The Journal of Supercomputing
https://doi.org/10.1007/s11227-021-03805-5

A survey of energy-saving technologies in cloud data centers

Huiwen Cheng, et al. ***[full author details at the end of the article]***

Accepted: 8 April 2021

Abstract
As an important part of the new infrastructure, the cloud data center is developing rapidly, and its energy consumption problem is becoming more and more prominent. Therefore, research on energy-saving technologies for cloud data centers has attracted widespread attention from academia and industry. Some studies have reviewed energy-saving optimization methods and technologies for data centers, but recently, many state-of-the-art optimization methods of energy consumption and energy-saving technologies have sprung out, which are still worth analyzing and discussing. Depending on the in-depth investigation and analysis of related research status, this article firstly focuses on analyzing and discussing the energy-saving technologies of the two components: IT equipment and cooling systems, both of which bring about the largest energy consumption in cloud data centers. As for IT equipment, its energy-saving technologies mainly include the energy saving of servers, storage systems, and network systems. While as for cooling systems, airflow organization in the computer room, thermal-aware scheduling technology, and other new energy-saving technologies are involved. Secondly, on the basis of analyzing the energy-saving technologies of the two major components, a new optimization scheme of energy consumption for the jointing computing system and cooling system is explained. Throughout this work, various energy-saving strategies and technologies have been described and compared. Finally, the future trends and development directions of energy saving for data centers are further promoted, such as integral optimization of energy consumption jointing multiple components, energy saving using artificial intelligence methods, energy saving based on novel hardware equipment, hybrid cooling energy saving, and comprehensive energy conservation with various energy technologies.

Keywords Data centers · Energy conservation · Optimization of energy consumption · Server electronics · Cooling systems

1 Introduction

In recent years, cloud computing, big data, Internet of Things (IoT), and other technologies have developed rapidly. Cloud computing realizes the on-demand deployment of resources through virtualization technology. Cloud users can quickly allocate computing power and resources according to their needs at any time without worrying about the underlying physical structure. The cloud computing platform provides information resources in a pay-as-you-go model. Cloud computing business has exploded in recent years. As the physical platform and infrastructure of cloud computing, the amount of data and business carried by the data center are also increasing day by day. IDC predicted that from 2018 to 2025, the amount of data stored by enterprises will increase by 61% to 175ZB, most of which will be stored in data centers [1]. The growth of massive data is accompanied by the energy consumption of data centers. According to a report, the energy usage of the world's data centers has doubled in the past ten years and will increase three or even four times in the next ten years [2]. The huge energy consumption generated by data centers has brought some problems to the society. On the one hand, it brings more energy consumption expense to data center operators, thus increasing the cost. The study has shown that the power consumption of data centers in 2017 accounted for about 2% of the total energy consumption, and it is predicted to reach 5% by 2024 [3]. On the other hand, the large amount of energy consumption generated by data centers has brought heavy environmental pressure to the whole society. The burning of fossil fuels has led to a sharp increase in carbon emissions, which are still the main source of electricity. According to statistics, the average construction cost of building a 100 MW power station is between \$60 million and \$100 million, and co_2 emissions are at least 50 million tons over the life cycle [4]. In this case, it is necessary to seek efficient energy-saving methods to enable the data center to reduce energy consumption while satisfying high-performance computing.

The energy consumption of a cloud data center depends on many aspects. As shown in Fig. 1 [5], it mainly includes the following parts: IT equipment, cooling systems and other infrastructures. Among them, IT equipment includes servers,

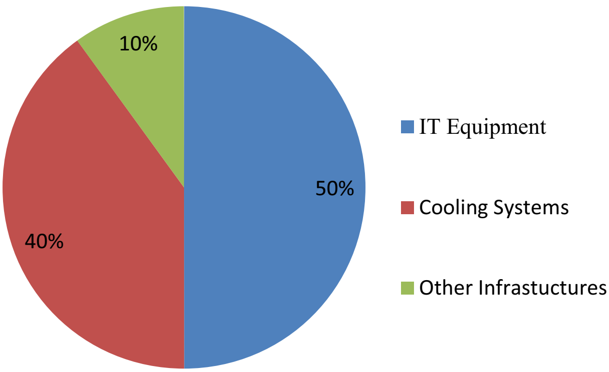

Fig. 1 Data center energy consumption composition

the network system, the storage system, etc.; the cooling systems include computer room air conditioner (CRAC), compressors, fans, etc.; other infrastructure facilities include power supply systems, lighting systems, etc. In the total energy consumption of the data center, the energy consumption of IT equipment is the largest, accounting for about half of the total energy consumption of the entire data center, of which the energy consumption of the server accounts for 40%, the memory system and network system both consume about 5%. In the past few years, the energy consumption of cooling systems has gradually approached or even exceeded the energy consumption of IT equipment [5]. Therefore, the energy saving of the data center focuses on the energy saving of IT equipment and cooling systems. The PUE is currently an energy efficiency index of data centers which is widely recognized by the industry. $PUE=$total energy consumption/IT equipment energy consumption. The value is greater than 1, and the closer it is to 1, the less energy consumption of non-IT equipment, that is, the higher the energy efficiency.

