OPEN ACCESS

REVIEW

Journal homepage: www.the-innovation.org/energy-use

Energy Use

Energy Use 2 (2025):100020

Coutents lists available at INNOVATION PRESS

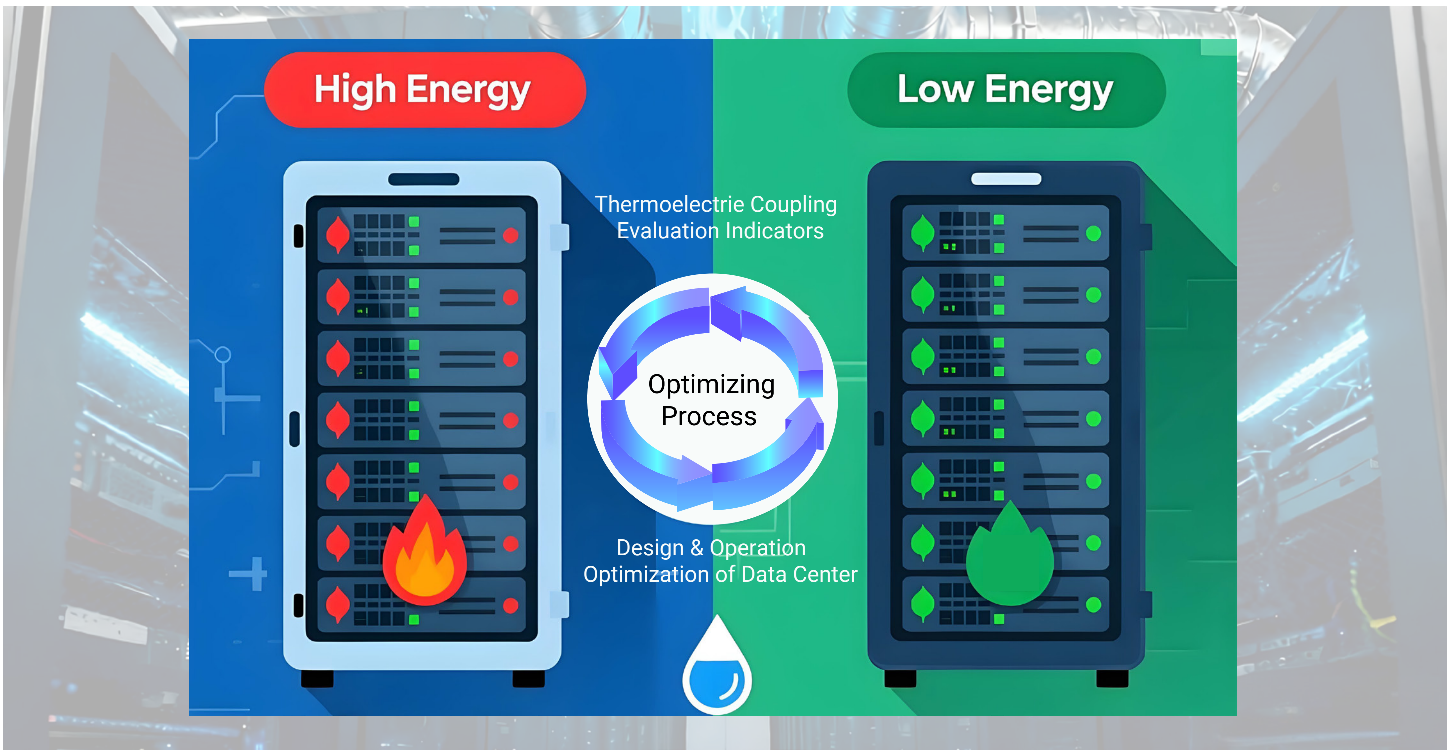
Joint IT-Cooling System Optimization for Data Center Energy Efficiency: A Critical Review

Kai Hu,[1] Puxi Guo,[1] Chengchu Yan,[1,*] Tao Deng,[2] Depeng Li,[2] and Chaoqun Zhuang[1]

*Correspondence: chengchu.yan@njtech.edu

Received: October 24, 2025; Accepted: December 3, 2025; Published Online: December 24, 2025; https://doi.org/10.59717/ipj.energy-use.2025.100020

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Focuses on thermoelectric coupling for holistic data center energy-efficiency optimization.
- Analyzes limitations of PUE and its derived indicators for energy saving.
- Reviews optimization schemes for IT and cooling systems at both design and operational stages.
- Proposes future research on "computational efficiency" and cross-scale thermoelectric models.

Journal homepage: www.the-innovation.org/energy-use
Energy Use
INNOVATION PRESS
Energy Use 2 (2025):100020
Coutents lists available at INNOVATION PRESS

Joint IT-Cooling System Optimization for Data Center Energy Efficiency: A Critical Review

Kai Hu,[1] Puxi Guo,[1] Chengchu Yan,[1,*] Tao Deng,[2] Depeng Li,[2] and Chaoqun Zhuang[1]
[1]College of Urban Construction, Nanjing Tech University, No.200, North Zhongshan Road, Nanjing, 210009, China
[2]China Construction Industrial & Energy Engineering Group Co., Ltd., No.6, Wenlan Road, Qixia District, Nanjing, Jiangsu, China
*Correspondence: chengchu.yan@njtech.edu
Received: October 24, 2025; Accepted: December 3, 2025; Published Online: December 24, 2025; https://doi.org/10.59717/ipj.energy-use.2025.100020

Citation: Hu K., Guo P., Yan C., et al. (2025). Joint IT-Cooling System Optimization for Data Center Energy Efficiency: A Critical Review. *Energy Use* **1**:100020.

Data center energy efficiency requires collaborative optimization of IT and cooling systems. However, comprehensive reviews that provide whole-process optimization guidance based on thermoelectric coupling are still scarce. This paper bridges this gap by analyzing the thermoelectric coupling characteristics of data centers and systematically reviewing evaluation indicators, design measures, and control methods for energy saving optimization, with the goal of providing theoretical guidance for holistic energy-efficiency improvement.Specifically, to determine the objective of collaborative energy saving optimization, this paper focuses on analyzing the limitations of PUE and its derived indicators. Secondly, to facilitate collaborative energy saving optimization at the design stage, this paper reviews the optimization schemes and measures for IT and cooling systems and highlights the key design aspects of cooling systems. Afterwards, to promote collaborative energy saving optimization at the operational stage, this paper reviews research into individual optimization as well as joint optimization between IT and cooling systems and points out the core problems of existing joint optimization methods. On this basis, this paper also proposes two future research directions in the field: the evaluation indicators of "computational efficiency" and a universal cross-scale thermoelectric coupling model.

INTRODUCTION

In recent years, the internet, big data, and artificial intelligence (AI) services have gradually infiltrated various industries. As the infrastructure supporting these services, data centers have played a crucial role. However, at the same time, their energy consumption issues have become increasingly prominent. It is projected that by 2030, global data centers will account for 7% of total electricity consumption[1] and generate 1.167 billion tons of carbon dioxide emissions.[2] Compared to the global scene, China's data center industry started relatively late, but it has witnessed extremely rapid development. By 2023, the number of data center racks nationwide had reached 7.6 million.[3] It is estimated that by 2025, China's data centers will consume approximately 120 billion kilowatt-hours of electricity annually and emit 100 million tons of carbon dioxide.[4] To facilitate the steady advancement of the dual-carbon (carbon peak and carbon neutrality) policy, energy conservation and emission reduction in data centers have become urgent priorities.

In terms of the energy consumption composition of data centers, IT systems and cooling systems consume the most electricity. They account for approximately 60% and 30% of the total energy consumption, respectively.[5,6] Many researchers have conducted extensive research on the individual energy saving optimization of both, including energy saving scheduling of IT systems, energy saving design and efficient control of cooling systems. However, in actual operation, data centers still face issues such as persistently high overall energy consumption, frequent local hotspots, severe overcooling, and insufficient utilization of natural cold sources, resulting in significant energy waste. A crucial reason for this lies in the insufficient collaboration between the energy saving efforts of IT and cooling systems.

Currently, the academic community has gradually recognized that energy saving in data centers is a work involving the collaborative optimization of IT and cooling systems. Several review papers have explored this topic from multiple perspectives, each providing valuable insights into different aspects of the field. For instance, Wang et al.[7] starts from the power characteristics of servers, proposes a collaborative concept of computing-electric-thermal in data centers, and focuses on reviewing server power and computing efficiency models, two-phase liquid cooling technology, and dynamic control technology. Lin and Cheng et al.[8,9] introduce various thermal modeling methods for air-cooled and liquid-cooled data centers and provide a comprehensive review of operational optimization research from three perspectives: IT load scheduling considering temperature (thermally aware scheduling), cooling system control optimization, and joint optimization of IT and cooling systems. Cao et al.[10] respectively reviews the joint operational optimization of servers and air conditioners, the joint optimization of IT systems and server fans/air conditioners, as well as the joint optimization of server fans/air conditioners and cooling systems.

Although the aforementioned review papers have made significant contributions from the perspectives of modeling and optimization algorithms,[11] they have not delved deeply into the coupled relationship between IT and cooling systems, nor have these papers analyzed the impact of this factor on the collaborative energy saving optimization of the two. Data centers exhibit pronounced thermoelectric coupling characteristics, meaning that server power is positively correlated with temperature, and temperature is a result of the combined effects of server heat dissipation and cooling system operation. However, current energy saving designs and the energy-efficient operation of the overall system have not fully taken this into account (i.e., there is a lack of comprehensive guidance for energy saving optimization in data centers based on thermoelectric coupling characteristics), making it difficult to achieve holistic energy efficiency improvements in both IT and cooling systems. As the amount and methods of server heat dissipation change, along with the gradual tightening of various policy and regulatory requirements, the architectural forms of cooling systems are also diversifying and differentiating, rendering the aforementioned collaboration particularly complex. Consequently, energy saving in data centers is gradually reaching a bottleneck.

To bridge this gap, this review reframes the problem through a thermoelectric coupling lens. It specifically focuses on the problems within existing evaluation indicators, design optimization measures, and operational optimization strategies, and proposes improvement suggestions. The aim is to

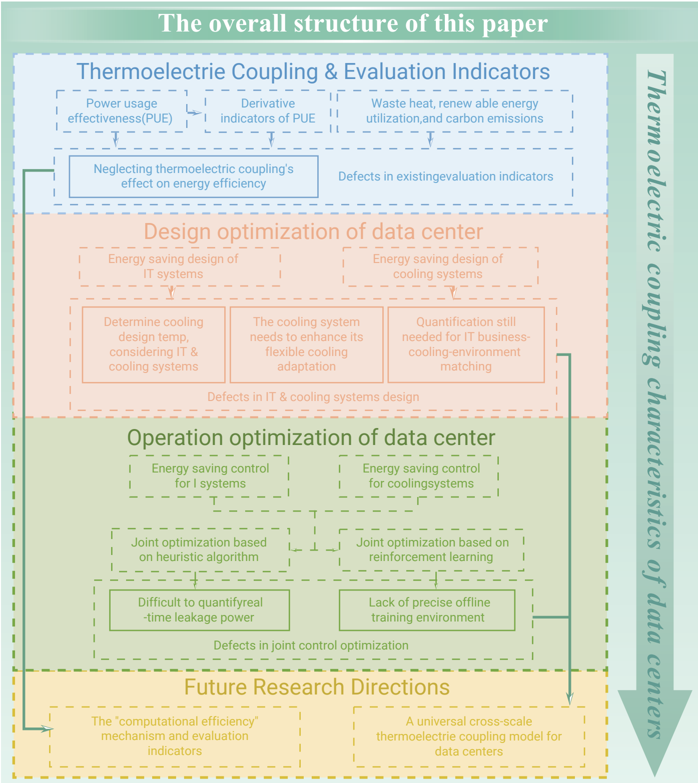


Figure 1. The overall structure of this paper

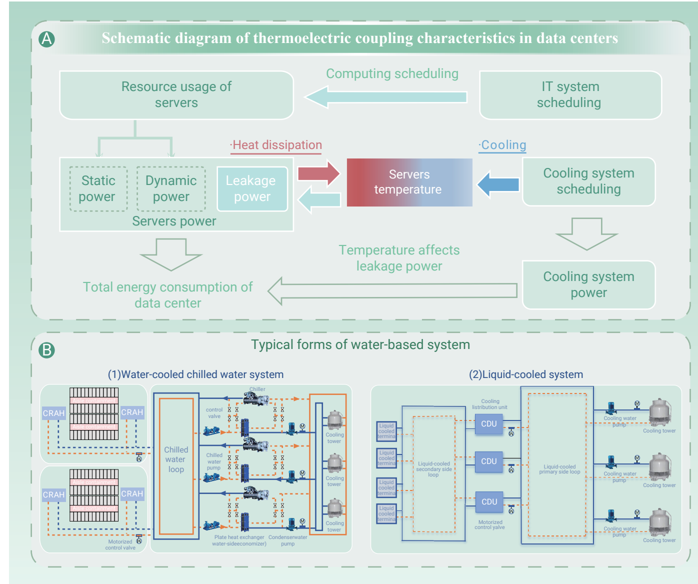


Figure 2. Schematic diagram (A) Thermoelectric coupling characteristics in data centers. (B) Typical forms of water-based system: (1) Water-cooled chilled water system. (2) Liquid-cooled system.

provide crucial directions and approaches for the collaborative energy saving optimization of IT and cooling systems, thereby achieving the overall optimal energy efficiency in data centers.

The overall structure of this paper is shown in Figure 1. Specifically, it reviews commonly used energy efficiency evaluation indicators in data centers, expounds on the issues with the currently most widely used Power Usage Effectiveness (PUE) indicator and its derivative indicators (Section 2). Secondly, it reviews the design optimization of IT and cooling systems, with a focus on elaborating the optimization measures for water-based, fluorine-based, and air-based cooling systems from the perspective of cooling system types. It elaborates on the problems in the design cooling temperatures of current systems and points out the research gaps in two other aspects. Afterwards, it reviews the individual and joint optimization strategies between IT and cooling systems, separately expounding on the research gaps in joint optimization methods based on optimal temperature and reinforcement learning (Section 4). Building on this, it identifies two key future research directions in the field of collaborative energy saving optimization (Section 5). Finally, it summarizes the main contributions of the entire paper (Section 6).

