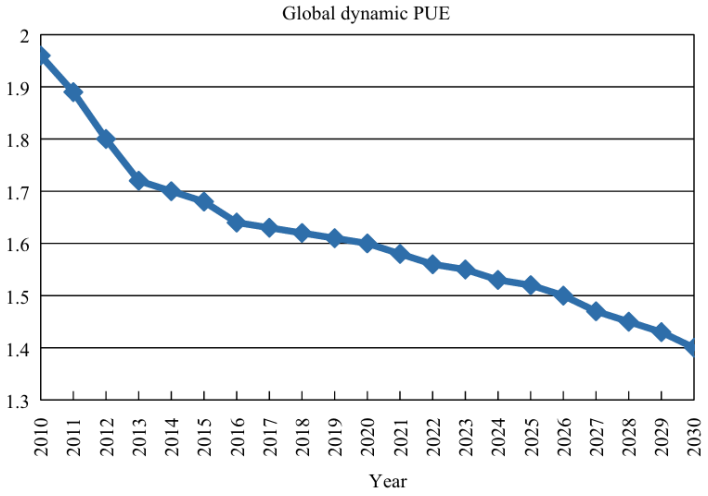

Figure 7. Regression results of the global average PUE of data centers

To systematically assess the practical performance of energy-saving technologies in data centers, this section summarizes key energy-saving metrics from recent representative studies covering multiple technical approaches. These include predictive neural networks, reinforcement learning, digital twins, and genetic algorithms. As presented in Table 4, the energy-saving outcomes and quantitative results achieved by these methods are compiled to offer critical reference information for subsequent technology selection and optimisation.

Table 4. Comparison of energy-saving technology benefits of data centers

Authors	Core Issues	Model	Methods	Results
Hsieh et al [111]	Elevated cooling costs	Gray - Markov	Predicting CPU utilization	9.6%-25.6% energy reduction vs baselines
Pakbaznia et al [112]	High cooling energy	TASP	Load allocation	5.8%-9.3% energy savings
Damme et al [113]	High cooling energy	Thermodynamic Model	Minimizing datacenter energy consumption	Up to 30% energy reduction
Zhou et al [114]	Elevated cooling costs and SLA violation rates	KMI - MPCU	Optimizing resource allocation	>40% energy + SLA violation reduction

Authors	Core Issues	Model	Methods	Results
Banerjee et al [115]	High cooling energy	HTS	Cooling-integrated thermal-aware scheduling	3.81%-16.0% savings vs EDF-LRH
Tang et al [116]	Elevated cooling costs	XInt - GA and XInt - SQP	Minimizing inlet temperature peaks	24%-35% cooling savings (50% util) vs UT/UOP
Lei et al [117]	High cooling energy	OL-PICEA-g	Intelligent scheduling strategies	Superior C-metric vs PICEA-g/OL-PICEA-g

As shown in the table, deep reinforcement learning and neural networks are particularly effective in multi-variable cooperative optimisation, delivering energy savings of up to 60%. Network-layer optimisation also shows significant potential; through topology reconfiguration and adaptive routing strategies, communication energy consumption can be reduced by 20% to 60%. Meanwhile, improving cooling systems remains a central challenge, with integrated control strategies consistently achieving average energy savings between 15% and 30%. It should be noted, however, that the applicability of each technology depends on factors including data center scale, workload profiles, and existing infrastructure configurations.

6. Conclusions and Outlook

The application of AI in data center energy management and optimisation continues to evolve. Techniques such as dynamic environmental modelling, closed-loop control, and integrated prediction-optimisation frameworks have significantly improved the energy efficiency of cooling systems and the precision of thermal management. AI methods—particularly deep learning and reinforcement learning—have led to notable advances in key areas including server fan control, CRAC operation, and coordinated management of multiple devices. These improvements are supported by multimodal model integration, adaptive algorithms, and hardware-in-the-loop optimisation. As a result, PUE values have been optimised—in some instances below 1.2—cooling energy consumption has been reduced by up to 55.7%, and operational stability under varying workloads has improved.

Globally, data centers are transitioning from conventional designs toward smarter and more sustainable infrastructures. Under hybrid deployment models, the co-existence of high-density computing equipment, heterogeneous hardware, and legacy systems intensifies challenges such as uneven thermal distribution, slow response to dynamic loads, and difficulties in modelling multi-scale thermal interactions. Although AI shows great promise, several limitations remain: current models struggle to adapt to emerging cooling technologies such as immersion cooling and edge computing; the lack of transparency in AI decision-making hinders trust in critical operations; and high computational costs for multi-physics simulations and real-time control continue to pose barriers.