In order to deal with the problems of high energy consumption and low energy efficiency in data centers, researchers have explored efficient energy-saving solutions. Energy-saving strategies can help reduce the huge energy consumption of data centers, which can reduce costs and are also conducive to environmental protection. Due to the extremely complex environment of the data center, the energy management of the data center has also become complicated. The energy consumption optimization strategy of the data center can start from various aspects. A variety of energy-saving methods and technologies have emerged so far, some of which are aimed at physical devices, and some used virtual machine (VM) technology to save energy. Although there have been a lot of studies that have reviewed the energy-saving methods of data centers from different aspects, there is still a lack of comprehensive investigation. To our knowledge, most of the relevant surveys focus on the power consumption of a single component. On the one side, some papers have studied the energy optimization technology of computing systems [6–9], but they have not considered the energy consumption of the cooling system and other systems. On the other side, some surveys have tried to reduce energy consumption through thermal management of data centers [10–13]. These works focus on the cooling energy consumption of data centers. They did not consider the energy consumption of IT equipment. In addition, some works tried to explain the energy consumption of the data center in a special way. Mastelic T et al. [14] provided a overview of data center ICT equipment with regards to energy efficiency. Malla S et al. [15] did a survey on power management techniques for oversubscription of multi-tenant data centers. Since there are many sources of energy consumption in data centers, in recent years, researchers have explored the energy consumption optimization of multiple components. They modeled the total energy consumption of the data center as a constraint problem or other mathematical problems and designed different algorithms to get the global optimum solution. Zhang et al. [16] investigated and summarized the joint optimization work of the data center. They envisioned that the learning-based approach can be a promising framework for the joint ICT and cooling management in the data center, from data profiling, learning, optimization to execution. In this regard, some heuristic intelligent algorithms are often used to solve joint scheduling problems. In recent years, intelligent optimization algorithms have made contributions to solving the optimal combination problem. For example, simulated annealing algorithm (SA) in [82], CRUZE algorithm based on Cuckoo algorithm in [83]. The energy consumption problem of data centers can be approximated as an optimization problem with constraints. Therefore, in the future, intelligent optimization algorithms can be used to solve the joint energy consumption optimization problem of data centers.

Unlike the above, this survey paper gives an overview of energy-saving strategy in data centers. We classify the energy-saving strategies into two aspects: the energy optimization methods of IT equipment and energy optimization methods of cooling systems (many of which are the most cutting-edge). Then we investigate the optimization methods of these two aspects to optimize the energy efficiency of the data center. On this basis, we review the latest joint optimization methods. Our innovation mainly lies in the perspective on topic selection. From the perspective of system components, we cover a wide range of the latest solutions and absorb the latest research results in this field. It is believed the survey can be helpful for further research on the energy conservation of cloud data centers in the future.

The main contributions of this survey are shown as follows:

(1) The energy-saving technologies of cloud data centers have been comprehensively introduced. From the perspective of system components, we discuss the existing energy-saving solutions for computing systems and cooling systems.
(2) We classify and summarize the energy-saving technologies of the two most important components in the cloud data center. Especially for the energy saving of the cooling system, this survey has discussed new technologies and methods that can reduce economic benefits as well as saving energy and environmental protection.
(3) By introducing the existing energy-saving technologies based on the computing system and cooling system, an energy-saving plan for jointly optimizing IT equipment and cooling equipment is obtained, and the existing research progress of joint energy-consuming optimization is discussed. Through the analysis and summary of existing research, we discuss the future directions of cloud data center energy saving at the end.

The rest of this paper is organized as follows. In Sect. 2, we focus on research on energy conservation of IT equipment in data centers. Section 3 summarizes and discusses the energy optimization research of cooling systems. The optimization study of the joint the computing system and cooling system is given in Sect. 4, followed by Sect. 5, which discusses the future research directions and prospects. Finally, we conclude the paper in Sect. 6. The organizational structure of our work is shown in Fig. 2.