THERMOELECTRIC COUPLING & EVALUATION INDICATORS

Thermoelectric coupling characteristics of data centers

Servers are the primary subjects of services in data centers. Their power can be broken down into three components: static power, dynamic power, and leakage power.[12] Static power is relatively stable; dynamic power is positively correlated with resource utilization; and leakage power, caused by leakage currents in semiconductor devices, is positively correlated with temperature.[7] Temperature, in turn, is the result of the combined effects of heat dissipation from servers and cooling provided by the cooling system. Therefore, there exists a complex coupling relationship among "server power - internal and external temperature fields of servers - cooling system power, which can also be referred to as the thermoelectric coupling characteristics of data centers (as shown in Figure 2A).

The thermoelectric coupling characteristics result in a "seesaw effect" in energy consumption between the IT system and the cooling system, where an increase in one leads to a decrease in the other. Moreover, due to the significant differences in scheduling speed and accuracy between the two systems, collaborative energy saving optimization becomes particularly complex.

Critical analysis of existing indicators

Evaluation indicators are the critical objectives for energy saving optimization in data centers. This section offers a review of commonly used indicators and analyzes their existing issues.

Power usage effectiveness (PUE). In terms of the overall energy efficiency of data center infrastructure, the most commonly used evaluation indicator is the PUE,[13] which calculates the ratio of the total energy consumption of a data center to the energy consumption of its IT equipment (as shown in Equation 1).

$$PUE = \frac{E_{DC}}{E_{IT}} \quad (1)$$

In the above content, E_{DC} is the total energy consumption of the data center; E_{IT} is the energy consumption of IT equipment.

Currently, the PUE indicator is widely applied in research on the design and operation of data center infrastructure. Governments and regions worldwide have also introduced numerous energy conservation policies based on this indicator. For instance, in 2018, the United States took the lead in issuing the *Data Center Optimization Initiative*. The initiative mandates that the energy consumption indicator PUE for data centers should be ≤1.5 (≤1.4 for newly built data centers).[14] In 2022, Singapore stipulated that the PUE for newly constructed data centers should be below 1.3. The Netherlands, on the other hand, imposed even stricter requirements, demanding that the PUE value for newly built data centers be below 1.2.

In China, since the introduction of the dual-carbon policy in 2020, various ministries and commissions have issued policies targeting energy consumption in data centers. *The Implementation Plan for Promoting the Green and High-Quality Development of New Infrastructure Such as Data Centers and 5G in Line with the Requirements for Achieving Carbon Peak and Carbon Neutrality Goals,*[15] released in November 2021, proposed that "by 2025, the nationwide average PUE of newly built large and super-large data centers should be reduced to below 1.3, with further reductions to below 1.25 for national hub nodes." Additionally, the *Code for Data Center Projects* issued by the Ministry of Housing and Urban-Rural Development in April 2024 requires that the annual average PUE of newly built and renovated large and super-large data centers should be less than 1.3. Although the PUE indicator measures the energy utilization efficiency of data center infrastructure, given the significant proportion of energy consumption attributed to cooling systems within data center infrastructure,[16] PUE is also commonly used as an approximate indicator to evaluate the energy efficiency of cooling systems.

Derivative indicators of PUE. Since the PUE indicator measures the energy utilization efficiency of data center infrastructure, to further specifically evaluate the energy efficiency of cooling/power systems, researchers have proposed derivative evaluation indicators based on PUE, including Data Center Infrastructure Efficiency (DCiE), Cooling Load Factor (CLF), Power Load Factor (PLF),[17] Mechanical Load Component (MLC),[18] Partial Power Usage Effectiveness (pPUE),[19] Cooling Efficiency Ratio (CER),[20] and Gross Coefficient of Performance of the cooling system (GCOP).[21] Among them, Data Center Infrastructure Efficiency (DCiE) is also a typical comprehensive energy efficiency indicator, which has a reciprocal relationship with PUE, as shown in Equation 2.

$$\mathrm{DCiE} = \frac{E_{IT}}{E_{DC}} = 1/\mathrm{PUE} \tag{2}$$

Considering that PUE and DCiE cannot accurately evaluate the performance of the cooling system in data centers, researchers have decomposed the PUE indicator into the CLF and the PLF. The calculation methods for CLF and PLF are shown in Equations 4 and 5, respectively. Compared with PUE, CLF can be used to measure the energy consumption level of the cooling system in data centers. A similar evaluation indicator, namely the MLC, has also been proposed in the reference[18]. This indicator is defined as the ratio of the energy consumption of the cooling system system (including all cooling-related equipment) to the energy consumption of IT equipment, with the calculation method shown in Equation 6.

$$\mathrm{PUE} = 1 + \mathrm{CLF} + \mathrm{PLF} \tag{3}$$

$$\mathrm{CLF} = \frac{E_{cooling}}{E_{IT}} \tag{4}$$

$$\mathrm{PLF} = \frac{E_{power}}{E_{IT}} \tag{5}$$

$$MLC = \frac{P_{HVAC}}{P_{IT}} \tag{6}$$

In the above content, E_{power} is the energy consumption of power system; P_{HVAC} is the energy consumption of cooling system system.

Information technology-Data centres-Key performance indicators-Part 2: Power usage effectiveness (PUE)[19] proposed the pPUE, as shown in Equation 7. It can be seen that the value of pPUE is also closely related to the energy consumption of IT equipment. Therefore, when used for evaluating the energy efficiency of the cooling system, there is no essential difference between pPUE and CLF. China Association of Refrigeration[20] introduced the CER indicator, which is defined as the ratio of the total heat rejected by the cooling system throughout the year to its annual energy consumption, as shown in Equation 8. Tian et al.[21] presented the GCOP for data center cooling systems, with its calculation process illustrated in Equation 9.

$$\mathrm{pPUE}_{sub} = \frac{E_{sub} + E_{IT}}{E_{IT}} \tag{7}$$

$$\mathrm{CER} = \frac{Q_{removed}}{E_{cooling}} \tag{8}$$

$$\mathrm{GCOP} = \frac{E_{DC} - E_{cooling}}{E_{cooling}} \tag{9}$$

In the above content, E_{sub} is the energy consumption of a subsystem; $Q_{removed}$ is the total annual heat rejected of the data center, including IT equipment, power system, and cooling system.

Other indicators. The aforementioned indicators are only applicable to evaluating the energy efficiency of data centers. To further measure the waste heat utilization in data centers, the Energy Reuse Rate (ERE) was proposed in Reference.[17] The ERE is primarily used to quantify the impact of waste heat utilization on the energy efficiency of infrastructure, with its calculation formula shown as Equation 10. Theoretically, when ERE reaches the optimal value of 1, it indicates that the waste heat rejected by the data center has been fully recovered and utilized. Currently, there are no explicit quantitative requirements for waste heat utilization in China's policies and regulations.

$$ERE = (1 - ERF) \times PUE \tag{10}$$

In the above content, the Energy Reuse Factor (ERF) is used to characterize the proportion of the recyclable part within the total input energy. If ERF = 0, it indicates that no waste heat is being utilized in the data center.

The Renewable Energy Ratio (RER) is used to evaluate the utilization of renewable energy/green electricity in data centers,[22] with its calculation method shown in Equation 11. Currently, only certain regions in China have established quantitative requirements for RER. For example, Beijing requires that for data center projects completed in 2021 and subsequent years, the proportion of annual renewable energy utilization to annual energy consumption should increase by 10% annually, reaching 100% by 2030. Hebei Province announced in 2022 that it would prioritize supporting the construction and expansion of data centers with a renewable energy utilization rate of 50% or higher. In 2024, Shenzhen mandates that the green electricity utilization rate of newly-built data centers must exceed 50%. Ningxia has relatively stricter requirements, stipulating that by 2025, the renewable energy utilization rate within the region should reach 60%, while that within the Zhongwei Data Center Cluster should reach 65%.

$$RER = \frac{E_R}{E_{DC}} \tag{11}$$

In the above content, E_R is the usage of renewable energy.

To further measure the carbon usage efficiency of data centers, the Carbon Usage Effectiveness (CUE) indicator has gradually drawn attention from the government and the academic community, with its calculation principle shown in Equation 12. Reference[17] further established a correlation between CUE and PUE (as demonstrated in Equation 14). Currently, there are no explicit quantitative requirements for the CUE indicator in China's policies and regulations.

$$CUE = \frac{C_e}{E_{IT}} \tag{12}$$

$$C_e = \sum_{1}^{8760} P_{total} \times \mu \tag{13}$$

$$CUE = CEF \times PUE \tag{14}$$

In the above content, C_e is the carbon dioxide emissions from the data center; μ is the carbon emission factor of primary energy; P_{total} is the total hourly power consumption of the data center; CEF is the carbon emissions generated for every 1kWh of electricity consumed.

Analysis on the Applicability of Indicators. Table 1 provides an overview of the types and purposes of the aforementioned evaluation indicators. It can be observed that, to evaluate the energy efficiency of data center infrastructure, indicators such as PUE and DCiE (which are reciprocally related) can be employed. Similarly, for assessing the carbon emissions of data center infrastructure, the CUE indicator can be used. For evaluating the energy efficiency of data center cooling systems, indicators including CLF, MLC, CER, and GCOP can be adopted (these four indicators share similar calculation principles without substantial differences). To assess the energy efficiency of data center power systems, the PLF indicator can be utilized. Moreover, the pPUE indicator is applicable to the energy efficiency evaluation of both power systems and cooling systems. Finally, indicators ERE and ERF can be used in combination to evaluate the renewable energy utilization rate of data centers.

Defects in existing evaluation indicators

At present, PUE and its derivative indicators (such as DCiE, CLF, etc.) are widely applied to guide the design and operation of data center cooling systems. However, all the aforementioned indicators have significant flaws because they overlook the "seesaw effect" of energy consumption between the IT and cooling systems caused by the thermoelectric coupling characteristics of data centers (details can be found in Section 2.1). This issue has greatly misled the current design and operation of data centers, that is, it has

Table 1. Summary of energy efficiency evaluation indicators.

Reference	Indicator	Indicator type	Evaluation object and main purpose
[13]	PUE	Infrastructure energy efficiency	It evaluates data center infrastructure efficiency and is often roughly used for cooling system efficiency assessment.
[17]	DCiE	Infrastructure energy efficiency	It is used to assess the energy efficiency of data center infrastructure, inversely related to PUE.
[17]	CLF	Cooling efficiency	It is used to evaluate the energy efficiency of data center cooling systems.
[17]	PLF	Power efficiency	It is used to assess the energy efficiency of the power system in data centers.
[18]	MLC	Cooling efficiency	It is used to evaluate the energy efficiency of the cooling system system within the cooling system.
[19]	pPUE	Subsystem efficiency	It is used to evaluate the energy efficiency of each subsystem in data centers.
[20]	CER	Cooling efficiency	It is used to evaluate the energy efficiency of data center cooling systems and is similar to CLF.
[21]	GCOP	Cooling efficiency	It is used to evaluate the energy efficiency of data center cooling systems and is similar to CLF.
[17]	ERE	Waste heat utilization efficiency	It is used to assess the impact of waste heat utilization on the energy efficiency of infrastructure.
[17]	ERF	Waste heat utilization efficiency	It is used to evaluate the waste heat recovery ratio in data centers.
[22]	RER	Renewable energy utilization efficiency	It is used to assess the proportion of green electricity in the electricity consumption of data centers.
[17]	CUE	Carbon emission	It is used to evaluate the operational carbon emissions of data centers.

led designers and operators to blindly pursue a reduction in PUE while neglecting the actual changes in the energy consumption of data centers. In practical engineering projects, to reduce PUE, designers and operation and maintenance personnel usually adopt methods such as increasing the inlet air/liquid temperature of servers to reduce the energy consumption of the cooling systems (and there may even be a phenomenon of over-implementing these measures). However, this measure will inevitably lead to an increase in server leakage power, to the extent that PUE may decrease, but the overall energy consumption of the data center may actually increase. Therefore, it is necessary to further study evaluation indicators that can directly guide the collaborative energy saving optimization of IT and cooling systems in data centers (this is also discussed in Section 5.1 of this paper).

DESIGN OPTIMIZATION OF DATA CENTER

Energy saving design serves as the foundation for energy saving control. This section first provides an overview of the energy saving design optimization for IT systems from two aspects: servers and switch-transmission networks. Subsequently, starting from the types of cooling systems, it offers a detailed review of the design optimization schemes and measures for systems such as water-based, fluorine-based, and air-based cooling systems. On this basis, it discusses three existing issues/gaps in current design optimization.

Energy saving design of IT systems

Energy saving design of servers. The server stands as the primary energy-consuming component within the IT system, with the processor serving as its core energy-draining element. There have been quite a few studies focusing on the embedded architectures, design methodologies, and optimization techniques for low-power processors.[23-27] For instance, Bartolini et al.[26] compared two basic layouts (single-sided and double-sided) of systems based on non-uniform cache architecture, aiming to explore the optimal combination of on-chip data management strategies and cache core layouts. The study evaluated the effectiveness of static and dynamic non-uniform cache architectures. The results revealed that, through a balance of hardware and software optimizations, the single-sided layout achieved the best performance and the lowest power. Momeni et al.[27] proposed an approximate network-on-chip architecture to address the high power consumption issue faced by three-dimensional networks-on-chip. This architecture implemented approximate logic in the network interface. The simulation results demonstrated that, compared with previous research work, the method in this study reduced energy consumption by 21.6%, decreased latency by 22%, and had a hardware overhead of only about 1.3%.

The energy saving design of heterogeneous computing systems is also a key focus.[28-30] Huang et al.[28] introduced a novel heterogeneous processing - in - memory hardware called Hegraph (which integrates a memristor based analog computing unit and a CMOS based digital computing core on the same logic layer of a 3D chip stacked storage device) to promote energy - efficient graphics processing. Karpowicz et al.[29] investigated the design of energy-efficient, application-specific CPU frequency governors in heterogeneous distributed computing systems. Its aim was to assign different tasks to servers with varying characteristics, thereby enhancing the operational efficiency of the system. The results showed that the controller designed in this study outperformed the standard general-purpose governor in the Linux kernel in terms of accessible server performance and energy saving capabilities.