Future research should focus on the following five priorities:

1. Multi-Physics Digital Twins: Develop integrated digital twin platforms that combine CFD-based thermal models, power system dynamics, and workload forecasting. These systems should address current inaccuracies—for example, in immersion cooling (where thermal conductivity errors may exceed 20%) or heterogeneous hardware setups. Incorporating reduced-order modelling techniques, such as tensor decomposition, could accelerate simulation speeds by more than tenfold.

2. Edge Intelligence: For edge data centers with limited resources, lightweight deep reinforcement learning frameworks—with under one million parameters—should be developed to enable local inference within 50 milliseconds using knowledge distillation. Federated learning could further support privacy-preserving collaboration across nodes.

3. Trust and Security: Improve the interpretability of AI decisions through causal inference models and establish probabilistic bounds for thermal risks with high confidence (e.g., >99.9%). Compliance with international standards such as ISO 30134 will be essential for deploying AI control in critical infrastructure.

4. Adaptation to New Cooling Technologies: AI should be tailored to support liquid cooling (offering 40 times greater thermal capacity than air) and phase-change materials (with latent heat over 200 kJ/kg). Physics-informed reinforcement learning can help achieve microsecond-level response times and suppress nonlinear thermal surges.

5. Green Energy Integration: Develop joint optimisation models coordinating cooling with renewable energy sources (wind, solar) and storage. Battery lifetime-aware reinforcement learning can increase renewable utilisation, and virtualisation technologies may help prototype zero-carbon data center systems.

Driven by policy goals—such as the EU requirement of PUE ≤ 1.3—and cross-sector collaboration, AI is expected to help data centers achieve annual energy savings of 8–12%, supporting a sustainable digital economy worldwide.

Acknowledgments

This work was supported by the Natural Science Foundation of Shaanxi Province (grant number 2025JC-639);Industrialization cultivation project of the Department of Education (23JC014).

References

[1] Nadjahi, C., Louahlia, H., & Lemasson, S. (2018). A review of thermal management and innovative cooling strategies for data center. *Sustainable Computing: Informatics and Systems, 19*, 14–28. https://doi.org/10.1016/j.suscom.2018.05.002

[2] Koomey, J. G. (2008). Worldwide electricity used in data centers. *Environmental Research Letters, 3*(3), 034008. https://doi.org/10.1088/1748-9326/3/3/034008

[3] Koomey, J. (2011). *Growth in data center electricity use 2005 to 2010* (A Report by Analytical Press, Completed at the Request of The New York Times). Analytical Press.

[4] Garimella, S. V., Persoons, T., Weibel, J., et al. (2013). Technological drivers in data centers and telecom systems: Multiscale thermal, electrical, and energy management. *Applied Energy, 107*, 66–80. https://doi.org/10.1016/j.apenergy.2013.02.047

[5] Basmadjian, R., Bouvry, P., Costa, G. D., & Khan, S. U. (2015). Green data centers. In *Large-Scale Distributed Systems and Energy Efficiency: A Holistic View* (pp. 159–196). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118981122.ch6

[6] Berreby, D. (2024). As use of AI soars, so does the energy and water it requires. *Yale Environment 360*.

[7] Eren, G., Hoe, S., Niklas, S., et al. (2024). *Electricity 2024—Analysis and forecast to 2026* (Report). International Energy Agency (IEA).

[8] Isazadeh, A., Ziviani, D., & Claridge, D. E. (2023). Global trends, performance metrics, and energy reduction measures in datacom facilities. *Renewable and Sustainable Energy Reviews, 174*, 113149. https://doi.org/10.1016/j.rser.2023.113149

[9] Meijer, G. I. (2010). Cooling energy-hungry data centers. *Science, 328*(5976), 318–319. https://doi.org/10.1126/science.1182769

[10] Salim, M., & Tozer, R. (2010). Data centers' energy auditing and benchmarking—Progress update. *ASHRAE Transactions, 116*(1). https://doi.org/10.1109/ITHERM.2010.5501366

[11] Thome, J. R., Marcinichen, J. B., & Olivier, J. A. (2012). Two-phase on-chip cooling systems for green data centers. In *Energy Efficient Thermal Management of Data Centers* (pp. 513–567). Springer. https://doi.org/10.1007/978-1-4419-7124-1_12