2 Energy saving for IT equipment

2.1 Energy consumption optimization methods for servers

The data center is a strategic resource as important as human resources and natural resources that calculates and stores a large amount of data on the network. It is

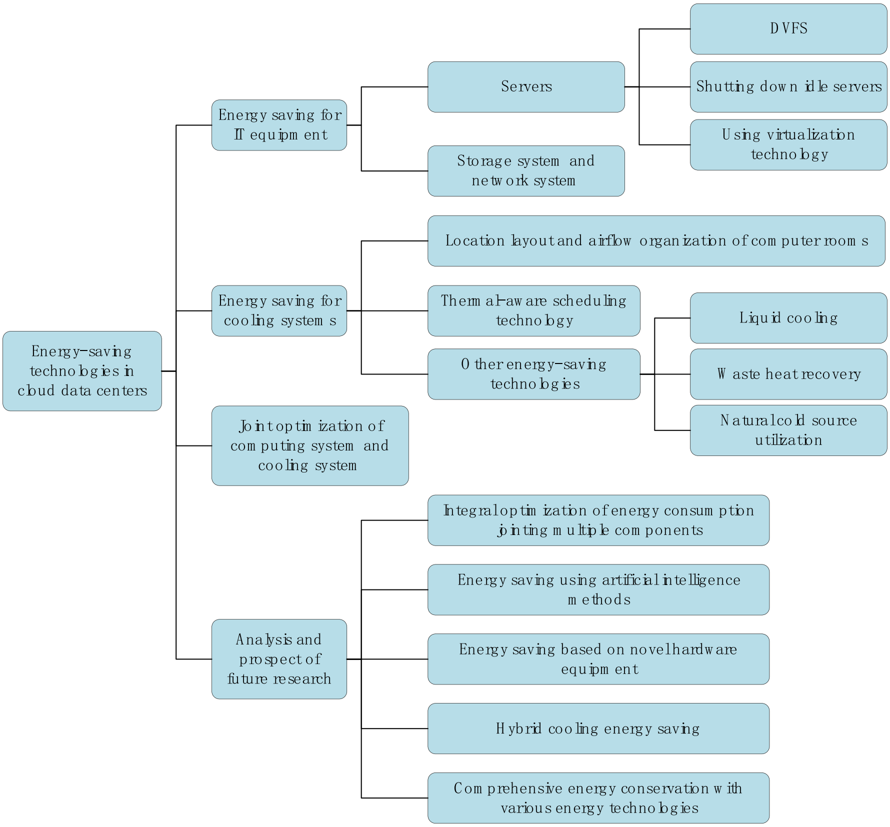

Fig. 2 Overview of energy-saving technologies for cloud data centers

a large-scale infrastructure with hundreds or even thousands of IT servers running around the clock. The servers are the core component of data centers which handle all computing services on the Internet, making it inevitable to generate energy consumption. However, unnecessary energy loss can be reduced through appropriate technical means. For server energy-saving technologies, academia and industry have conducted in-depth research. These technologies can be roughly divided into three aspects: dynamic voltage and frequency scaling (DVFS) technology, shutting down idle servers, and using virtualization technology.

2.1.1 DVFS

DVFS technology dynamically adjusts the operating frequency and voltage of CPU according to the application's requirements for different computing capabilities to reduce energy consumption. It realizes energy saving according to the relationship between the energy consumption of CMOS circuit and the square of voltage and frequency. DVFS is currently an optimization technology commonly used in industry and academia. In large computing scenarios, such as cluster computing in data centers, DVFS technology has been applied to reduce power consumption and save large amounts of expenses for enterprises.

Kyong et al. [17] used DVFS technology to propose three data center energy-saving solutions, namely Lowest-DVFS, δ-Advanced-DVFS and Adaptive-DVFS. Lowest-DVFS adjusts the processor speed to the lowest speed required by the task, which means that the VM executes the task at the minimum speed required. This solution has the lowest energy consumption but reduces the performance of the data center to a certain extent. In order to overcome this defect, the authors proposed an improved Lowest-DVFS called δ-Advanced-DVFS, which increases the execution speed by $\delta\%$, and can adjust the processor scale. The Adaptive-DVFS solution obtains the optimal scale according to the service request arrival rate and service time and adjusts the CPU parameters according to the load situation. Among these three schemes, Lowest-DVFS has the best effect in reducing energy consumption, but it does not meet Quality of Service (QoS). The Adaptive-DVFS scheme has good adaptability and can effectively reduce the level of energy consumption.