Energy saving design of switches and transmission networks. Network devices typically account for 10% -20% of IT system energy consumption. Widjaja et al.[31] proposed a method for optimizing switch sizing at the deployment design stage of data center networks, focusing on how to select an appropriate switch scale to achieve maximum energy savings during the network's operational period. The results indicate that when traffic demand is relatively low or when servers exchanging traffic are in close proximity, deploying a large number of small switches is more energy-efficient than deploying a small number of large switches. Li et al.[32] proposed a data center network solution based on optical switching technology and network technology, which effectively enhanced data transmission efficiency. Wang et al.[33] proposed a comprehensive overview of the network topology architectures in data centers, classifying them into hierarchical models, recursive models, or rack-to-rack models. Building on this, it discussed their individual impacts on cloud computing performance, with the goal of providing guidance for the design of network infrastructure. Xiao et al.[34] proposed an energy-efficient encryption acceleration system for HTTPS based on OpenSSL. This system aimed for energy-efficient encryption via hardware-software co-design, enhancing both web server energy efficiency and system resource utilization. Results showed that, using only AES-256-CBC, compared to original software encryption, instruction set acceleration, and hardware accelerators, it boosted performance per watt by 1637.13%, 84.82%, and 966.23% respectively.

Energy saving design of cooling systems

The cooling system provides cooling services for IT equipment, yet it also has a significant impact on the energy consumption of IT equipment. Therefore, a rational design of the cooling system serves as a crucial foundation for the collaborative energy saving optimization of the IT system and the cooling system. Given that most cooling systems in current data centers adopt water-based systems, fluorine-based systems and air-based systems, this section mainly provides an overview of the design optimization of these three types of systems and their derivative systems, and analyze the existing problems.

Water-based systems

The water-based system refers to a cooling system that utilizes water as the secondary refrigerant. Its main types encompass water-cooled chilled water systems and liquid-cooled systems, with the basic configurations of both illustrated in Figure 2B.

Water-cooled chilled water systems. The water-cooled chilled water system is a commonly adopted cooling solution for large/ultra-large data centers (as shown in Figure 2B(1)). It consists of equipment such as chillers, cooling towers, water pumps, cold storage tanks, and air-cooled terminals (including room-level air conditioners, row-level air conditioners, and rack-level air conditioners). Numerous studies have been conducted to optimize its system architecture and equipment selection. Some people[35] investigated the piping in the chiller plant and water pump selection for water-cooled chilled water systems. Based on a steady-state model of an actual data center cooling system, the studies compared the energy efficiency performance of five different chiller plant designs under various ambient wet-bulb temperatures and cooling loads. The results indicated that incorporating headers and a mixed configuration of large and small water pumps significantly enhances the cooling system's energy efficiency. With an appropriate control strategy, the annual operational energy consumption can be reduced by 3% to 15%. Building on the traditional water-cooled chilled water system architecture, Wang et al.[36] proposed a cooling system utilizing dual-temperature chilled water to improve energy efficiency through cascading utilization of cooling energy. The results showed that compared to the conventional water-cooled chilled water system, the staged treatment system achieved an annual energy savings rate of 27.4%, with the initial investment recoverable within 2.55 years. Zhang et al.[37] addressed the partial load issues caused by insufficient server rack utilization in data centers by proposing an optimization method for chiller configuration tailored to progressive loads. The results demonstrated that this optimization method could reduce the data center's PUE by 0.02 to 0.11 (depending on the climate zone). Xiang et al.[38] conducted an optimization study on the volume of cold storage tanks in the system. The results revealed that as the volume of the cold storage tank increases, the overall PUE of the system tends to decrease. When the volume is increased to 7.5 times the initial value, an annual energy savings of 1.12% can be achieved. To fully leverage natural cold sources, Zhao et al.[39] introduced a novel water-mediated indirect evaporative cooling chiller designed to replace the cooling water loop (chiller and cooling tower) in the system. When this chiller was applied to the Changji Data Center project in China, the study found that its wet - bulb efficiency ranged from 120% to 152.8%, with a dew - point efficiency of 27% to 62.7%. Moreover, the PUE of the cooling system could be reduced to 1.212.

Liquid-cooled systems. The liquid-cooled system (as shown in Figure 2B(2)) employs liquid-cooled terminals represented by cold plates and liquid-cooled tanks. It boasts a higher heat-transfer efficiency than the water-cooled chilled water system, so it generally does not require chillers, and its overall solution is simpler. To further enhance its operational energy efficiency, numerous experts have also conducted research on the optimization of its solution/design parameters. Wang et al.[40] compared the heat-transfer performance, flow field characteristics, and energy consumption of blade servers under three different cooling methods. It also explored the influence mechanism of different oil-based coolant characteristics on the performance of single-phase immersion liquid-cooled. The results showed that compared with forced air-cooled and cold plate schemes, single-phase immersion liquid-cooled could reduce the system's PUE by 20.8% and 17.6%, respectively. Based on the concept of cascading utilization of cooling energy, Zhang and Hnayno et al.[41,42] proposed a hybrid liquid-cooled system that combines cold-plate liquid-cooled with single-phase immersion cooling technology. Specifically, it first cools the CPU and GPU using cold plates and then uses the return water to cool other components on the server (via single-phase immersion liquid cooling). The study focused on testing the impact of dielectric fluid types, water-loop configurations, and server power on system performance. The results indicated that compared with existing air-cooled and liquid-cooled technologies, this new liquid-cooled system could reduce the energy consumption of the cooling system by 20% and 7%, respectively. Li et al.[43] focused on the pipeline optimization of cold-plate liquid-cooled systems, comparing two different architectural schemes: series and parallel. The results showed that while the cold-plate parallel system had higher control complexity, its energy consumption could be reduced by more than 30% under the same test environment. In Muneeshwaran et al.,[44] a single-phase liquid immersion cooling system using FC-40 insulating fluid was developed, and design parameters such as the inlet/outlet layouts (T-type and Z-type layouts) and different heat sink bases were optimized. The results demonstrated that compared with the Z-type layout, the T-type layout could reduce the thermal resistance by 12.6% and the casing temperature by 0.5-2.8°C. Additionally, heat sinks with heat pipes and vapor chambers as bases had minimum casing temperatures of 56.7°C and 55.2°C, respectively, while the heat sink with a flat-plate base had a casing temperature as high as 60°C.

Due to the high heat-transfer efficiency of liquid-cooled systems, their designed supply and return water temperatures are much higher than those of water-cooled chilled water systems. Consequently, experts and scholars have explored the waste-heat utilization potential of liquid-cooled systems through various approaches. Ebrahimi et al.[45] attempted to use chip waste heat to drive absorption refrigeration and proposed an absorption refrigeration architecture based on two-phase heat transfer. On this basis, the study compared the performance of two refrigerant-absorbent pairs: lithium bromide-water and water-ammonia. The results indicated that for the waste heat characteristics of data centers, the absorption refrigerator using the lithium bromide-water refrigerant-absorbent pair performed better. Haywood et al.[46] proposed an absorption refrigeration architecture based on cold plates. This architecture utilizes the high-quality waste heat (up to 125°C) from the CPUs inside blade servers, which is transferred through the cold plates, to drive an absorption refrigerator (with a required temperature range of 70-95°C). The results showed that this architecture could recover more than 85% of the waste heat from the CPUs of blade servers.

In addition to absorption refrigeration, the waste heat in data centers can also be utilized for adsorption refrigeration, which has lower requirements for waste-heat quality (with a driving temperature that can be below 60°C). In Pan et al.,[47] a silica gel-water adsorption refrigerator was employed for waste-heat recovery and refrigeration in data centers, and the feasibility of applying this solution in the Shanghai region was studied. The results showed that the payback period of the system was less than one year. Given its ten-year service life, it has great potential for market application. Du et al.[48] proposed a waste-heat recovery solution suitable for an air-liquid co-existing architecture, that is, recovering the waste heat of the liquid-cooled system through a silica gel-water adsorption refrigerator to provide cooling for the air-cooled system. The results indicated that under typical summer conditions (with inlet temperatures of hot water/cooling water/chilled water at 50°C/30°C/23°C respectively), the coefficient of performance (COP) of this adsorption refrigerator could reach 0.50.

It should be noted that while the high waste heat temperature represents a significant advantage of liquid-cooled systems, it also introduces new challenges. Compared to water-cooled chilled water systems, the supply and return water temperatures in liquid-cooled systems are relatively high (falling within the optimal activity range of mesophilic bacteria at 40–60°C), rendering them more susceptible to the synergistic detrimental effects of bio-mineralization deposition and electrochemical corrosion induced by the metabolic activities of microbial communities such as sulfate-reducing bacteria and iron bacteria.[49,50] Given that the heat exchangers in the cooling terminals (cold plates or liquid-cooled tanks) and the Cooling Distribution Unit (CDU) are highly precise, they are particularly vulnerable to damage. In practical applications, liquid-cooled systems typically employ a three-stage heat exchange architecture (either liquid-cooled terminal - CDU-plate heat exchanger - open cooling tower or liquid-cooled terminal – CDU - closed cooling tower), with chemical dosing and filtration devices installed in each loop to mitigate the impact of water quality issues. However, this approach increases deployment costs, and the multi-stage heat exchange inevitably affects the operational energy efficiency of the system. Therefore, our team believes that it is essential to further investigate the heat exchange characteristics of liquid-cooled data centers and develop appropriate water treatment technologies based on electrochemical methods to further enhance the operational energy efficiency of liquid-cooled systems.[51]

Fluorine-based systems

The fluorine-based system generally refers to a cooling system that uses fluorinated hydrocarbons as refrigerants and does not require water as the secondary refrigerant. Currently, it is also widely applied in the field of data center cooling systems. Based on the system composition, it can be roughly divided into heat-pipe systems and hybrid systems (heat pipe coupled vapor compression refrigeration systems).

Heat-pipe systems. A heat pipe is a highly efficient heat - transfer device that utilizes phase - change processes to rapidly conduct heat within a closed system. It enables the full utilization of natural cold sources. Based on types, it can be roughly classified into separated heat pipes, fluorine-pump heat pipes, and pulsating heat pipes.

(1) Separated heat pipes

A separated heat pipe is a highly efficient heat-transfer device that uses gravity as the driving force. Its evaporation section and condensation section are arranged separately and connected via pipes to form a closed loop, and it is also known as a gravity heat pipe. Liu et al.[52] conducted research on the application of separated heat pipes in data rooms, focusing on exploring the relationship between heat-transfer efficiency and changes in air volume and windward area. Key parameters such as the maximum allowable outdoor temperature for system operation, annual operating time, power -, and annual electricity savings were derived. Qian et al.[53] investigated the optimal charge ratio of separated heat pipe cooling system systems in data rooms. The study found that the optimal charge ratios for both R22 and R134a refrigerants were around 80%. Zou et al.[54] studied the enhancement of heat-transfer performance and stability of separated heat pipes. The results showed that: ① Within the optimal range, a lower charge ratio leads to better fault-resistance performance of separated heat pipes; ② Under the optimal charge ratio, the position of a faulty fan has a relatively minor impact on the overall heat-transfer performance of separated heat pipes; ③ The server arrangement can cause uneven inlet air temperatures at the evaporator, which has a certain influence on the optimal charge ratio range of separated heat pipes. Wang et al.[55] explored the effects of different refrigerant flow rates, inclination angles, and inlet gas-phase fractions on a micro-channel separated heat pipe. The results indicated that for the R134a medium: ① When the inlet flow velocity is 0.005 m/s, the outlet exhibits a superheated vapor state; ② Setting the evaporator at an inclination angle of 0 - 20° can not only significantly reduce the pressure difference between the inlet and outlet but also improve the heat-transfer coefficient and refrigerant flow velocity; ③ Increasing the inlet steam volume fraction can promote bubble nucleation and reduce the pressure drop, but it will result in outlet superheating.

(2) Fluorine-pump heat pipes

Since the separated heat pipe relies on gravity as the driving force, a sufficient height difference between the condenser and the evaporator is required. When the height difference cannot be met, or when the pipeline in the adiabatic section of the heat pipe is excessively long and the flow resistance is high, there may be insufficient driving force. In such cases, an additional driving force is often provided by adding a fluorine-pump (liquid pump/gas pump).[56]

Gong et al.[57] combined a separated heat pipe with a fluorine-pump to develop a water-cooled micro-channel separated fluorine-pump heat pipe system, which can operate in various modes, including gravity circulation and variable-frequency fluorine-pump operation. Research shows that compared to the gravity circulation mode, the cooling capacity of the fluorine-pump operating at speeds of 800 RPM, 1600 RPM, and 2400 RPM can increase by 0.98%, 4.45%, and 9.04%, respectively. Vasta[58] developed a fluorine-pump heat pipe system suitable for large-capacity heat exchange using R600 as the working fluid. The study indicates that the system can achieve a heat exchange capacity of 3 kW when using an 8-millimeter diameter pipeline, maintaining a reasonable temperature difference between the evaporator and the condenser, and with a liquid filling rate of 46%. Li et al.[59] proposed a novel rack-level cooling system that adds an auxiliary fluorine-pump to the backplane evaporator and is equipped with a large-capacity liquid storage tank, aiming to address the common issues of insufficient and uneven refrigerant supply in fluorine-based systems. Experimental results show that the system can provide an adequate refrigerant supply and facilitate uniform refrigerant distribution. Compared to traditional compression refrigeration air conditioners, this system can achieve energy savings of over 50% in northern China. Lin et al.[60] investigated the performance variations of a composite air conditioning system based on fluorine-pump under different configuration options, as well as its application effects in different climatic regions of China. The results indicate that in typical cities in northern China, an Annual Energy Efficiency Ratio (AEER) of 6.6-10.7 can be achieved, while in typical cities in southern China, an AEER of 4.6-5.3 can be realized.