[12] Zhang, Q., Meng, Z., Hong, X., et al. (2021). A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization. *Journal of Systems Architecture, 119*, 102253. https://doi.org/10.1016/j.sysarc.2021.102253

[13] Zhan, B., Shao, S., Lin, M., et al. (2021). Experimental investigation on ducted hot aisle containment system for racks cooling of data center. *International Journal of Refrigeration, 127*, 137–147. https://doi.org/10.1016/j.ijrefrig.2021.02.006

[14] Chu, W. X., Wang, R., Hsu, P. H., et al. (2020). Assessment on rack intake flowrate uniformity of data center with cold aisle containment configuration. *Journal of Building Engineering, 30*, 101331. https://doi.org/10.1016/j.jobe.2020.101331

[15] Song, P., Zhang, Z., & Zhu, Y. (2021). Numerical and experimental investigation of thermal performance in data center with different deflectors for cold aisle containment. *Building and Environment, 200*, 107961. https://doi.org/10.1016/j.buildenv.2021.107961

[16] Cho, J., & Lim, S. (2023). Balanced comparative assessment of thermal performance and energy efficiency for three cooling solutions in data centers. *Energy, 285*, 129370. https://doi.org/10.1016/j.energy.2023.129370

[17] Cui, X., Yang, C., Yan, W., et al. (2024). Climatic applicability of indirect evaporative cooling strategies for data centers in China. *Journal of Building Engineering, 83*, 108431. https://doi.org/10.1016/j.jobe.2023.108431

[18] Chethana, G. D., & Gowda, B. S. (2021). Thermal management of air and liquid cooled data centres: A review. *Materials Today: Proceedings, 45*, 145–149. https://doi.org/10.1016/j.matpr.2020.10.396

[19] Du, Y., Zhou, Z., Yang, X., et al. (2023). Dynamic thermal environment management technologies for data center: A review. *Renewable and Sustainable Energy Reviews, 187*, 113761. https://doi.org/10.1016/j.rser.2023.113761

[20] Wan, J., Gui, X., Kasahara, S., et al. (2018). Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey. *IEEE Access, 6*, 48867–48901. https://doi.org/10.1109/ACCESS.2018.2866840

[21] Patel, D., Sharma, R., Bash, C., et al. (2002). Thermal considerations in data center design. In *Proceedings of the Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems* (pp. 767–776). IEEE.

[22] Cui, D., Zhou, L., Luo, Q., et al. (2023). Multi-scale modeling and fast inference for thermal environment analysis of air-cooled data center. *Journal of Building Engineering, 78*, 107722. https://doi.org/10.1016/j.jobe.2023.107722

[23] Dai, Y., Zhao, J., Shi, J., et al. (2021). A multi-scale thermal analysis method for data centers with application in a ship data center. *Journal of Thermal Science, 30*, 1973–1985. https://doi.org/10.1007/s11630-021-1494-4

[24] Liu, W., Tong, X., Wang, J., et al. (2022). Real-time temperature predictions via state-space model and parameters identification within rack-based cooling data centers. *Journal of Building Engineering, 58*, 105013. https://doi.org/10.1016/j.jobe.2022.105013

[25] Zhabelova, G., Vesterlund, M., Eschmann, S., et al. (2018). A comprehensive model of data center: From CPU to cooling tower. *IEEE Access, 6*, 61254–61266. https://doi.org/10.1109/ACCESS.2018.2875623

[26] Beghi, A., Cecchinato, L., Danza, L., et al. (2017). Modeling and control of a free cooling system for data centers. *Energy Procedia, 140*, 447–457. https://doi.org/10.1016/j.egypro.2017.11.156

[27] Wang, Z., Bash, C., Tolia, N., et al. (2009). Optimal fan speed control for thermal management of servers. In *Proceedings of the ASME 2009 InterPACK Conference* (pp. 709–719). ASME. https://doi.org/10.1115/InterPACK2009-89074

[28] Gao, X., Liu, G., Xu, Z., et al. (2022). Investigating security vulnerabilities in a hot data center with reduced cooling redundancy. *IEEE Transactions on Dependable and Secure Computing, 19*, 208–226. https://doi.org/10.1109/TDSC.2020.2977292

[29] Gao, P., Liu, H., Luo, H., et al. (2024). Discussion on the technical path of data center information and communication thermal management. *Energy Reports, 11*, 2704–2714. https://doi.org/10.1016/j.egyr.2024.02.003

[30] Santos, A. F., Gaspar, P. D., & Souza, H. J. L. (2020). New data center performance index: Perfect design data center—PDD. *Climate, 8*(10), 110. https://doi.org/10.3390/cli8100110

[31] The Green Grid. (2012). *PUE: A comprehensive examination of the metric* (White Paper 49).