Some studies have studied task scheduling algorithms based on DVFS technology, which improve system resource utilization by adjusting calculation frequency and performing reasonable scheduling. The scheduling algorithm designed in [18] considered the frequency of maximum task F_{max} and minimum task F_{min}, as well as multiple server Si running at maximum $Si\ F_{max}$ and minimum $Si\ F_{min}$ frequencies. For a specific task, the scheduling algorithm effectively allocates an appropriate server for the job according to the requirements of the task frequency. Gu et al. [19] studied a data center management and a request scheduling algorithm based on DVFS, which uses the joint optimization of task deployment and frequency adjustment to minimize power costs. The authors also considered the geographical distribution of different data centers and the difference in electricity prices. This algorithm tries to activate as many servers in areas with low electricity prices as possible, and shuts down servers in areas with high electricity prices. Tang et al. [20] proposed an energy-efficient workflow task scheduling algorithm (DEWTS) based on DVFS. By recycling the slack time to merge relatively inefficient processors, the DEWTS algorithm can use DVFS technology to allocate parallel applications in the workflow to the appropriate processors within a given deadline, so as to meet the performance and reduce energy consumption at the same time.

Unlike the previous offline scheduler, Ge et al. [21] described an online scheduler named CPU MISER based on DVFS technology. Figure 3 shows the system function implementation of CPU MISER, which is a DVFS online scheduler for multi-core or power-aware based on symmetric multiprocessing. In CPU MISER, there are three important roles, namely performance monitor, workload predictor and DVFS scheduler. The performance monitor uses hardware counters to periodically collect performance events during each interval. The workload predictor predicts the next time workload by calculating the data obtained by the performance detector, and then the DVFS scheduler decides the target frequency of each processor and performs scheduling based on the load prediction.

In order to derive a general method to design an automatic, system-wide DVFS scheduler, the authors described the DVFS scheduling problem as: given an

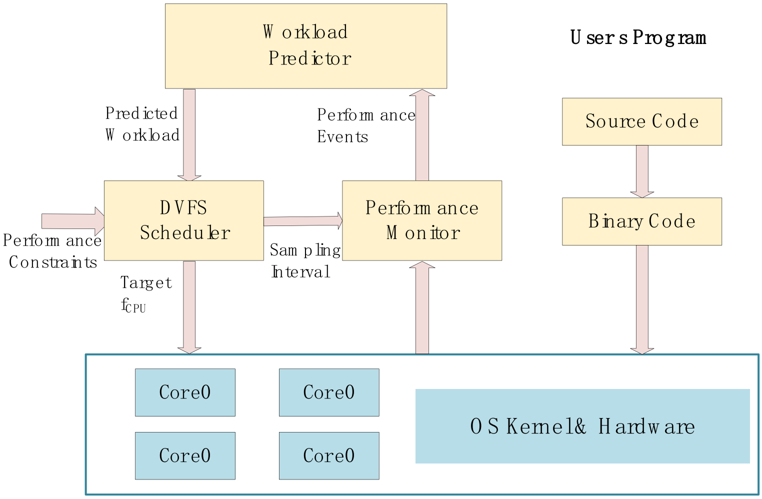

Fig. 3 The implementation of CPU MISER

energy-aware system and a workload *W*, schedule a sequence of CPU frequencies over time that is guaranteed to finish computing the workload within a time duration $(1 + \delta *) \cdot T$ and minimize the total energy. Where $\delta *$ is the user-specified performance loss constraint, and *T* is the execution time when the system runs continuously at the highest frequency f_{max}. The authors modeled this problem, which captured the correlation between workload, frequency, and performance loss due to frequency scaling. In this performance model, the standardized performance loss function is defined as follows:

$$\delta(f) = \frac{t' - t}{t} \tag{1}$$

where *t'* is the time required to complete the same workload when the system is running at a lower frequency *f*, and *t* is the running time of the system. They are calculated as follows:

$$t = \alpha \cdot w_{on} \cdot \frac{CPI_{on}}{f_{\max}} + w_{mem} \cdot t_{mem} + tIO + t0 \tag{2}$$

$$t' = \alpha \cdot w_{on} \cdot \frac{CPI_{on}}{f} + w_{mem} + t_{mem} + tIO + t0 \tag{3}$$

here, w_{on} is the number of on-chip memory (including register and on-chip cache) accesses, and CPU_{on} is the average cycles per on-chip access, w_{mem} and t_{mem} are, respectively, defined as the number of main memory accesses and the main memory access time, t_{IO} is the IO access time,t_0 is the time when the system is idle. Since on-chip access is often overlapped with off-chip access on modern computer architecture, the authors introduced an overlapping factorα ($0 \leq \alpha \leq 1$) into (2). Substituting formula (2) and (3) into formula (1) we can see:

$$\delta(f) = \left(\alpha \cdot w_{on} \cdot \frac{CPI_{on}}{f_{\max}} \right) \cdot \frac{1}{t} \cdot \frac{f_{\max} - f}{f} \tag{4}$$

Equation (4) indicates that performance loss is determined by both the processor frequency and workload characteristics. The workload characteristics is defined as:

$$k = \left(\alpha \cdot w_{on} \cdot \frac{CPI_{on}}{f_{\max}} \right) \cdot \frac{1}{t} \tag{5}$$

k is considered as an index of CPU intensiveness. When $k = 1$, it means that the CPU load and the system are computationally intensive. When $k \approx 0$, the system is either idle or memory and IO load.