(3) Pulsating heat pipes

The pulsating heat pipe, first proposed by Japanese scholar Akachi in the 1990s, is an advanced heat conduction device that achieves high heat flux density transfer by utilizing the interaction mechanism between vapor-liquid plugs of the internal working fluid.[61] Currently, this technology has been explored for applications in data centers and electronic thermal control fields. Chen et al.[62] studied the start-up, heat transfer, and flow characteristics of a pulsating heat pipe under uniform and non-uniform heating, using HFE7100 (50% fill ratio, 10-45 W heating power). Results showed vapor-liquid conversion frequency rose with higher input power, but due to fill ratio and initial fluid position, some tubes had higher frequencies during local pulsation than stable pulsation. Liu et al.[63] proposed a novel pulsating heat pipe with alternating tube diameters and grooved structures, and constructed an experimental system without an additional cold source. The research results indicated that at a heating power of 30 W, its thermal resistance was 0.05°C/W, which was 1 - 2°C/W lower than that of traditional heat pipes. Lu et al.[64] conducted research on data center rack-level cooling technology based on a plate-type pulsating heat pipe, focusing on analyzing the effects of temperature, heat source power, filling ratio, and condenser fin configurations on the start-up and stable operation performance of the pulsating heat pipe. The study found that for cabinets with a power not exceeding 1380 W, a structurally optimized cabinet cooling system could keep the CPU temperature below 60°C.

Hybrid systems (heat pipe coupled vapor compression refrigeration systems). Given that the cooling effect of a single heat pipe system is significantly and easily influenced by weather conditions, some researchers have combined it with vapor compression refrigeration to form a hybrid system (as shown in Figure 3A). Chen et al.,[65] addressing issues such as high costs, switching failures, and refrigerant leakage in fluorine-pump (liquid-phase) heat pipes, proposed a variable frequency multi-split computer room air conditioning system incorporating fluorine-pump

(gas-phase) heat pipes. It experimentally explored the application effects of this system in seven typical cities across China. The results showed that, compared to traditional compression refrigeration air conditioners, this system could achieve an annual energy saving rate of nearly 50% in Harbin and over 30% in Guangzhou. Yang et al.[66] proposed using a heat pipe-based indirect evaporative cooling system as a pre-cooling system (to pre-cool the return air in computer rooms) and combining it with an air-cooled precision air conditioner to enhance energy efficiency in data centers located in hot and humid climatic regions. Huang et al.[67] proposed integrating evaporative condensation with fluorine-pump heat pipe technology and developed an air conditioning prototype with a cooling capacity of 30 kW. The energy efficiency ratios (EER) of this unit in dry mode, wet mode, and supplementary cooling mode could reach 12.33, 10.55, and 5.52, respectively. Bai et al.[68] designed a novel hybrid cooling system that combines vapor compression refrigeration with a pump-driven loop heat pipe using an intermediate plate heat exchanger, enabling three operating modes: heat pipe mode, refrigeration mode, and dual mode. Experimental results showed that when the temperature differences between indoors and outdoors were 10°C and 18°C, respectively, the EER of the heat pipe mode was 5.6 and 10.4. Moreover, the AEER of the hybrid system could reach 8.3. Wang et al.[69] proposed a hybrid computer room air conditioning system, which is formed by cascading a fluorine-pump heat pipe system with a vapor compression refrigeration system. The research results indicated that, compared to air-cooled direct expansion units and air-cooled dual-cooling-source chiller units, this hybrid air conditioner could achieve energy savings of 4.8% to 46%, with a maximum AEER of up to 7.9. Han et al.[70] proposed a data center hybrid air conditioning system incorporating evaporative condensation technology and explored the impact of evaporative condensation technology on heat pipes

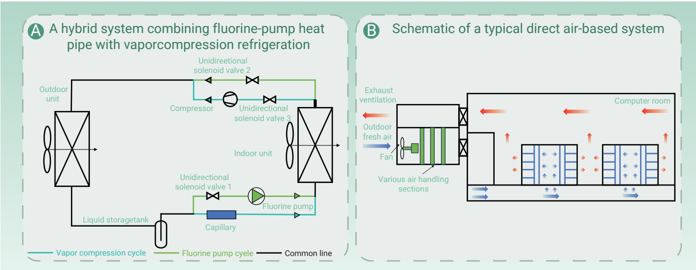


Figure 3. (A) Hybrid system combining fluorine-pump heat pipe with vapor compression refrigeration.[57] (B) Schematic of a typical direct air-based system

and vapor compression refrigeration. The results showed that evaporative condensation technology helps reduce the condensation temperature, and under a constant cooling supply, the system's COP could increase by 9.82%.

Air-based systems

Direct air-based systems. In some regions with favorable natural cold source endowments, data centers also adopt air-based systems. This involves installing air handling units on the exterior walls of the computer rooms to reduce the number of heat exchange stages and shorten the cooling supply path. Air-based systems can be classified into direct air-based systems (as shown in Figure 3B) and indirect air-based systems based on whether outdoor air directly exchanges heat with the equipment inside the computer room. In a direct air-based system, outdoor air is processed (including evaporative cooling, dust removal, humidification/ dehumidification, etc.) and then introduced into the computer room for cooling. This system has the fewest heat exchange stages and the shortest cooling supply path. Siriwardana et al.[71] investigated the potential of direct air-side free cooling in Australia and analyzed meteorological data from 2000 to 2011 at 20 meteorological stations. The results indicated that southern cities in Australia and Tasmania have free cooling potential exceeding 5,500 hours per year. Oró et al.[72] explored the potential of direct air-side free cooling for data centers in different regions of Europe, as well as its integration with thermal energy storage technologies. The results showed that, under a time-of-use electricity pricing mechanism, the integration of direct air-side free cooling with energy storage could reduce the operating costs of cooling systems by over 50%. Although direct air-based systems offer high efficiency, their applicability is easily restricted by regional natural conditions (for example, direct air-based systems are generally not suitable for areas with high humidity/poor air quality) and require additional maintenance costs.

Indirect air-based systems. In contrast to direct air-based systems, in indirect air-based systems, the indoor and outdoor air are not directly connected; instead, heat exchange occurs within the air handling unit. Although this increases the number of heat exchange stages and the transfer path, the operation and maintenance of this system are more stable and reliable. Ham and Yu et al.[73,74] pointed out that, compared with traditional vapor compression refrigeration cycles, the use of indirect air-based systems can still achieve energy savings of 63.6% to 67.2%. Kim et al.[75] proposed a multi-stage free cooling scheme that integrates an indirect air-based system with a water-based system. The results showed that, compared with a standalone indirect air-based system, the new system could reduce the energy consumption of the cooling system by approximately 50%.

Other systems

To meet the diverse demands of different IT business types, as well as energy and resource utilization requirements, the aforementioned systems can be flexibly combined or integrated with other energy systems. To reduce the number of heat exchange stages and improve the efficiency of free cooling, Zou et al.[76] proposed a hybrid thermosyphon cooling system (serving as an auxiliary to the water-cooled chilled water system). This system employs an air-side economizer to assist water-cooled in-row thermosyphon tubes (terminal equipment) in achieving free cooling. The results showed that, compared with traditional cooling methods, the new system could increase the duration of free cooling by up to 4,589 hours. Wang et al.[77] integrated loop thermosyphon tubes into a conventional water-cooled chilled water system, creating a dual-cold-source cooling system. This system combines air-side cooling with water-side cooling to extend the year-round utilization time of free cooling. Experimental results demonstrated that vapor compression refrigeration was only required when the outdoor air temperature exceeded 30°C and the chilled water temperature exceeded 25°C.

In addition to natural cold source, the cold from some industrial sectors can also be applied to data center cooling. Liu et al.[78] investigated the integration of liquid air energy storage with data center cooling systems, aiming to utilize the cold released during the evaporation of liquid air for data center cooling, and proposed an immersion cooling system for data centers based on liquid air. Ting et al.[79] explored the combination of the cascade utilization of liquefied natural gas (LNG) cold energy and data center cooling. Yuan et al.[80] further discussed the utilization of data center waste heat for seawater desalination and agricultural production. Weng and Demir et al.[81,82] proposed a composite system that employs separated heat pipes to utilize data center waste heat for domestic heating.

Discussion and Analysis

Cooling Design Temp.: Determined by IT & Cooling Systems. With the enhancement of heat exchange performance of internal components in servers, the upper operating temperature limit of servers has also increased. Currently, it is widely acknowledged in the HVAC industry that the supply air temperature range (18-27°C) specified in ASHRAE TC 9.9 2015[83] is outdated and fails to meet the parameter requirements of new-generation servers. To reduce energy consumption in cooling systems, designers tend to raise the cooling design temperature, even using the server's safe temperature as an optimization constraint, aiming to maximize the free cooling time and the potential for waste heat recovery. However, most of the above studies have not considered the energy consumption issue of IT systems. Due to the influence of transistor leakage current, an increase in the cooling supply temperature of the cooling system will inevitably lead to an increase in server leakage power. IT equipment constitutes the largest portion of energy consumption in data centers, and blindly raising the cooling supply temperature may result in an overall increase, rather than a decrease, in energy consumption for the data center, making it difficult to achieve collaborated energy saving optimization between IT and cooling systems.

Taking an individual server as the research subject, Figure 4A shows the comparative variations in total power and PUE. It is evident that there is an inverse correlation between the cooling power (specifically, fan power) and the leakage power. As the fan power diminishes, the PUE experiences a rapid decline; however, the total power demonstrates a trend of initial decrease followed by an increase, reaching its minimum value when the CPU temperature is at 68°C.

Furthermore, Reference[7] elaborates that for specific types of servers and their corresponding cooling systems, there must be an optimal temperature that achieves an overall minimum trade-off between the power of IT and cooling systems. Therefore, during the cooling system design process, this optimal temperature should serve as the design basis for collaborative energy saving optimization between IT and cooling systems.

Cooling System Needs Flexibility for AI Loads. Compared to traditional IT loads, AI loads are characterized by strong periodicity, high suddenness, and significant volatility. Taking large-scale model applications as an example, the load variations during the inference phase are particularly intense, leading to dramatic changes in the server's heat dissipation (as shown in Figure 4B). Limited by fluid distribution speed and active control methods, the cooling supply struggles to match these power changes in real time. Typically, it can only ensure the safety of IT equipment by overcooling (lowering the temperature excessively), which results in higher energy consumption in the cooling system and reduces the control range for cooling system optimization during operation. Consequently, the effectiveness of collaborated energy saving optimization between IT and cooling systems is constrained. Currently, IT equipment manufacturers have recognized the importance of this issue and have begun developing server thermal management technologies based on phase-change thermal storage.[85,86]

However, constrained by factors such as volume space, material thermal properties, and stability, the above technologies have seen limited application in practical engineering. As AI load proportion and server rated power rise, the response speed gap between IT system heat dissipation and cooling system supply will keep growing. Therefore, it is necessary to develop

Table 2. Summary of cooling system design optimization.

Reference	System type	Optimize objects/measures	Main contribution
[35]	water-cooled chilled water	Pipeline and water pump	The free cooling time increases through pipeline optimization and a mixed setup of large and small water pumps.
[36]	water-cooled chilled water	Dual-temperature chilled water system	Using two cooling tower types allows staged air treatment, enhancing energy efficiency. The Grading Treatment System exhibits markedly higher Specific Cooling Power (SCOP) than the Waterside Economizer System, with its annual average SCOP increasing from 8.20 to 11.30.
[37]	water-cooled chilled water	Chiller plant configuration	To reduce total life-cycle costs, optimal designs for various climates were selected by assessing energy efficiency at full loads and different ambient temperatures. Under constant pressure control, the global cooling system's COP improves by 0.7–2.9, saving 13–22% in cooling energy.
[38]	water-cooled chilled water	Chilled water storage tank volume	The effect of water storage tank volume on energy savings was analyzed. With the same tank size, energy savings reach 27.1%, and electricity cost savings reach 26.7%.
[39]	water-cooled chilled water	A novel water-mediated indirect evaporative cooling chiller	It empirically verifies the use of indirect evaporative cooling water refrigeration units in hot, arid regions, reducing the cooling system's PUE to 1.212.
[40]	Single-phase immersion liquid-cooled	Oil-based coolants	It elucidates how thermophysical properties affect single-phase immersion liquid cooling performance. Compared to forced air and cold plate cooling, it reduces system PUE by 20.8% and 17.6%, respectively.
[44]	Single-phase immersion liquid-cooled	Inlet/outlet layout & heat sink base parameters for single-phase immersion cooling	Developed a FC-40 dielectric fluid-based single-phase immersion cooling system for 1U servers; the T-type layout reduces thermal resistance by 12.6% and casing temperature by 0.5-2.8°C compared to the Z-type.
[45]	Two-phase immersion liquid-cooled	Absorption refrigeration architecture	The performance of two working fluid pairs, lithium bromide-water and water-ammonia, in data center waste heat recovery was compared.
[46]	Cold plate liquid-cooled	Absorption refrigeration architecture	This system can recover at least 85% of high-quality waste heat from CPUs within servers through cold plates.
[48]	Air-liquid co-existing architecture	A compact silica gel–water adsorption chiller	This design harnesses waste heat from liquid cooling to power adsorption refrigeration, satisfying air-cooling needs; its COP reaches 0.50 in typical summer conditions.
[52]	Separated heat pipe	The frontal area and design airflow rate of precision air conditioning	An integrated analysis method, performance prediction model, and experimental setup for separated heat pipe air conditioning systems have been established.
[54]	Separated heat pipe	The server layout, refrigerant type, and charge ratio	The heat transfer of separated heat pipes under non-uniform conditions and optimal charge ratios for R32, R22, and R134a are clarified; R22 and R134a's optimal ratios are around 80%.
[59]	Fluorine-pump heat pipe	A rack-level cooling architecture	It addresses refrigerant supply and distribution issues in fluorine systems, saving over 50% energy vs. traditional air conditioners in northern China.
[64]	Pulsating heat pipe	A rack-level cooling architecture	The effects of temperature, heat source power, liquid fill ratio, and condenser fins on pulsating heat pipe performance were analyzed; an optimized cabinet cooling system maintains CPU temperature below 60°C for cabinets ≤1380 W.
[65]	Hybrid system combining fluorine-pump heat pipe with vapor compression refrigeration	Fluorine-pump (gas-phase) heat pipes	A 40kW VRF air conditioning prototype with free cooling was developed and tested in seven Chinese cities, achieving nearly 50% annual energy savings in Harbin and over 30% in Guangzhou.
[70]	Heat pipe-vapor compression refrigeration hybrid system based on evaporative condensation	Hybrid system architecture	Evaporative condensation tech was studied for its impact on heat pipes and vapor compression systems; it lowers condensation temperature, boosting system COP by 9.82% at constant cooling.
[71]	Direct air-side free cooling	The application performance of direct air-side free cooling was investigated	Analysis of direct air-side free cooling duration in major Australian cities showed southern cities and Tasmania exceed 5,500 free cooling hours annually.
[73]	Indirect air-side free cooling	The performance of nine types of air-side heat exchangers was compared	It offers a guide for direct air-side cooling in mid-latitude data centers; indirect air-based systems save 63.6%-67.2% energy vs. traditional refrigeration.
[76]	Water-fluorine conversion	Air-side heat exchangers were employed to assist water-cooled rack-level thermosyphons for free cooling	It bypasses multi-stage heat exchange in water-cooled free cooling, extending free cooling duration by up to 4,589 hours.
[77]	Water-fluorine conversion	Loop thermosyphons were integrated into the chilled-water cooling system	It merges air- and water-side cooling to prolong natural cold source use, requiring vapor compression only when outdoor air exceeds 30°C and chilled water exceeds 25°C.