[32] Liu, Y., Wei, X., Xiao, J., et al. (2020). Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers. *Global Energy Interconnection, 3*(3), 272–282. https://doi.org/10.1016/j.gloei.2020.07.008

[33] Pogorelskiy, S., & Kocsis, I. (2023). BIM and computational fluid dynamics analysis for thermal management improvement in data centres. *Buildings, 13*(10), 2636. https://doi.org/10.3390/buildings13102636

[34] Dayarathna, M., Wen, Y., & Fan, R. (2015). Data center energy consumption modeling: A survey. *IEEE Communications Surveys & Tutorials, 18*(1), 732–794. https://doi.org/10.1109/COMST.2015.2481183

[35] Azevedo, D., Belady, S. C., & Pouchet, J. (2011). *Water usage effectiveness (WUE): A green grid datacenter sustainability metric* (The Green Grid White Paper 32).

[36] Lei, N., & Masanet, E. (2020). Statistical analysis for predicting location-specific data center PUE and its improvement potential. *Energy, 201*, 117556. https://doi.org/10.1016/j.energy.2020.117556

[37] Evans, T. (2004). *The different types of air conditioning equipment for IT environments* (White Paper 59). APC.

[38] ASHRAE Technical Committee (TC) 9.9. (2011). *Thermal guidelines for data processing environments—expanded data center classes and usage guidance* (Whitepaper). ASHRAE.

[39] Sharma, R., Bash, C., & Patel, C. (2002). Dimensionless parameters for evaluation of thermal design and performance of large-scale data centers. In *Proceedings of the 8th AIAA/ASME Joint Thermophysics and Heat Transfer Conference*. AIAA. https://doi.org/10.2514/6.2002-3091

[40] Wan, J., Gui, X., Kasahara, S., et al. (2018). Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey. *IEEE Access, 6*, 48867–48901. https://doi.org/10.1109/ACCESS.2018.2866840

[41] Wang, Z., Dong, R., Ye, R., et al. (2024). A review of thermal performance of 3D stacked chips. *International Journal of Heat and Mass Transfer, 235*, 126212. https://doi.org/10.1016/j.ijheatmasstransfer.2024.126212

[42] Shrivastava, S., Sammakia, B., Schmidt, R., et al. (2005). Comparative analysis of different data center airflow management configurations. In *Proceedings of the International Electronic Packaging Technical Conference and Exhibition* (pp. 329–336). ASME. https://doi.org/10.1115/IPACK2005-73234

[43] Rani, R., & Garg, R. (2021). A survey of thermal management in cloud data centre: Techniques and open issues. *Wireless Personal Communications, 118*, 679–713. https://doi.org/10.1007/s11277-020-08039-x

[44] Du, Y., Zhou, Z., Yang, X., et al. (2023). Dynamic thermal environment management technologies for data center: A review. *Renewable and Sustainable Energy Reviews, 187*, 113761. https://doi.org/10.1016/j.rser.2023.113761

[45] Aghasi, A., Jamshidi, K., Bohlooli, A., et al. (2023). A decentralized adaptation of model-free Q-learning for thermal-aware energy-efficient virtual machine placement in cloud data centers. *Computer Networks, 224*, 109624. https://doi.org/10.1016/j.comnet.2023.109624

[46] Liu, Z., Wierman, A., Chen, Y., et al. (2013). Data center demand response: Avoiding the coincident peak via workload shifting and local generation. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems* (pp. 341–342). ACM. https://doi.org/10.1145/2465529.2465740

[47] Armoni, M. (2024). *Tensor processing units (TPU): A technical analysis and their impact on artificial intelligence*.

[48] Jacquet, P. (2024). *Enhancing IaaS consolidation with resource oversubscription* [Doctoral dissertation, Université de Lille].

[49] Hota, A. (2024). AI-enhanced cooling systems: Innovations in heat management for hyperscale data centers. *International Journal of Engineering Research & Technology (IJERT), 13*.