The above studies are based on DVFS technology to achieve energy saving in data centers, which proved that DVFS technology is an effective method to reduce server energy consumption, but these studies have not taken into account the communication overhead and memory consumption in the actual heterogeneous environment.

2.1.2 Shutting down idle servers

Shutting down idle servers refers to assigning tasks to as few servers as possible and turning off low-load servers to reduce energy consumption. It is reported that the server will still generate a lot of energy consumption under low load or even idle conditions [22]. Therefore, shutting down idle or low-load servers can effectively reduce data center energy consumption, especially when the data center load is predicted. But frequent opening and closing operations of servers will reduce the performance of the server. It is important to consider the balance of system availability and energy consumption when shutting down idle servers.

There are many studies on shutting down idle servers. Due to the characteristics of real-time changes in application load, directly shutting down low-load servers is less efficient. For this reason, many studies used artificial intelligence methods to predict future workload demands and operate low-load servers on this basis. Berral et al. [23] used machine learning (ML) to predict the power consumption level after load migration. The authors changed the schedule based on the predicted results to achieve smarter and more effective scheduling. This method can effectively reduce the energy consumption of the data center. However, the authors did not consider the time cost and energy cost caused by the migration workload. Similarly, Duy et al. [24] used neural network (NN) to predict the future load of servers, and designed a green scheduling algorithm based on the predicted results. The algorithm shut down unused servers and minimized the number of running servers, it could reduce energy consumption by about 46.7%.

Different from [23, 24], Beloglazov et al. [25] designed an overall optimization framework. The system is mainly divided into four main functional modules: (1) consumption/agent; (2) green service distributor: acting as the interface between cloud infrastructure and users, interacting with energy-saving schedulers, service analyzers, energy monitors, service schedulers, VM managers and other components to achieve energy-saving scheduling; (3) VM: the virtual resource of the cloud service. By dynamically migrating virtual machines (VMs) to physical machines (PMs), users can consolidate workloads, switch unused resources to low-power mode, and shut down VMs running at low-performance levels to save energy consumption; (4) physical machines: hardware infrastructure for computing. In addition, the authors also proposed an energy-aware allocation heuristic method for reasonable scheduling of VMs. According to the proposed energy consumption model, the scheduling algorithm was used to obtain the VM allocation method with the least energy consumption. The simulation results show that compared with static resource allocation technology, this method can significantly reduce the energy consumption of cloud data centers.

Meisner et al. [26] discussed a method to reduce server idle power consumption by quickly switching ultra-low power consumption states, called PowerNap. In order to improve the energy efficiency of the processor during sleeping and minimize the transition time of entering and exiting the sleeping state, the authors designed the entire system based on PowerNap. The whole system quickly transits between high load state and idle state. Since the power supply efficiency of current servers decreases as server utilization decreases, PowerNap operated in areas where the power supply efficiency of current blade server centers was low. To solve this problem, the authors introduced a redundant array (RAILS) to realize cheap load sharing. This power supply method can provide high conversion efficiency within the entire power demand range of PowerNap. The overall concept of the system is shown in Fig. 4 below.

Figure 4 illustrates the concept of PowerNap. Whenever the server completes all work, it transits to the Nap state. In this state, the power consumption is very low, and the system components do not perform any processing. When a new job arrives, the system will wake up and switch back to the active state. When the work is completed again, the system returns to the Nap state. The authors used utilization tracking collected from commercial deployments, and experiments show that PowerNap and RAILS can reduce energy consumption of servers by about 74% in total.

2.1.3 Using virtualization technology

As mentioned earlier, the extremely high energy consumption and low resource utilization pose challenges to the operation of the data center. In view of this situation, by using virtualization technology, the tasks of multiple servers can be integrated into fewer VM server, thereby shutting down the idle machines to reduce energy consumption. Virtualization technology uses the four functions of the virtualization platform: A physical machine (PM) can support multiple running VMs, each VM can be dynamically started and shut down, the resources of the VM can be dynamically adjusted and allocated, the last is that the VM can support dynamic migration [27]. These virtualization functions provide a huge optimized space for cloud computing resource scheduling. Existing research in this direction focuses on two aspects: the initial placement of VMs and the dynamic migration of VMs. The