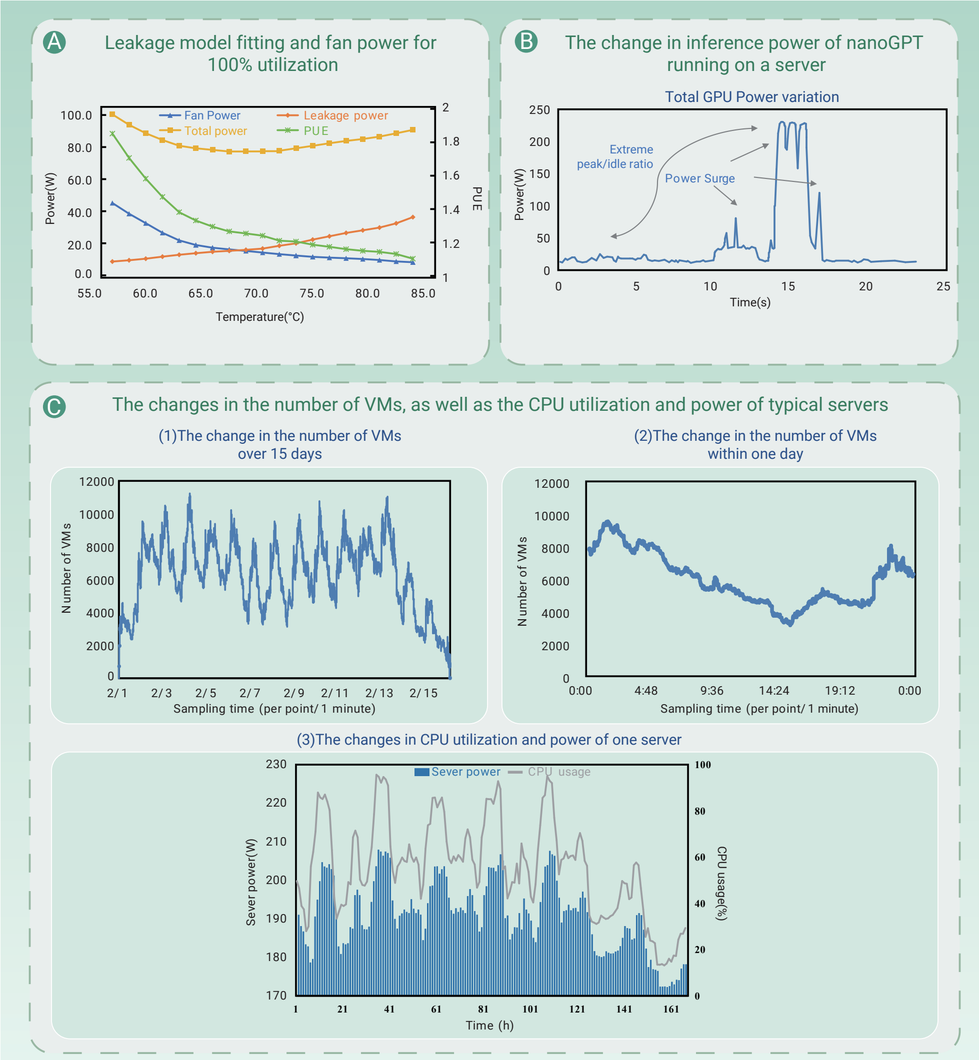


Figure 4. (A) Leakage model fitting and fan power for 100% utilization.[84] **(B) The change in inference power of nanoGPT running on a server.**[87] **(C) The changes in the number of VMs, as well as the CPU utilization and power of typical servers.**[105,106]

passive cold storage and release devices at the terminal side, tailored to the characteristics of various cooling systems and incorporating suitable cold storage and release technologies. This aims to increase the thermal inertia of the server room/loop without affecting heat exchange efficiency, thereby enhancing the cooling system's flexible adaptability to the dynamic and high-intensity AI loads, and laying the foundation for collaborated energy saving optimization between IT and cooling systems.

Matching Quantification: IT, Cooling & Environment. This section introduces research on the design optimization of various mainstream cooling systems. However, the effectiveness of cooling systems in practical applications is simultaneously influenced by IT equipment loads and outdoor weather parameters, and the applicable scope of each system has not yet been fully clarified. In practical engineering projects, designers typically define the range of selectable cooling system types based on constraints such as the air/liquid-cooled form of servers, the rated power per rack, and the PUE indicator. Nevertheless, due to the lack of quantitative comparisons of the performance of various system types, further decision-making on schemes often relies solely on subjective experience or other business requirements, which can easily limit the effectiveness of collaborated energy saving optimization between IT and cooling systems. Influenced by the development of AI technology, the power characteristics of servers also exhibit features closely related to IT business types, yet the appropriate cooling system solutions for various IT business types remain unclear. Additionally, the geographical applicability of different cooling system types still requires further research.

Taking water-cooled chilled water systems and air-based systems (which are most commonly used with air-cooled servers) as examples, air-based systems have fewer heat exchange stages and shorter cooling supply paths, offering superior utilization of natural cold sources, but their mechanical refrigeration efficiency is often lower than that of water-cooled chilled water systems. The selection between the two should be based on the annual duration of free cooling in the region, but this boundary has not yet been clearly quantified. Therefore, it is necessary to quantify the matching characteristics among IT business types, cooling system schemes, and outdoor environments to lay the foundation for collaborated energy saving optimization between IT and cooling systems.

OPERATION OPTIMIZATION OF DATA CENTER

A reasonable cooling system design serves as the foundation for the collaborative energy saving optimization between IT and cooling systems, while the operational control of the two is the key to facilitating such collaborative energy saving optimization. To this end, this section reviews existing research from the perspectives of individual and joint energy saving control optimization respectively and analyzes the issues within them.

Individual energy saving control for the IT/cooling system

Research on energy saving control for IT systems. Energy saving control for IT systems is one of the research priorities in the field of cloud computing. Its main optimization direction is to enhance the overall utilization efficiency of server resources under a given workload, and the key lies in reducing the number of active servers. Feng, D.[88] proposes a workflow low-power scheduling method relying on consolidation and allocation. It optimizes workflow scheduling to cut the number of active servers, dividing the process into three stages: task consolidation based on the longest path, server scheduling based on consolidated task sequences, and task slackening according to scheduling results. The method's effectiveness was validated with both scientific and random workflows. Taking the enhancement of energy efficiency in cloud systems and the assurance of cloud users' Quality of Service as its foundation, Cui, Y.[89] thoroughly accounts for the differing requirements between real-time and non-real-time tasks. It then puts forward a task scheduling strategy incorporating priority and partial synchronization multiple semi-sleep modes. In a heterogeneous cloud environment, this strategy is capable of considering multiple key metrics, including task waiting time, system blocking rate, and system energy saving rate. Han et al.[90] addresses the large-scale virtual machine (VM) allocation problem by developing a heuristic algorithm, the GammaFF algorithm, and tests its performance using Huawei Cloud datasets. The test results demonstrate that the GammaFF algorithm outperforms traditional algorithms such as FirstFit, BestFit, and HLEM-VMP in terms of reducing the number of active servers and minimizing VM migration overhead.

Considering the time-varying characteristics of IT loads, as well as the IT system response delays caused by data transmission and server sleep/active switching,[91] researchers employ load-prediction-based virtualized resource scheduling techniques in energy-efficient operation and maintenance work. This technique can alleviate the lag problem in VM scheduling that relies solely on current load conditions, further reduce the idle rate of hardware resources, and minimize the number of active servers while ensuring that service quality remains unaffected. Dabbagh et al.[92] proposed an integrated and energy-efficient resource allocation framework specifically tailored for resource oversubscribed cloud environments. This framework achieves energy consumption optimization through a dual-mechanism approach: Firstly, it conducts real-time monitoring and accurate prediction of VM (VM) resource usage, effectively reducing the frequency of server overloads. Secondly, it employs intelligent migration and deployment strategies to dynamically reduce the number of active servers. Wu, J.[93] focuses on the selection and placement aspects during the process of VM migration. To address the issue of inefficient migration caused by the use of fixed objective functions in VM selection, it proposed a prediction-based resource-aware algorithm. This algorithm combines current and predicted resource utilization rates to determine the status of servers and dynamically defines the selection function, thereby improving efficiency with fewer migrations. The results demonstrate that this prediction-based resource-aware algorithm achieves over 25% lower energy consumption than the Minimum Utilization method. Xu, J.[94] presents a combined prediction model based on Gated Recurrent Unit-Long Short-Term Memory (GRU-LSTM) to enhance the accuracy and speed of current load prediction models. It also incorporates monitoring, load prediction, and elastic scaling modules. Experiments demonstrate that this elastic scaling solution outperforms traditional Horizontal Pod Autoscaling.

However, consolidating workloads will also increase the heat dissipation of the remaining active servers. If the cooling system fails to promptly increase the cooling supply at their locations, local hotspots are prone to emerge, leading to a rise in the leakage power of the servers. Cheng et al.[12] tests the performance of servers equipped with Xeon Silver 4210 processors. The results show that when the core temperature rises from 56.8°C to 59.1°C, the server power under the same CPU utilization rate increases by approximately 15.1%. Zapater et al.[95] finds that under full-load conditions, as the CPU temperature rises from 55°C to 85°C, the proportion of leakage power in the total power increases from 18% to 46%. Although some researchers have developed energy-efficient VM scheduling algorithms that take both temperature and resource utilization into account,[96,97] these algorithms mainly focus on balancing the energy consumption and security of IT systems.

Research on energy saving control for cooling systems.. The primary optimization direction for energy-efficient control of cooling systems is on-demand cooling,[98] which involves optimizing the timing, quality, and quantity of cooling based on the actual heat dissipation requirements of active servers to maintain their temperatures within the target range. Considering the inherent time-lag characteristics in the cooling supply of cooling systems, many researchers have adopted prediction-based control methods for cooling supply. Compared to traditional feedback control, these methods can achieve over 20% energy savings in cooling systems.[99,100] Qu et al.[101] uses the safe temperature of chips as an optimization constraint for predictive cooling control, enabling a 42.7% reduction in cooling energy consumption under variable load conditions. The aforementioned studies have all achieved notable results in terms of energy saving in cooling systems. However, reducing the cooling supply also leads to an increase in server leakage power. To achieve a trade-off between server energy consumption and cooling energy consumption, Zapater et al.[95] takes a single server as the research object, clarifies its leakage power variation characteristics through experimental methods, and proposes a fan control strategy that considers both leakage power and fan power, resulting in a 9% reduction in total energy consumption under given operating conditions. Nevertheless, unlike fan control for a single server, data center cooling systems typically need to provide cooling services for thousands or even hundreds of thousands of servers, and IT loads are characterized by high volatility and frequent changes.[102,103] Using Alibaba Cloud's dataset as an example, the number of VMs can abruptly increase from dozens to over a thousand within a few minutes (as shown in Figure 4C). Due to the significant differences in response speed and control accuracy between the IT and cooling systems,[104] it is difficult for unilateral adjustments by the cooling system to achieve on-demand cooling, let alone optimize the overall energy efficiency of both IT and cooling systems.

Joint energy-efficient optimization of the IT and cooling systems

To address the issue that independent energy-efficient optimization for IT and cooling systems cannot achieve optimal overall energy efficiency, researchers have conducted studies on the joint energy-efficient optimization of the two. In Gupta et al.,[107] server utilization, chilled water temperature, and air-conditioning outlet airflow rates are taken as optimization variables. The results show that through solution via optimization algorithms, the computing efficiency ratio can be improved by up to 11.6%. Zhou et al.[108] carries out joint optimization of server load rates and outlet air temperatures of air conditioners. The results indicate that through solution via optimization algorithms, up to 62% energy savings can be achieved under specific low-load operating conditions. Most of the aforementioned studies employ heuristic algorithms for iterative optimization and have achieved remarkable results. However, the joint energy-efficient optimization of IT and cooling systems is a high-dimensional discrete-continuous optimization problem with an extremely large solution space.[109] Due to the lack of a clear optimization direction, heuristic algorithms often perform poorly under actual operating conditions with dynamically changing IT loads (where the solution time is extremely limited).