[50] Masoudi, J., Barzegar, B., & Motameni, H. (2021). Energy-aware virtual machine allocation in DVFS-enabled cloud data centers. *IEEE Access, 10*, 3617–3630. https://doi.org/10.1109/ACCESS.2021.3136827

[51] Vangilder, J. W., & Schmidt, R. R. (2005). Airflow uniformity through perforated tiles in a raised-floor data center. In *Proceedings of the International Electronic Packaging Technical Conference and Exhibition* (pp. 493–501). ASME. https://doi.org/10.1115/IPACK2005-73375

[52] Ling, Y. Z., Zhang, X. S., Zhang, K., et al. (2017). On the characteristics of airflow through the perforated tiles for raised-floor data centers. *Journal of Building Engineering, 10*, 60–68. https://doi.org/10.1016/j.jobe.2017.01.002

[53] Sunkara, K. N. K. C. (2025). *Power consumption and heat dissipation in AI data centers: A comparative analysis*.

[54] Tang, Q., Mukherjee, T., Gupta, S. K. S., et al. (2006). Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters. In *Proceedings of the 2006 Fourth International Conference on Intelligent Sensing and Information Processing* (pp. 203–208). IEEE. https://doi.org/10.1109/ICISIP.2006.4286097

[55] Van Der Broeck, C. H., Conrad, M., & De Doncker, R. W. (2015). A thermal modeling methodology for power semiconductor modules. *Microelectronics Reliability, 55*(9-10), 1938–1944. https://doi.org/10.1016/j.microrel.2015.06.102

[56] Wang, Y., Zhang, Y., Nörtershäuser, D., et al. (2022). Model and data driven transient thermal system modelings for contained data centers. *Energy and Buildings, 258*, 111790. https://doi.org/10.1016/j.enbuild.2021.111790

[57] Skach, M., Aurora, M., Hsu, C. H., et al. (2017). Thermal time shifting: Decreasing datacenter cooling costs with phase change materials. *IEEE Internet Computing*. https://doi.org/10.1109/MIC.2017.2911418

[58] Gao, L. J., Xu, H. J., Zhang, X., et al. (2023). Numerical investigation on thermal performance of thermoelectric-cooler integrated cold plate of thermal control liquid loop in spacecraft. *International Communications in Heat and Mass Transfer, 142*, 106620. https://doi.org/10.1016/j.icheatmasstransfer.2023.106620

[59] Gao, T., Sammakia, B. G., Geer, J., et al. (2015). Comparative analysis of different in row cooler management configurations in a hybrid cooling data center. In *Proceedings of the International Electronic Packaging Technical Conference and Exhibition*. ASME. https://doi.org/10.1115/IPACK2015-48069

[60] Ran, Y., Hu, H., Zhou, X., et al. (2019). Deepee: Joint optimization of job scheduling and cooling control for data center energy efficiency using deep reinforcement learning. In *Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)* (pp. 645–655). IEEE. https://doi.org/10.1109/ICDCS.2019.00070

[61] Ran, Y., Hu, H., Wen, Y., et al. (2022). Optimizing energy efficiency for data center via parameterized deep reinforcement learning. *IEEE Transactions on Services Computing, 16*(2), 1310–1323. https://doi.org/10.1109/TSC.2022.3184835

[62] Zhou, X., Wang, R., Wen, Y., et al. (2021). Joint IT-facility optimization for green data centers via deep reinforcement learning. *IEEE Network, 35*(6), 255–262. https://doi.org/10.1109/MNET.011.2100101

[63] Wang, Y., Li, Y., Wang, T., et al. (2022). Towards an energy-efficient data center network based on deep reinforcement learning. *Computer Networks, 210*, 108939. https://doi.org/10.1016/j.comnet.2022.108939

[64] Yi, D., Zhou, X., Wen, Y., et al. (2019). Toward efficient compute-intensive job allocation for green data centers: A deep reinforcement learning approach. In *Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)* (pp. 634–644). IEEE. https://doi.org/10.1109/ICDCS.2019.00069

[65] Li, Y., Wen, Y., Tao, D., et al. (2019). Transforming cooling optimization for green data center via deep reinforcement learning. *IEEE Transactions on Cybernetics, 50*(5), 2002–2013. https://doi.org/10.1109/TCYB.2019.2927410

[66] Chi, C., Ji, K., Song, P., et al. (2021). Cooperatively improving data center energy efficiency based on multi-agent deep reinforcement learning. *Energies, 14*(8), 2071. https://doi.org/10.3390/en14082071

[67] Yang, D., Wang, X., Shen, R., et al. (2024). Global optimization strategy of prosumer data center system operation based on multi-agent deep reinforcement learning. *Journal of Building Engineering, 91*, 109519. https://doi.org/10.1016/j.jobe.2024.109519