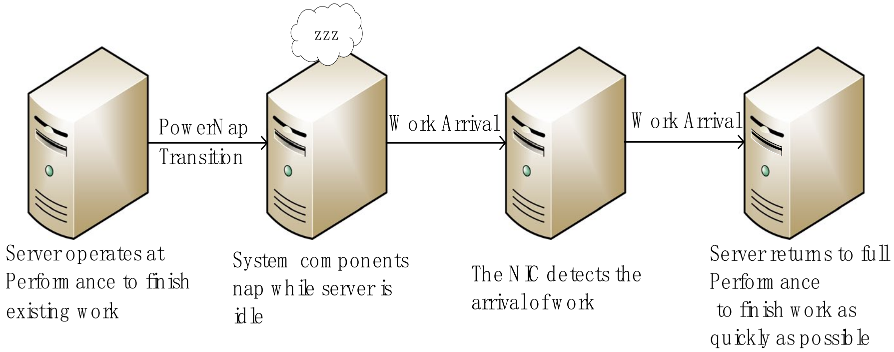

Fig. 4 The concept and principle of PowerNap

initial placement of VMs refers to the allocation of VMs to physical machines when the data center is just started and there is no load. The dynamic migration of VMs can achieve load balancing and VM integration.

In this paper, we divide the current solutions to the VM placement problem into three types: constraint specification, the bin-packing problem, and random integer programming. Constraint specification method is a programming method to solve the problem of complex VM optimization configuration combination. It solves the feasible solution of the problem by defining a set of constraints. Zhang et al. [28] designed a cloud resource allocation model according to constraint specifications, which can meet the requirements of service quality and reduce the costs of resource usage. Dong et al. [29] used the constraints of network link capacity and physical machine size to schedule VMs through a two-stage virtual machine scheduling algorithm. At first, the bin-packing best-fit heuristic algorithm and the minimum cut hierarchical clustering algorithm are combined to locate the VM, which minimizes the number of active physical machines, and then through the maximum link utilization optimization to further avoid network congestion. The bin-packing problem means modeling the VM placement problem as a boxing problem. It is described as: There are many boxes, each box has a capacity of V, there are N items, and the volume of each item is C. The purpose is to pack all items into the box to make the number of boxes is minimized. The VM placement problem is modeled as a boxing problem, means that each item is regarded as a VM, which is tightly packaged in a minimum number of boxes, and each box is regarded as a physical machine. Song et al. [30] presented a dynamic resource allocation algorithm based on binning to optimize the number of running servers. The authors designed an easy online packing algorithm named Variable Item Size Packing (VISBP). The core of VISBP is to change the load size of the VM at runtime. Although VISBP has advantages in load balancing and hot spot detection, it violates the Service Level Agreement (SLA) to a certain extent. Random integer programming is used to study decision-making problems with uncertainty. It uses the estimation model of probability distribution with relevant data to obtain the optimal expected solution of the problem according to the mathematical model. Here, the future needs of VMs or applications are unknown.

Therefore, some VM placement technologies use this method to predict the appropriate physical machine. Chen et al. [31] assigned VMs to PMs and proposed a new VM placement method. By associating the dynamic load of the VM with the fixed demand, the stochastic optimization problem is simplified. This method determines the resource demand of a VM through the principle of statistical multiplexing. The principle estimates the resource demand of the VM as the sum of the inherent demand and the related demand. On this basis, the authors designed a VM placement algorithm. Evaluations driven by real data center's load tracking show that this method saves 10% to 23% of energy compared to previous integration methods.

The dynamic migration of VMs can achieve load balancing and reduce the appearance of hot spots. A survey in [32] analyzed the existing load balancing scheduling algorithm and comprehensively summarized the virtual machine scheduling technology of cloud data centers. This dynamic migration can be implemented in combination with DVFS technology. For example, Ding et al. [33] designed an energy optimization algorithm called EEVS, which used DVFS technology to schedule VMs with a given deadline. However, frequent migration of VMs would cause degradation of system performance. For this reason, the migration of VMs should be minimized while ensuring energy consumption and satisfying QoS. In addition, too much physical machine load leads to poor service quality, so a prediction-based mechanism is needed to solve this problem. Methods such as ML and NN have played an important role here. Barthwal et al. [34] decided to use regression model to predict CPU utilization. Based on the predicted value, the virtual machine with the shortest migration time was selected for migration, and the physical machine with the least predicted CPU utilization was selected to host the migrated VM. Khoshkholghi et al. [35] developed an adaptive energy-saving technology to manage VMs. They developed an algorithm that uses iterative weighted linear regression to detect overloaded and underutilized physical machines. Shaw et al. [36] proposed a virtual machine-physical machine mapping mechanism using reinforcement learning (RL) method. Their method enables the optimal allocation of VMs and significantly improves energy efficiency. Liu et al. [37] designed a dynamic integration method (EQVC) to improve energy efficiency and service quality and migrate redundant VMs. They used an autoregressive integrated moving average model (ARIMA) to predict the utilization of overloaded hosts. Besides, modeling can also be used to optimize the energy consumption of the data center. The Google data center received sensor data and input the data into the deep NN to predict how different choices will affect future energy consumption so that the system can take effective measures to reduce energy consumption and reduce the *PUE* value in time [38].