To tackle the problem that joint energy-efficient optimization based on heuristic algorithms is difficult to solve efficiently, some researchers have attempted to employ reinforcement learning for joint optimization. In Ran et al.,[104] a deep Q-network algorithm based on a parameterized action space is proposed to jointly optimize job scheduling and air volume control. The results show that compared with heuristic global optimization, this method can improve energy efficiency by 10%. Zhou et al.[110] improves upon the aforementioned reinforcement learning algorithm to further enhance the control collaboration between IT and cooling systems, as well as the algorithm's applicability to multi-dimensional IT resources. Chi et al.[111] further corroborates the energy saving effects of reinforcement learning algorithms in the joint scheduling optimization of IT and cooling systems. The study indicates that compared with energy-efficient optimization conducted solely on a single system, joint optimization based on reinforcement learning algorithms can improve energy efficiency by up to 15%. From the above analysis, it can be seen that reinforcement learning algorithms hold great potential in joint optimization. However, current research is limited to the simulation level and lacks practical engineering applications.

Discussion and Analysis

Lack Accurate Real-Time Server Leakage Power Quantification. From the above content, it can be seen that to achieve the overall optimal energy efficiency of the IT system and the cooling system in a data center, it is necessary to carry out energy saving optimizations on both systems simultaneously. Given that joint energy saving optimization represents a high-dimensional discrete-continuous optimization problem with an extremely large solution space, finding its key direction is crucial for solving it. Based on the power mechanisms of servers and cooling systems, some researchers have currently proposed using server temperature as the key direction for joint optimization (as shown in Figure 5A).[7,112] Specifically, the optimal server operating temperature corresponding to the lowest overall energy consumption under various operating conditions is first calculated in advance based on IT load and outdoor parameters. Subsequently, during actual operation, the IT system and the cooling system are regulated separately or in a collaborated manner with the goals of maintaining the number of active servers and their operating temperatures at optimal values.

The key to the validity of the above approach lies in identifying temperature related indicators that can quantify changes in server leakage power. Only in this way can the optimal operating point be determined based on the "energy consumption seesaw effect" between the IT system and the cooling system. However, there are significant differences in the response speeds and control precision between the IT system and the cooling system, leading to severe non-uniformity in the temperature field within servers. Furthermore, given the extremely complex composition of electronic components inside servers, it is difficult to directly calculate the real-time leakage power of servers through weighted summation.

At present, our team's preliminary approach is to introduce a comprehensive indicator termed "equivalent temperature" for characterizing the leakage

Table 3. Summary of data center operation optimization.

Reference	Optimize objects	Optimize measures	Main contribution
[88]	IT system	Reduce the number of active servers by optimizing workflow scheduling	A workflow scheduling framework based on consolidation and allocation is proposed
[89]	IT system	Balance server energy consumption and response speed through a semi-dormant mode	A task scheduling strategy with priority and partial synchronization multiple semi-dormant modes is proposed
[90]	IT system	Decrease the number of active servers and reduce the overhead of VM migration	GammaFF algorithm
[92]	IT system	Achieve over-allocation of IT hardware resources through load forecasting	A resource allocation framework for oversubscribed cloud environments is proposed, reducing total energy consumption by 40% versus no overcommitment.
[93]	IT system	Combining current and predicted resource utilization rates to reduce the number of VM migrations	An energy-aware scheduling strategy is proposed, reducing energy consumption by over 25% compared to the Minimum Utilization method.
[94]	IT system	By classifying IT loads, prediction accuracy is improved	A combined prediction model based on GRU-LSTM is proposed.
[95]	Server's CPU and fan	A fan controller has been designed that can minimize the total energy consumption	The trade-off between server leakage and fan power is examined; at full load, as CPU temp rises from 55°C to 85°C, leakage power's share increases from 18% to 46%.
[96]	IT system and air-cooled air conditioner	VMs are placed and migrated based on a greedy strategy	A GRANITE VM scheduling algorithm is proposed, GRANITE consumes 4.3% - 43.6% lesstotal energy in comparison to the state-of-the-art.
[100]	Water-cooled chilled water system	The timing control of the chiller and storage system has been optimized	A mixed-integer linear programming-based model predictive control method is proposed, achieving over 20% energy savings in cooling systems versus traditional feedback control.
[101]	Cold plate liquid-cooled system	Optimizing system power with chip temperature constraints, we adjust cooling tower airflow and primary/secondary flow rates	A feedback-regulated liquid-cooled system strategy is proposed, optimizing predictive cooling with chip temperature constraints to cut energy use by 42.7% under variable loads.
[107]	IT system and a special row-level cooling unit	Server utilization, water temperature, and air-conditioning airflow rates are optimized	A method balancing energy, exergy, and computational efficiency is proposed, improving computing efficiency by up to 11.6% via optimization.
[108]	IT system and room-level air conditioner	Server load distribution and air conditioning supply air temperature are optimized	A machine learning energy management scheme achieves up to 62% energy savings under low-load conditions via optimization.
[104]	IT system and room-level air conditioner	Jointly optimize job scheduling and air conditioning air volume	A deep reinforcement learning framework, DeepEE, is proposed, enhancing energy efficiency by 10% over heuristic global optimization.
[111]	IT system and room-level air conditioner	Jointly optimize IT load, supply air temperature, and flow rate	A deep reinforcement learning-based joint optimization method, MAD3C, improves energy efficiency by up to 15% compared to single-system optimization.
[112]	IT and cold plate liquid-cooled system	Jointly optimize VM deployment, migration, CDU water pump speed, cooling water pump speed, and cooling tower fan speed	A joint optimization strategy based on optimal temperature control is proposed, avoiding optimization for high-dimensional action spaces

power of servers. This indicator is based on the multi-physics field parameters of servers, encompassing resource utilization, inlet and outlet temperatures, as well as flow rate. Specifically, through a data-driven methodology, we directly quantify the correlation between the aforementioned parameters and the real-time leakage power of servers. To achieve this objective, temperature and flow rate sensors must be installed at the inlets and outlets of each server, and data communication between the IT system and the cooling system within the data center should be established.

Missing Precise Offline Training on IT-Cooling Joint Scheduling. Reinforcement learning is another approach to achieving joint energy saving optimization of IT and cooling systems, which can avoid directly solving high-dimensional discrete-continuous optimization problems. The application of reinforcement learning involves collecting operational data across all conditions and extensive trial-and-error training for the agent. Unlike entertainment scenarios such as large language models and short videos, data centers have extremely high requirements for safety and stability. Consequently, the available space for online training of the agent is extremely limited, necessitating the establishment of an offline training environment that precisely describes the dynamic characteristics of joint scheduling between IT and cooling systems. However, due to factors such as server heterogeneity and changes in server rack utilization rates, there is currently a lack of high-precision methods for constructing offline training environments tailored to real-world scenarios, which severely restricts the practical application of reinforcement learning algorithms.

FUTURE RESEARCH DIRECTIONS

The "computational efficiency" mechanism and evaluation indicators

In previous section of this paper, the defects inherent in the existing PUE indicator and its derivative indicators are elucidated, specifically highlighting their oversight of the impact exerted by thermoelectric coupling characteristics. Consequently, pursuing the optimal individual energy efficiency of IT/cooling systems alone may actually lead to an increase in the overall energy consumption of data centers. To address this, existing standards and research have proposed the concept of the "computational efficiency" indicator,[113] aiming to evaluate the overall energy efficiency of data centers, as shown in Equation 15.[114] However, how to quantify *CP* in Equation 15? Some studies use floating point operations per second (FLOPs), but this indicator is generally only suitable for measuring the theoretical peak computational capacity of processors. The demands of IT operations on hardware resources are usually complex, and a single FLOPs indicator clearly cannot summarize the actual task execution efficiency. (for example, there may be abnormal resource occupation). Furthermore, the actual task execution performance and hardware resource utilization are also closely related to scheduling algorithms, meaning that two servers performing the same task

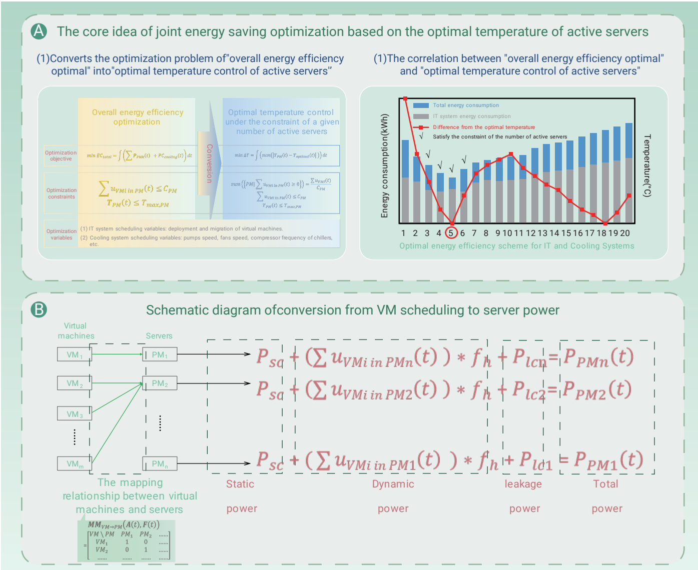


Figure 5. (A) The core idea of joint energy saving optimization based on the optimal temperature of active servers.[112] (B) Schematic diagram of conversion from VM scheduling to server power.

may have the same *CE* but significantly different P_{total} values. Therefore, it is unreasonable to evaluate "computational efficiency" solely based on the ratio of hardware resource usage to power.

$$CE = \frac{CP}{P_{total}} \tag{15}$$

In the above content, *CP* is the *f*(IT business, including computing, storage, and communication, etc.) of the data center; P_{total} is the total power of the data center; *CE* is the computational efficiency of the data center.

In summary, the evaluation of "computational efficiency" should focus on the business types of IT systems. Subsequently, starting from the underlying mechanisms of the three dimensions—software, hardware, and cooling—we should investigate the complex interrelationships among the execution quality of IT operations, the execution time of IT operations, and the total energy consumption/total carbon emissions of data centers. This particular research endeavor will constitute a central focus in our team's forthcoming investigations.

Universal cross-scale thermoelectric coupling model

In Sections 3 and 4 of this paper, the significance of optimal temperature in collaborative energy saving design and operation is expounded. However, how can we obtain this optimal temperature? This typically necessitates detailed optimization calculations based on a thermoelectric coupling model that spans from servers to the entire data center. Our research team has presented a simplified matrix-based modeling methodology for liquid-cooled data centers in Reference,[112] as shown in Figure 5B and Equations 15-20.

$$\boldsymbol{P}_{PM}(t) = \boldsymbol{P}_{sc}(t) + \boldsymbol{P}_{d}(t) + \boldsymbol{P}_{lc}(t) \tag{16}$$

$$\boldsymbol{P}_{d}(t) = \boldsymbol{u}_{VM}(t) * \boldsymbol{MM}_{VM \to PM}(\boldsymbol{A}(t), \boldsymbol{F}(t)) * f_h \tag{17}$$

$$\boldsymbol{P}_{lc}(t) = \boldsymbol{D}(\boldsymbol{T}_{CPU}(t)) \tag{18}$$

In Figure 5B and equations (15) - (20), $\boldsymbol{P}_{PM}(t)$is the total power matrix composed of servers in the data center; $\boldsymbol{P}_{sc}(t)$is the static power matrix of servers (active servers are constant values, dormant servers are 0); $\boldsymbol{P}_{d}(t)$is the dynamic power matrix of servers; $\boldsymbol{u}_{VM}(t)$ is the real-time resource utilization matrix of VMs; $\boldsymbol{MM}_{VM \to PM}$ is the mapping matrix between VMs and servers, which is a function of the real-time deployment $\boldsymbol{A}(t)$ of and the migration matrix $\boldsymbol{F}(t)$; $\sum u_{VMiinPMk}(t)$ represents the CPU usage of server *k* at time *t*; f_h is the conversion coefficient between server CPU usage and power; $\boldsymbol{P}_{lc}(t)$ is the leakage power matrix of servers, which is a function of the server CPU operating temperature matrix $\boldsymbol{T}_{CPU}(t)$.

$$Q_{CDU,i}(t) = (Q_{central}(t) - \rho_{water} * V_{suplop} * C_{water} * \frac{dT_{suplop}}{dt} - \rho_{water} * V_{relop} * C_{water} * \frac{dT_{relop}}{dt}) * \varphi_i \tag{19}$$

$$Q_{PM,i,k}(t) = (Q_{CDU,i}(t) - \rho_{water} * V_{suplop,i} * C_{water} * \frac{dT_{suplop,i}}{dt} - \rho_{water} * V_{relop,i} * C_{water} * \frac{dT_{relop,i}}{dt}) * \varphi_{i,k} \tag{20}$$

$$EC_{cooling} = \int \left(PC_{central}(t) + \sum PC_{CDU}(t)\right) dt \tag{21}$$

Equations 18 - 20 demonstrate the matrix model framework of the liquid-cooled system (refer to Figure 2B(2) for the composition of the liquid-cooling system). $Q_{CDU,i}(t)$ is the real-time cooling of the *i* CDU (Cooling Distribution Unit) in the data center; $Q_{central}(t)$ is the real-time cooling of the cooling plant; T_{suplop} and T_{relop} are the average water temperature of the supply/return loop, respectively; V_{suplop} and V_{relop} are the water capacity of the supply/return loop, respectively; $\varphi_i$is the cooling distribution coefficient, representing the amount of cooling allocated by the cooling plant to the *i* CDU; $Q_{PM,i,k}(t)$ is the submodule where the *i* CDU is located, providing cooling for the *k* server; $V_{suplop,i}$ and $V_{relop,i}$are the water capacity of the supply/return loop (the submodule where the *i* CDU is located), respectively; $T_{suplop,i}$ and $T_{relop,i}$ are the average water temperature of the supply/return loop (the submodule where the *i* CDU is located), respectively; $\varphi_{i,k}$is the cooling allocation coefficient, representing the cooling allocated from the *i* CDU to the *k* server.