[68] Chu, W. X., Lien, Y. H., Huang, K. R., et al. (2021). Energy saving of fans in air-cooled server via deep reinforcement learning algorithm. *Energy Reports, 7*, 3437–3448. https://doi.org/10.1016/j.egyr.2021.06.003

[69] Wan, J., Zhou, J., & Gui, X. (2021). Intelligent rack-level cooling management in data centers with active ventilation tiles: A deep reinforcement learning approach. *IEEE Intelligent Systems, 36*(6), 42–52. https://doi.org/10.1109/MIS.2021.3049865

[70] Wan, J., Duan, Y., Gui, X., et al. (2023). SafeCool: Safe and energy-efficient cooling management in data centers with model-based reinforcement learning. *IEEE Transactions on Emerging Topics in Computational Intelligence, 7*(6), 1621–1635. https://doi.org/10.1109/TETCI.2023.3234545

[71] Shaw, R., Howley, E., & Barrett, E. (2017). An advanced reinforcement learning approach for energy-aware virtual machine consolidation in cloud data centers. In *Proceedings of the 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)* (pp. 61–66). IEEE. https://doi.org/10.23919/ICITST.2017.8356347

[72] Farahnakian, F., Liljeberg, P., & Plosila, J. (2014). Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning. In *Proceedings of the 2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing* (pp. 500–507). IEEE. https://doi.org/10.1109/PDP.2014.109

[73] Yan, L., Liu, W., Jiang, W., et al. (2021). Deep reinforcement learning based optimization of battery charging and discharging management for data center. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–9). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533476

[74] Asgari, S., Mirhoseininejad, S. M., Moazamigoodarzi, H., et al. (2021). A gray-box model for real-time transient temperature predictions in data centers. *Applied Thermal Engineering, 185*, 116319. https://doi.org/10.1016/j.applthermaleng.2020.116319

[75] Mirhoseininejad, S. M., García, F. M., Badawy, G., et al. (2019). ALTM: Adaptive learning-based thermal model for temperature predictions in data centers. In *Proceedings of the 2019 IEEE Sustainability through ICT Summit (StICT)* (pp. 1–6). IEEE. https://doi.org/10.1109/STICT.2019.8789370

[76] Alkharabsheh, S., Sammakia, B., Shrivastava, S., et al. (2013). Utilizing practical fan curves in CFD modeling of a data center. In *Proceedings of the 29th IEEE Semiconductor Thermal Measurement and Management Symposium* (pp. 211–215). IEEE. https://doi.org/10.1109/SEMI-THERM.2013.6526831

[77] Simon, V. S., Siddarth, A., & Agonafer, D. (2020). Artificial neural network based prediction of control strategies for multiple air-cooling units in a raised-floor data center. In *Proceedings of the 2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* (pp. 334–340). IEEE. https://doi.org/10.1109/ITherm45881.2020.9190431

[78] De Lorenzi, F., & Vömel, C. (2012). Neural network-based prediction and control of air flow in a data center. *Journal of Heat Transfer, 134*(7). https://doi.org/10.1115/1.4005605

[79] Choi, Y. J., Park, B. R., Hyun, J. Y., et al. (2022). Development of an adaptive artificial neural network model and optimal control algorithm for a data center cyber-physical system. *Building and Environment, 210*, 108704. https://doi.org/10.1016/j.buildenv.2021.108704

[80] Huang, J., Zhang, Z., Yang, X., et al. (2023). Data-driven adaptive control of CRAC in data center based on online incremental RVFL. In *Proceedings of the 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)* (pp. 962–967). IEEE. https://doi.org/10.1109/DDCLS58216.2023.10166712

[81] Pao, Y. H., & Takefuji, Y. (1992). Functional-link net computing: Theory, system architecture, and functionalities. *Computer, 25*(5), 76–79. https://doi.org/10.1109/2.144401

[82] Mohsenian, G., Khalili, S., Tradat, M., et al. (2021). A novel integrated fuzzy control system toward automated local airflow management in data centers. *Control Engineering Practice, 112*, 104833. https://doi.org/10.1016/j.conengprac.2021.104833

[83] Berezovskaya, Y., Yang, C. W., & Vyatkin, V. (2020). Towards multi-agent control in energy-efficient data centres. In *Proceedings of the IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society* (pp. 3574–3579). IEEE. https://doi.org/10.1109/IECON43393.2020.9255232

[84] Berezovskaya, Y., Yang, C. W., & Vyatkin, V. (2021). Towards reinforcement learning approach to energy-efficient control of server fans in data centres. In *Proceedings of the 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 1–4). IEEE. https://doi.org/10.1109/ETFA45728.2021.9613639