In recent years, meta-heuristics and heuristic algorithms have been used to solve large-scale scheduling problems in cloud data centers [39–41]. In order to ensure the reasonable scheduling of VMs, Ragmani et al. [42] proposed a fuzzy ant colony optimization algorithm (FACO). The function of the fuzzy module is to evaluate historical information to calculate the pheromone value so as to select the appropriate server while ensuring the best calculation time. Similarly, Kruekaew et al. [43] proposed an artificial bee colony heuristic task scheduling algorithm (HABC). The HABC algorithm combines artificial bee colony algorithm with heuristic scheduling algorithm for scheduling. The authors compared the method with other heuristic algorithms (including ACO algorithm, PSO algorithm), experiments results prove that HABC provides the best performance in scheduling and load balancing. Vila et al. [44] discussed a BLEMO (Blacklist Evolution Multi-Objective) method based on the GA to schedule resources on Iaas cloud resources to minimize the makespan and energy consumption. A Chhabra et al. [45] presented a hybrid multi-objective policy (CSPSO) for scheduling a set of parallel tasks in IaaS cloud systems to meet QoS and energy-efficiency expectations.

2.2 Energy consumption optimization methods for storage system and network system

The storage system is an important part of the data center, and the optimization of the storage system greatly affects the energy-saving development of the data center. In this field, current research focuses on finding new high-speed and low-power storage technologies. For example, O'Connor et al. [46] proposed a new high-bandwidth, low-power DRAM architecture (FGDRAM) in 2017. Compared with previous DRAMs, this fine-grained DRAM has twice the energy efficiency. In addition, some works explored new memory energy-saving strategies. Some studies [47–49] allowed servers to remotely use the remaining memory of other servers in order to improve the utilization of memory resources while reducing energy consumption. A memory decomposition strategy proposed by Nitu in [50], this strategy decoupled the CPU and memory at the power level, which allows remote access to the memory while suspending the server, thereby reducing system energy consumption.

With the development of network technology during recent years, cloud computing networks have greatly changed from the previous traditional data center networks in terms of network architecture and network traffic models. Energy-saving research on network system has gradually become a hot topic in the industry. Some studies [51, 52] have investigated in the data center's network by turning off unnecessary network devices (routers, switches, etc.) or putting them into sleeping mode to save power consumption of network. Inspired by the hierarchical data communication network (DCN) topology and data center traffic patterns, Zhang et al. [53] studied a hierarchical model (HERO) to reduce energy consumption, specifically, by turning off idle switches or links to optimize the power of network components. The power optimization consists of two levels: core-level and pod-level. The authors extended this model in [54], they added switching power loss, and proved that the HERO problem was an NP problem so that the authors designed lots of heuristic algorithms to solve this NP problem.

Data center network traffic integration is an effective method to reduce network energy consumption. The traditional method is centralized traffic integration. However, this integration method expresses the energy-saving problem as a linear programming model with high computational complexity and low response speed. Different from the previous centralized traffic integration, Zhou et al. [55] considered the need to respond to changes in data center's network traffic and proposed a distributed stream-level traffic integration framework (DREAM) to achieve energy efficiency. In DREAM, distributed working nodes and hosts work together to merge traffic into a part of the data center's network and close the part with no traffic load. It can quickly respond to traffic emergencies. The results have proved that DREAM saves at least 15.8% of energy in data center networks on average, which is more energy efficient than centralized traffic integration, and application-level latency is also reduced by 30%.

2.3 Comparison of energy-saving methods for IT equipment

As the server handles a large number of computing tasks in the data center, it has become the most important energy-consuming component of the IT equipment in data centers. Therefore, most of the existing energy-saving research based on IT equipment is aimed at servers, but the energy consumption generated by storage system and network system cannot be ignored. For this reason, Table 1 provides a comprehensive comparison for the energy saving of IT equipment.