Nevertheless, this approach exhibits limitations in precisely characterizing the thermoelectric coupling behaviors of servers, along with the distribution characteristics of cooling. Although some researchers have investigated the thermoelectric coupling characteristics of specific server models and proposed corresponding modeling methods,[12,115] the performance of different servers varies significantly due to factors such as hardware architecture and manufacturing processes. Consequently, the universality of the aforementioned models remains to be studied. Furthermore, cooling system solutions and equipment configurations are complex, and there is currently a lack of a universal model to characterize its performance under all operating conditions. Additionally, the overall model of a data center encompasses multiple scales, including servers, computer rooms, and the entire system. The coupling process between models at different scales inevitably involves the issue of error accumulation, which also requires further research to address.

In summary, subsequent research should focus on the further development of a universal cross-scale thermoelectric coupling model for data centers, thereby laying a crucial foundation for collaborative energy saving optimization.

SUMMARY

This paper focuses on the collaborative energy saving optimization of IT and cooling systems in data centers. Starting from the thermoelectric coupling characteristics of data centers, it reviews commonly used energy efficiency evaluation indicators, design optimization measures, and operational optimization methods, and identifies existing research gaps, aiming to provide theoretical guidance for achieving overall optimal energy efficiency in data centers. The main contributions of this paper are as follows:

(1) Review of energy efficiency evaluation indicators for data centers

This paper conducts a comprehensive review of evaluation indicators, including PUE, PUE-derived indicators, waste heat utilization, renewable energy utilization, and carbon emissions. It emphasizes that the key factor preventing PUE and its derived indicators from guiding the collaborative energy saving optimization of IT and cooling systems is the neglect of the influence of thermoelectric coupling characteristics on energy efficiency.

(2) Review of design optimization for cooling systems

This paper reviews research on the scheme and selection optimization of water-based systems, fluorine-based systems, air-based systems, and various composite systems. It indicates that to lay the foundation for the collaborative energy saving optimization of IT and cooling systems, the following points should be noted during the design process of cooling systems: ① The optimal temperature for balancing the energy consumption (minimizing

Nomenclature	
IT	Information Technology
PUE	Power Usage Effectiveness
AI	Artificial Intelligence
AEER	Annual Energy Efficiency Ratio
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CDU	Coolant Distribution Unit
CER	Cooling Efficiency Ratio
CLF	Cooling Load Factor
CUE	Carbon Usage Effectiveness
COP	Coefficient of Performance
CPU	Central Processing Unit
CRAH	Computer Room Air Handler
DCiE	Data Center Infrastructure Efficiency
DC	Data Center
EER	Energy Efficiency Ratio
ERF	Energy Reuse Factor
ERE	Energy Reuse Effectiveness/Rate
GCOP	Gross Coefficient Of Performance (data-center cooling)
GPU	Graphics Processing Unit
GRU-LSTM	Gated Recurrent Unit-Long Short-Term Memory
HVAC	Heating, Ventilation and Air Conditioning
LNG	Liquefied Natural Gas
MLC	Mechanical Load Component
MPC	Model Predictive Control
OpenSSL	Open Secure Sockets Layer
PCM	Phase Change Material
PLF	Power Load Factor
PUE	Power Usage Effectiveness
pPUE	Partial Power Usage Effectiveness
RER	Renewable Energy Ratio
RPM	Revolutions Per Minute
VM	Virtual Machine
WUE	Water Usage Effectiveness

overall) of IT and cooling systems should be used as the basis for cooling design; ② Passive cold storage and release devices should be developed at the terminal side to enhance the cooling system's flexible adaptability to AI loads; ③ The matching characteristics between IT business, cooling system schemes, and outdoor environments should be quantified to determine the applicable scope of various cooling system schemes.

(3) Review of individual and joint optimization of IT and cooling systems

This paper reviews research on the individual and joint optimization between IT and cooling systems and identifies research gaps in two highly promising joint energy saving optimization methods: ① It is necessary to study temperature-related indicators that can accurately quantify the real-time leakage power of servers to achieve joint energy saving optimization based on optimal temperature; ② It is required to establish an offline training environment that accurately describes the dynamic characteristics of IT and cooling systems to achieve joint energy saving optimization based on reinforcement learning.

(4) Prospects for future research directions

Based on the aforementioned review, this paper further identifies two prospective research directions in the field of collaborative energy saving optimization: ① the "computational efficiency" mechanism and its corresponding evaluation indicators; ② a universal cross-scale thermoelectric coupling model.

REFERENCES

1. Nicola J. (2018). How to stop data centres from gobbling up the world's electricity. *Nature* **561**:163–166. DOI:10.1038/d41586-018-06610-y
2. Andrae A. and Edler T. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges* **6**:117–157. DOI:10.3390/challe6010117
3. Ministry of Industry and Information Technology of the People's Republic of China (Department of Information and Communication Development; General Office; MIIT Weibo) (2023). The 2023 China Computing Power Conference was held in

Yinchuan.
4. China Academy of Information and Communications Technology (2023). White Paper on Data Centers.
5. Xuan Y., (2023). Current status and challenges of data center energy management.
6. Wang S., Tu R., Chen X., et al. (2023). Thermal performance analyses and optimization of data center centralized-cooling system. *Appl. Therm. Eng.* **222**:119817. DOI:10.1016/j.applthermaleng.2022.119817
7. Wang Y., Song K., Han K., et al. (2025). Full-stack linkage study of high-density racks in data centers from a computing power–electric power–thermal power coordination perspective. *Sci. China Tech. Sci.* **55**:213–232. DOI:10.1360/SST-2024-0170
8. Lin J., Lin W., Huang H., et al. (2023). Thermal modeling and thermal-aware energy saving methods for cloud data centers: A review. *IEEE Trans. Sustain. Comput.* **9**:571–590. DOI:10.1109/TSUSC.2023.3346332
9. Cheng H., Liu B., Lin W., et al. (2021). A survey of energy-saving technologies in cloud data centers. *J. Supercomput.* **77**:13385–13420. DOI:10.1007/s11227-021-03805-5
10. Cao K., Li Z., Luo H., et al. (2024). Comprehensive review and future prospects of multi-level fan control strategies in data centers for joint optimization of thermal management systems. *J.Build. Eng.* **94**:110021. DOI:10.1016/j.jobe.2024.110021
11. Xu Y., Yan C., Liu H., et al. (2020). Smart energy systems: A critical review on design and operation optimization. *Sustain. Cities Soc.* **62**:102369. DOI:10.1016/j.scs.2020.102369
12. Zhu J., Liu D., Li B., et al. (2023). A data-center server power consumption model for temperature estimation based on CPU operating states. *Auto. Elect. Power Sys.* **47**:140–148.
13. Cheng H., Chen H., Shao S., et al. (2020). Comprehensive COP evaluation of data center cooling systems. *J. Refrig.* **41**:77–84.
14. Gusmann C., Bacon M. and Cassidy D. The Data Center Optimization Initiative (DCOI) and its Potential Effects on Government Contracting.
15. Implementation Plan for Promoting Green and High-Quality Development of Data Centers, 5G and Other New Infrastructure in Accordance with the Carbon Peaking and Carbon Neutrality Goals.pdf
16. Zhang X., Lindberg T., Xiong N., et al. (2017). Cooling Energy Consumption Investigation of Data Center IT Room with Vertical Placed Server. *Energy Procedia* **105**:2047-2052. DOI:https://doi.org/10.1016/j.egypro.2017.03.581
17. Zhang Z., Shao X., Song P., et al. (2022). Influencing factors and evaluation indicators of data center energy efficiency. *Heating Ventilating & Air Conditioning* **52**:148–156+199. DOI:10.19991/j.hvac1971.2022.03.24
18. Ren F. (2020). Optimization and application study of cooling-tower free cooling technology for data centers. M.S. thesis.
19. Information technology-Data centres-Key performance indicators-Part 2: Power usage effectiveness (PUE) : ISO/IEC30134-2: 2016/AMD 1: 2018. (2016). *Ceneva: InternationalOrganization for Standardization [ISO].*
20. Information technology-Data centre facilities and infrastruc-tures Part 4-7: Cooling efficiency ratio: BS EN 50600-47:2020 [S]. (2020). *London: European Committee for Electrotechnical Standardization.*
21. Data center cooling systems-Part 1: Overall coefficient of performance (GCOP) test and calculation method:T/CAR 9.1-2021[S]. (2021). Beijing: China Standards Press.
22. Tian Z. (2020). Theory and analysis of applying a novel evaporative-cooling air-conditioning system in data center computer rooms. M.S. thesis.
23. Puri R., Choudhury M., Qian H., et al. (2014). Bridging high performance and low power in processor design. *IEEE*. DOI:10.1145/2627369.2631642
24. Malik M. and Homayoun H. (2015). Big data on low power cores: Are low power embedded processors a good fit for the big data workloads? *IEEE*. DOI:10.1109/ICCD.2015.7357128
25. Ziegler M.M., Puri R., Philhower B., et al. (2014). POWER8 design methodology innovations for improving productivity and reducing power. *IEEE*. DOI:10.1109/CICC.2014.6946042
26. Bartolini S., Foglia P. and Prete C. A. (2018). Exploring the relationship between architectures and management policies in the design of NUCA-based chip multicore systems. *Future Gen. Comp. Sys.* **78**:481–501. DOI:10.1016/j.future.2017.06.001
27. Momeni M. and Shahhoseini H. S. (2022). Energy efficient 3D network-on-chip based on approximate communication. *Computer Networks* **203**:108652. DOI:10.1016/j.comnet.2021.108652
28. Huang Y., Zheng L., Yao P., et al. (2020). A heterogeneous PIM hardware-software co-design for energy-efficient graph processing. *IEEE*. DOI:10.1109/IPDPS47924.2020.00076
29. Karpowicz M. P., Arabas P. and Niewiadomska-Szynkiewicz E. (2018). Design and implementation of energy-aware application-specific CPU frequency governors for the heterogeneous distributed computing systems. *Future Gen. Comp. Sys.* **78**:302–315. DOI:10.1016/j.future.2016.05.011
30. Rethinagiri S.K., Palomar O., Sobe A., et al. (2015). ParaDIME: Parallel Distributed Infrastructure for Minimization of Energy. *IEEE*. **39**:1174-1189. DOI:10.1016/j.micpro.2015.06.005
31. Widjaja I., Walid A., Luo Y., et al. (2014). The importance of switch dimension for energy-efficient datacenter design. *Comp. commun.* **50**:152–161. DOI:10.1016/j.comcom.2014.02.017
32. Li D. (2024). Virtualization and energy management optimization of high speed computer network data centers based on optical switching and network technology. *Therm. Sci. Engi. Prog.* **55**:102918. DOI:10.1016/j.tsep.2024.102918
33. Wang B., Qi Z., Ma R., et al. (2015). A survey on data center networking for cloud computing. *Computer Networks* **91**:528–547. DOI:10.1016/j.comnet.2015.08.040
34. Xiao C., Zhang L., Liu W., et al. (2019). Energy-efficient crypto acceleration with HW/SW co-design for HTTPS. *Future Gen. Comp. Sys.* **96**:336–347. DOI:10.1016/j.future.2019.02.023
35. Handbook A. (1996). HVAC systems and equipment (chapter).
36. Wang W., Liang C., Zha F., et al. (2023). Air conditioning system with dual-temperature chilled water for air grading treatment in data centers. *Energy Build.* **290**:113073. DOI:10.1016/j.enbuild.2023.113073.
37. Zhang Y., Li H. and Wang S. (2024). Life-cycle optimal design and energy benefits of centralized cooling systems for data centers concerning progressive loading. *Renew. Energy* **230**:120847. DOI:10.1016/j.renene.2024.120847.
38. Xiang K., Tian Z., Zhou C., et al. (2025). Performance analysis and optimization of a hybrid mechanical and free cooling system with cold water storage for a data center. *Appl. Therm. Eng.* **267**:125831. DOI:10.1016/j.applthermaleng.2025.125831.
39. Zhao C., Jing Y., Xie X., et al. (2025). Application and simulation of IEC water chillers for summer free cooling in data centers. *Appl. Therm. Eng.* **260**:124915. DOI:10.1016/j.applthermaleng.2024.124915.
40. Wang H., Yuan X., Zhang K., et al. (2024). Performance evaluation and optimization of data center servers using single-phase immersion cooling. *Int. J. Heat Mass Trans.* **221**:125057. DOI:10.1016/j.ijheatmasstransfer.2023.125057
41. Zhang Y., Zhao Y., Dai S., et al. (2022). Cooling technologies for data centres and telecommunication base stations – A comprehensive review. *J. Clean. Product.* **334**:130280. DOI:10.1016/j.jclepro.2021.130280
42. Hnayno M., Chehade A., Klaba H., et al. (2023). Experimental investigation of a data-centre cooling system using a new single-phase immersion/liquid technique. *Case Studies Therm. Eng.* **45**:102925. DOI:10.1016/j.csite.2023.102925
43. Li J. (2023). Design and energy-saving study of liquid-cooling heat dissipation systems for data centers. M.S. thesis.
44. Muneeshwaran M., Lin Y.-C. and Wang C.-C. (2023). Performance analysis of single-phase immersion cooling system of data center using FC-40 dielectric fluid. *Int. Commun. Heat Mass Transf.* **145**:106843. DOI:10.1016/j.icheatmasstransfer.2023.106843
45. Ebrahimi K., Jones G. F. and Fleischer A. S. (2015). Thermo-economic analysis of steady state waste heat recovery in data centers using absorption refrigeration. *Appl. Energy* **139**:384–397. DOI:10.1016/j.apenergy.2014.10.067
46. Haywood A., Sherbeck J., Phelan P., et al. (2012). Thermodynamic feasibility of harvesting data center waste heat to drive an absorption chiller. *Energy Conver. Manag.* **58**:26–34. DOI:10.1016/j.enconman.2011.12.017
47. Pan Q., Peng J. and Wang R. (2021). Application analysis of adsorption refrigeration system for solar and data center waste heat utilization. *Energy Conver. Manag.* **228**:113564. DOI:10.1016/j.enconman.2020.113564
48. Du S., Cui Z., Wang R. Z., et al. (2024). Development and experimental study of a compact silica gel-water adsorption chiller for waste heat driven cooling in data centers. *Energy Conver. Manag.* **300**:117985. DOI:10.1016/j.enconman.2023.117985.
49. Hao T. (2023). Performance and mechanism study of treating recirculating cooling water using electrochemical advanced oxidation. M.S. thesis (North China Electric Power University).
50. Zhou Y. (2024). Study on the scale-inhibition effects of microbial agents in reclaimed-water reuse recirculating cooling water. M.S. thesis (North China Electric Power University).
51. Liu Q. (2025). Development of a high-energy-efficiency electrochemical descaling technology system for recirculating cooling-water systems and its reaction mechanisms. M.S. thesis (Qingdao University of Technology).
52. Liu H. (2017). Design and energy-efficiency analysis of a data-center heat dissipation system based on separated heat pipes. M.S. thesis.
53. Qian X., Li Z. and Li Z. (2012). Experimental study on a heat-pipe air-conditioning system for data rooms. *J. Eng. Thermophys.* **33**:1217–1220.
54. Zou S. (2022). Heat-transfer performance and energy-saving study of a separated heat-pipe cooling system for data rooms. Ph.D. dissertation.
55. Wang N., Guo Y., Yu H., et al. (2024). Investigation on performance enhancement of micro-channel separated heat pipe in data center: A coupled heat-mass-flow characterization approach. *Appl. Therm. Eng.* **248**:123327. DOI:10.1016/j.applthermaleng.2024.123327
56. Li L., Tong Z., Liu X., et al. (2023). Review of applications of separated heat pipes in the HVAC field. *Heating Ventilating & Air Conditioning* **53**:1–12. DOI:10.19991/j.hvac1971.2023.08.01
57. Gong B. (2024). Performance study of a water-cooled microchannel separated fluorocarbon-pump heat-pipe system for data centers. M.S. thesis.
58. Vasta S., Palomba V., La Rosa D., et al. (2020). Experimental assessment and numerical study of a pump-assisted loop heat pipe for high capacity thermal systems. *Appl. Therm. Eng.* **180**:115828. DOI:10.1016/j.applthermaleng.2020.115828.
59. Li X., Zhang C., Dong J., et al. (2023). Feasibility investigation on a novel rack-level cooling system for energy-saving retrofit of medium-and-small data centers. *Appl. Therm. Eng.* **229**:120644. DOI:10.1016/j.applthermaleng.2023.120644
60. Lin Y. (2020). Experimental study of a hybrid air-conditioning system with fluorocarbon-pump boosting. M.S. thesis.
61. Mehta K., Mehta N. and Patel V. (2020). Influence of the channel profile on the thermal resistance of closed-loop flat-plate oscillating heat pipe. *J. Brazilian Soc. Mechan. Sci. Eng.* **42**:123. DOI:10.1007/s40430-020-2213-x