[85] Brännvall, R., Sarkinen, J., Svartholm, J., et al. (2019). Digital twin for tuning of server fan controllers. In *Proceedings of the 2019 IEEE 17th International Conference on Industrial Informatics (INDIN)* (pp. 1425–1428). IEEE. https://doi.org/10.1109/INDIN41052.2019.8972291

[86] Wang, Z., Bash, C., Tolia, N., et al. (2009). Optimal fan speed control for thermal management of servers. In *Proceedings of the International Electronic Packaging Technical Conference and Exhibition* (pp. 709–719). ASME. https://doi.org/10.1115/InterPACK2009-89074

[87] Zapater, M., Risco-Martín, J. L., Grosso, P., et al. (2013). Temperature-aware server control for improving energy efficiency in data centers. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 266–269). IEEE. https://doi.org/10.7873/DATE.2013.067

[88] Lucchese, R., & Johansson, A. (2019). On energy efficient flow provisioning in air-cooled data servers. *Control Engineering Practice, 89*, 103–112. https://doi.org/10.1016/j.conengprac.2019.05.019

[89] Eiland, R., Fernandes, J. E., Nagendran, B., et al. (2017). Effectiveness of rack-level fans—Part I: Energy savings through consolidation. *Journal of Electronic Packaging, 139*(4), 041011. https://doi.org/10.1115/1.4038235

[90] Fernandes, J. E., Eiland, R., Nagendran, B., et al. (2017). Effectiveness of rack-level fans—Part II: Control

strategies and system redundancy. *Journal of Electronic Packaging, 139*(4), 041012. https://doi.org/10.1115/1.4038014

[91] Li, G., Li, M., Azarm, S., et al. (2007). Optimizing thermal design of data center cabinets with a new multi-objective genetic algorithm. *Distributed and Parallel Databases, 21*, 167–192. https://doi.org/10.1007/s10619-007-7009-9

[92] Nadjahi, C., Louahlia, H., & Lemasson, S. (2018). A review of thermal management and innovative cooling strategies for data center. *Sustainable Computing: Informatics and Systems, 19*, 14–28. https://doi.org/10.1016/j.suscom.2018.05.002

[93] Moss, D. L. (2011). *Data center operating temperature: The sweet spot* (White Paper). Dell Inc./Data Center Infrastructure.

[94] Chen, J., Tan, R., Xing, G., et al. (2014). PTEC: A system for predictive thermal and energy control in data centers. In *Proceedings of the 2014 IEEE Real-Time Systems Symposium* (pp. 218–227). IEEE. https://doi.org/10.1109/RTSS.2014.27

[95] Ayoub, R., Nath, R., & Rosing, T. (2012). JETC: Joint energy thermal and cooling management for memory and CPU subsystems in servers. In *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture* (pp. 1–12). IEEE. https://doi.org/10.1109/HPCA.2012.6169035

[96] Gao, P., Liu, H., Luo, H., et al. (2024). Discussion on the technical path of data center information and communication thermal management. *Energy Reports, 11*, 2704–2714. https://doi.org/10.1016/j.egyr.2024.02.003

[97] Bai, F., Tang, Z., Yin, R. J., et al. (2024). A novel '3D+ digital twin+ 3D' upscaling strategy for predicting the detailed multi-physics distributions in a commercial-size proton exchange membrane fuel cell stack. *Applied Energy, 374*, 124012. https://doi.org/10.1016/j.apenergy.2024.124012

[98] Xu, W., Zhao, B., Zeng, Y., et al. (2023). Intelligent data center safety status prediction based on algorithm ensemble. In *Proceedings of the 2023 International Conference on Mobile Internet, Cloud Computing and Information Security (MICCIS)* (pp. 63–68). IEEE. https://doi.org/10.1109/MICCIS58901.2023.00016

[99] Yao, J., Guan, H., & Luo, J. (2014). Adaptive power management through thermal aware workload balancing in internet data centers. *IEEE Transactions on Parallel and Distributed Systems, 26*(9), 2400–2409. https://doi.org/10.1109/TPDS.2014.2353051

[100] Song, Z., Murray, B. T., & Sammakia, B. (2013). Airflow and temperature distribution optimization in data centers using artificial neural networks. *International Journal of Heat and Mass Transfer, 64*, 80–90. https://doi.org/10.1016/j.ijheatmasstransfer.2013.04.017

[101] Zapater, M., Risco-Martín, J. L., Arroba, P., et al. (2016). Runtime data center temperature prediction using grammatical evolution techniques. *Applied Soft Computing, 49*, 94–107. https://doi.org/10.1016/j.asoc.2016.07.041

[102] Simin, W., Yifei, K., Yixuan, X., et al. (2024). Data center temperature prediction and management based on a two-stage self-healing model. *Simulation Modelling Practice and Theory, 132*, 102883. https://doi.org/10.1016/j.simpat.2023.102883

[103] Liu, N., Lin, X., & Wang, Y. (2017). Data center power management for regulation service using neural network-based power prediction. In *Proceedings of the 2017 18th International Symposium on Quality Electronic Design (ISQED)* (pp. 367–372). IEEE. https://doi.org/10.1109/ISQED.2017.7918343

[104] Song, Z., Murray, B. T., & Sammakia, B. (2014). A dynamic compact thermal model for data center analysis and control using the zonal method and artificial neural networks. *Applied Thermal Engineering, 62*(1), 48–57. https://doi.org/10.1016/j.applthermaleng.2013.09.006

[105] Jin, S. Q., Li, N., Bai, F., et al. (2023). Data-driven model reduction for fast temperature prediction in a multi-variable data center. *International Communications in Heat and Mass Transfer, 142*, 106645. https://doi.org/10.1016/j.icheatmasstransfer.2023.106645

[106] Wang, R., Zhou, X., Dong, L., et al. (2020). Kalibre: Knowledge-based neural surrogate model calibration for data center digital twins. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (pp. 200–209). ACM. https://doi.org/10.1145/3408308.3427982

[107] Zhu, H., & Lin, B. (2024). Digital twin-driven energy consumption management of integrated heat pipe

cooling system for a data center. *Applied Energy, 373*, 123840. https://doi.org/10.1016/j.apenergy.2024.123840

[108] Liu, Y., Wei, X., Xiao, J., et al. (2020). Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers. *Global Energy Interconnection, 3*(3), 272–282. https://doi.org/10.1016/j.gloei.2020.07.008

[109] He, Z., Xi, H., Ding, T., et al. (2021). Energy efficiency optimization of an integrated heat pipe cooling system in data center based on genetic algorithm. *Applied Thermal Engineering, 182*, 115800. https://doi.org/10.1016/j.applthermaleng.2020.115800

[110] Jia, C., Wang, H., & Wei, L. (2016). Study of smart transportation data center virtualization based on vmware vsphere and parallel continuous query algorithm over massive data streams. *Procedia Engineering, 137*, 719–728. https://doi.org/10.1016/j.proeng.2016.01.309

[111] Hsieh, S. Y., Liu, C. S., Buyya, R., et al. (2020). Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers. *Journal of Parallel and Distributed Computing, 139*, 99–109. https://doi.org/10.1016/j.jpdc.2019.12.014

[112] Pakbaznia, E., Ghasemazar, M., & Pedram, M. (2010). Temperature-aware dynamic resource provisioning in a power-optimized datacenter. In *Proceedings of the 2010 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 124–129). IEEE. https://doi.org/10.1109/DATE.2010.5457223

[113] Van Damme, T., De Persis, C., & Tesi, P. (2018). Optimized thermal-aware job scheduling and control of data centers. *IEEE Transactions on Control Systems Technology, 27*(2), 760–771. https://doi.org/10.1109/TCST.2017.2783366

[114] Zhou, Z., Abawajy, J., Chowdhury, M., et al. (2018). Minimizing SLA violation and power consumption in cloud data centers using adaptive energy-aware algorithms. *Future Generation Computer Systems, 86*, 836–850. https://doi.org/10.1016/j.future.2017.07.048

[115] Banerjee, A., Mukherjee, T., Varsamopoulos, G., et al. (2011). Integrating cooling awareness with thermal aware workload placement for HPC data centers. *Sustainable Computing: Informatics and Systems, 1*(2), 134–150. https://doi.org/10.1016/j.suscom.2011.02.003

[116] Tang, Q., Gupta, S. K. S., & Varsamopoulos, G. (2008). Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach. *IEEE Transactions on Parallel and Distributed Systems, 19*(11), 1458–1472. https://doi.org/10.1109/TPDS.2008.111

[117] Lei, H., Wang, R., Zhang, T., et al. (2016). A multi-objective co-evolutionary algorithm for energy-efficient scheduling on a green data center. *Computers & Operations Research, 75*, 103–117. https://doi.org/10.1016/j.cor.2016.05.014