62. Chen M. (2024). Heat transfer and flow characteristics of oscillating heat pipes under non-uniform heating conditions. M.S. thesis.
63. Liu X. (2024). Flow and heat transfer performance of oscillating heat pipes with novel irregular-channel structures. M.S. thesis.
64. Lu Q. (2016). Study on a data-center rack cooling system based on oscillating heat pipes. M.S. thesis.
65. Chen C. (2020). Design and experimental study of a variable-frequency multi-split precision air-conditioning system for data rooms with free cooling. M.S. thesis.
66. Yang Z., Fan C. and Liao Y. (2025). Applicability and influencing-factor analysis of evaporative pre-cooling systems for data centers in hot-humid climates. *Refrig. Air Condi.* **25**:27–33. DOI:10.20245/j.issn.1009-8402.2025.03.005
67. Huang X., Yang L., Chu J., et al. (2025). Energy-saving characteristics of an evaporative-condensation and fluorocarbon-pump heat-pipe air-conditioning system. *Refrig. Air Condi.* **25**:40–43. DOI:10.20245/j.issn.1009-8402.2025.03.007
68. Bai K., Ma G., Zhou F., et al. (2016). Performance characteristics of a year-round hybrid cooling system combining pump-driven loop heat pipes and mechanical refrigeration. *Heat. Ventilat. Air Condi.* **46**:109–115. DOI:10.19991/j.hvac1971.2016.09.024
69. Wang F., Wang J., Shi Z., et al. (2017). Research and experiments on heat-pipe hybrid precision air-conditioning for data rooms. *Refrig.Air Condi.* **17**:37–41.
70. Han Z., Ji Q., Wei H., et al. (2020). Simulation study on performance of data center air-conditioning system with novel evaporative condenser. *Energy* **210**:118521. DOI:10.1016/j.energy.2020.118521
71. Siriwardana J., Jayasekara S. and Halgamuge S. K. (2013). Potential of air-side economizers for data center cooling: A case study for key Australian cities. *Appl. Energy* **104**:207–219. DOI:10.1016/j.apenergy.2012.10.046
72. Oró E., Depoorter V., Pflugradt N., et al. (2015). Overview of direct air free cooling and thermal energy storage potential energy savings in data centres. *Appl. Therm. Eng.* **85**:100–110. DOI:10.1016/j.applthermaleng.2015.03.001
73. Ham S.-W., Kim M.-H., Choi B.-N., et al. (2015). Energy saving potential of various air-side economizers in a modular data center. *Appl. Energy* **138**:258–275. DOI:10.1016/j.apenergy.2014.10.066
74. Yu Y. and Wang L. (2019). Solid sorption heat pipe coupled with direct air cooling technology for thermal control of rack level in internet data centers: Design and numerical simulation. *International Journal of Heat and Mass Transfer* **145**:118714. DOI:10.1016/j.ijheatmasstransfer.2019.118714
75. Kim J.-Y., Chang H.-J., Jung Y.-H., et al. (2017). Energy conservation effects of a multi-stage outdoor air enabled cooling system in a data center. *Energy Build.* **138**:257-270. DOI:10.1016/j.enbuild.2016.12.057
76. Zou S. and Pan Y. (2024). Performance of a hybrid thermosyphon cooling system using airside economizers for data center free cooling under different climate conditions. *J. Build. Eng.* **98**:111235. DOI:10.1016/j.jobe.2024.111235
77. Wang J., Cao X., Yuan Y., et al. (2025). Experimental study on free cooling performance of an integrated cooling system based on loop thermosyphon for data centers. *Energy* **322**:135706. DOI:10.1016/j.energy.2025.135706
78. Liu C., Hao N., Zhang T., et al. (2024). Optimization of data-center immersion cooling using liquid air energy storage. *J. Energy Storage* **90**:111806. DOI:10.1016/j.est.2024.111806
79. Ting Y., Zhen P., Zhenbo L., et al. (2025). Design, analysis, and multi-objective optimization of LNG cold energy cascade utilization CCHP system including ORC, cold storage, and data center cooling. *Cryogenics* **150**:104152. DOI:10.1016/j.cryogenics.2025.104152
80. Yuan X., Liang Y., Hu X., et al. (2023). Waste heat recoveries in data centers: A review. *Renew. Sustain. Energy Rev.* **188**:113777. DOI:10.1016/j.rser.2023.113777
81. Weng Y.-C., Cho H.-P., Chang C.-C., et al. (2011). Heat pipe with PCM for electronic cooling. *Appl. Energy* **88**:1825-1833. DOI:10.1016/j.apenergy.2010.12.004
82. Demir M. E. and Erden H. S. (2025). Effective energy use and passive cooling in data centers using heat pipes and PCM-integrated TEG systems. *J. Energy Storage* **122**:116625. DOI:10.1016/j.est.2025.116625
83. Data Center Storage Equipment Thermal Guidelines, Issues, and Best Practices. (2015).
84. Zapater M., Ayala J.L., Moya J.M., et al. (2013). Leakage and temperature aware server control for improving energy efficiency in data centers. *IEEE.* DOI:10.7873/DATE.2013.067
85. Moshe J. (2023). Device for cooling data center computer servers using phase change materials. *Patent.* DOI:CA3265979A1
86. Zhou E. and Pan Y. (2024). Liquid-cooled server heat dissipation device coupling a composite cold plate with phase-change thermal storage.
87. Li Y., Mughees M., Chen Y., et al. (2024). The unseen AI disruptions for power grids: LLM-induced transients. *arXiv preprint arXiv:2409.11416.* DOI:10.48550/arXiv.2409.11416
88. Feng D. (2021). Low-energy workflow scheduling method based on merge allocation in cloud environments. M.S. thesis.
89. Cui Y. (2023). Research on energy-efficient task scheduling strategies and performance optimization in heterogeneous cloud environments. Ph.D. dissertation.
90. Han X. M., Wang J. X., Wu J. X., et al. (2025). Energy-efficient cloud systems: Virtual machine consolidation with G-robustness optimization. *iScience* **28**:111897. DOI:10.1016/j.isci.2025.111897
91. Kim Y. G., Kim S. Y., Choi S. H., et al. (2021). Thermal-aware adaptive VM allocation considering server locations in heterogeneous data centers. *J. Sys. Archi.* **117**. DOI:10.1016/j.sysarc.2021.102071
92. Dabbagh M., Hamdaoui B., Guizani M., et al. (2018). An Energy-Efficient VM Prediction and Migration Framework for Overcommitted Clouds. *IEEE Trans. Cloud Comput.* **6**:955–966. DOI:10.1109/tcc.2016.2564403
93. Wu J. (2022). Research and system implementation of energy-consumption-aware virtual machine scheduling strategies in cloud computing environments. M.S. thesis.
94. Xu J. (2022). Research and implementation of load-prediction-based elastic scaling schemes in cloud environments. M.S. thesis.
95. Zapater M., Ayala J.L., Moya J.M., et al. (2013). Leakage and temperature aware server control for improving energy efficiency in data centers. *DATE.* DOI:10.7873/DATE.2013.067
96. Li X., Garraghan P., Jiang X., et al. (2018). Holistic Virtual Machine Scheduling in Cloud Datacenters towards Minimizing Total Energy. *IEEE TPDS.* **29**:1317–1331. DOI:10.1109/tpds.2017.2688445
97. Chen R., Liu B., Lin W., et al. (2023). Power and thermal-aware virtual machine scheduling optimization in cloud data center. *Future Gen. Comput. Sys.* **145**:578–589. DOI:10.1016/j.future.2023.03.049
98. Li X., Li M., Zhang Y., et al. (2024). Rack-level cooling technologies for data centers – A comprehensive review. *J. Build. Eng.* **90**. DOI:10.1016/j.jobe.2024.109535
99. Lin J., Lin W., Lin W., et al. (2024). Multi-objective cooling control optimization for air-liquid cooled data centers using TCN-BiGRU-Attention-based thermal prediction models. *Build. Simul.* **17**:2145–2161. DOI:10.1007/s12273-024-1185-7
100. Zhang Q., Zheng H., Zhu Y., et al. (2024). Optimization of data center cooling energy consumption based on mixed-integer programming. *J. Hunan Uni.* **51**:188–197. DOI:10.16339/j.cnki.hdxbzkb.2024099
101. Qu S., Duan K., Guo Y., et al. (2024). Real-time optimization of the liquid-cooled data center based on cold plates under different ambient temperatures and thermal loads. *Appl. Energy* **363**. DOI:10.1016/j.apenergy.2024.123101
102. Jin C., Bai X., Yang C., et al. (2020). A review of power consumption models of servers in data centers. *Applied Energy* **265**:114806–114806. DOI:10.1016/j.apenergy.2020.114806
103. Xue P. N., Shi L. Y., Zhou Z. G., et al. (2024). An online fault detection and diagnosis method of sensors in district heating substations based on long short-term memory network and adaptive threshold selection algorithm. *Energy Build.* **308**:22. 114009,DOI:10.1016/j.enbuild.2024.114009.
104. 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. *ICDCS.* DOI:10.1109/ICDCS.2019.00070
105. Alibaba. (2020). cluster-trace-v2020[EB/OL]. https://github.com/alibaba/clusterdata/tree/master/cluster-trace-gpu-v2020
106. Ham S.-W., Kim M.-H., Choi B.-N., et al. (2015). Simplified server model to simulate data center cooling energy consumption. *Energy Build.* **86**:328–339. DOI:10.1016/j.enbuild.2014.10.058
107. Gupta R., Asgari S., Moazamigoodarzi H., et al. (2021). Energy, exergy and computing efficiency based data center workload and cooling management. *Appl. Energy* **299**:117050. DOI:10.1016/j.apenergy.2021.117050
108. Zhou B., Song X., Shi X., et al. (2019). Bridging Server and Cooling: Toward Effective Energy Management in Data Centers. *ICPADS.* DOI:10.1109/ICPADS47876.2019.00015
109. Lin J., Lin W., Huang H., et al. (2024). Thermal Modeling and Thermal-Aware Energy Saving Methods for Cloud Data Centers: A Review. *IEEE TSC* **9**:571–590. DOI:10.1109/tsusc.2023.3346332
110. Zhou X., Wang R., Wen Y., et al. (2021). Joint IT-Facility Optimization for Green Data Centers via Deep Reinforcement Learning. *IEEE Network* **35**:255–262. DOI:10.1109/MNET.011.2100101
111. Chi C., Ji K., Song P., et al. (2021). Cooperatively improving data center energy efficiency based on multi-agent deep reinforcement learning. *Energies* **14**:2071. DOI:10.3390/en14082071
112. Xu Y., Tian M., Yan C., et al. (2025). A data center energy efficiency optimization method based on optimal temperature control of designated active servers. *Energy Build.* **345**:116126. DOI:10.1016/j.enbuild.2025.116126
113. Guidelines for Shanghai Intelligent Computing Centers (2025).pdf
114. Han H. and Ke Q. (2022). On the optimal granularity for improving data center efficiency. *Inf. Commun. Techn. Policy.*47-54. DOI:Cnki:Sun:Dxwj.0.2022-03-009
115. Zhou H., Zhu J., Zhang J., et al. (2025). An intelligent virtual-machine scheduling framework based on a server CPU temperature–power model. *Elect. Power Auto. Equip.*:1-13. DOI:10.16081/j.epae.202507027

FUNDING AND ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the project “Research and Application of Key Technologies for EPC Design of Data Centers” (AZ-2025-12).

DECLARATION OF INTERESTS

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

Readers may contact the corresponding authors for data upon reasonable request.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/ipj.energy-use.2025.100020