3 Energy saving for cooling systems

In data centers, IT equipment processes a large number of computing tasks, causing its temperature to rise, which affects the availability of the equipment. In order to reduce the heat emitted by computing equipment and ensure the long-term stable operation of the equipment, cooling systems such as CRAC are required. However, the cooling system running at full load or some unreasonable use will lead to higher energy consumption in the data center. The energy consumption generated by the refrigeration system takes about 40% of the total energy consumption of the data center. In some data centers, this proportion even exceeds 50% [56]. During these years, the world's data centers have been growing into "mega-data-centers." The new generation of data centers is more prominently manifested as: larger scale, higher density, and higher cooling requirements. However, a short-term cooling interruption in the air conditioning system will cause IT equipment to overheat and shut down. Therefore, how to ensure that the data center's refrigeration system can provide the required environmental temperature and humidity for the computer room for a long time has become a matter of widespread concern in the academic and industrial circles. In this paper, we summarize the research in recent years from three main aspects, namely location layout and airflow organization of computer rooms, thermal-aware scheduling technology and other energy-saving technologies.

3.1 Location layout and airflow organization of computer rooms

At present, the commonly used cooling scheme in the data center industry is air cooling. Air cooling involves the air supply method of the computer room. The air supply mode of the computer room is mainly divided into two types: upper air supply and lower air supply. Early data centers did not consider the way of airflow organization, and most of them adopted the upward air supply method. The upper

Table 1 Comparison for the energy saving of IT equipment

Literature	Specific strategies and methods	Optimization object	Implementation effect	Experimental data
17	Three energy-saving solutions based on DVFS: Lowest-DVFS, δ-Advanced-DVFS, Adaptive-DVFS	Servers	The experimental results show that these three solutions can obtain higher economic efficiency with lower energy consumption	CloudSim simulation
18, 19, 20	Task scheduling algorithm based on DVFS technology	Servers	The results show that the proposed methods can significantly improve data center energy efficiency	Real Google data center/CloudSim simulation
21	A runtime scheduler (CPU MISER) based on DVFS	Servers	Experimental results show that when using CPU MISER as the DVFS scheduler, up to 20% energy can be saved	Real clusters
23, 24	Using the scheduling strategy of ML algorithms to shut down servers with less load and try to consolidate tasks to the same server to reduce energy consumption	Servers	Simulation experiments show that these methods can significantly improve energy efficiency	Simulator
25	An energy-saving optimization framework	Servers	Experiments show that this method can significantly reduce server energy consumption compared with static resource allocation methods	Simulator
26	A method to eliminate server idle power consumption by quickly switching ultra-low power consumption states (PowerNap)	Servers	Experiments show that PowerNap can reduce server energy consumption by about 74% in total	Actual commercial deployment tracking utilization
28, 30, 31	The problem of VM initialization and placement is considered as a constraint specification problem, a binning problem, and a random integer programming problem	Servers	Experiments show that optimizing the initial placement of VMs can effectively reduce energy consumption	Real data center load tracking or simulation
46	A new high-bandwidth, low-energy DRAM architecture (FGDRAM)	Storage system	Twice the energy efficiency of previous DRAM	Simulate a GPU system to handle workload tasks

Table 1 (continued)

Literature	Specific strategies and methods	Optimization object	Implementation effect	Experimental data
53, 54	A hierarchical energy optimization model (HERO) to reduce energy consumption	Network system	The simulation results show that HERO can effectively reduce the power consumption and complexity of network elements	Simulator
55	A distributed stream-level traffic integration framework (DREAM) to merge traffic into a part of the data center network and close the part with no traffic load	Network system	Testing and evaluation proved that DREAM has saved about 15.8% of energy consumption for data center networks	Wikipedia and Facebook traffic tracking test

air supply method is to send the cold air processed by the air conditioning unit out through the top of the machine room. Based on the jet flow principle, the supply air flow forms a whirlpool in the entire machine room, fully mixes with the waste heat and returns to the air conditioning system through the air return port. The upper air supply method should choose an air conditioning unit with a higher residual pressure outside the machine, and at the same time, ensure that the return air is smooth and not blocked by the equipment. This method has obvious short-circuit phenomenon of cold and hot air flow, and the cooling efficiency is low. The downward air supply method is currently the main form of air conditioning and cooling air supply for data centers, and it is widely used in enterprise data centers and operator IDCs. This kind of air supply method usually lays an electrostatic floor in the data center's computer room. The height of the electrostatic floor is usually 20–100 cm, even as high as 2 m. The airflow organization structure is shown in Fig. 5 [9]. Firstly, the cold air of the special air conditioner in the computer room is sent to the underside of the electrostatic floor to form a large static pressure box. The static pressure box can reduce the dynamic pressure of the air supply system and increase the static pressure, steady airflow, etc. And then send the cold air to the server rack through the perforated floor. The return air can be returned through the floor space in the computer room or the dedicated return air duct (space above the ceiling) so as to form obvious hot and cold channels. The downward air supply method has high cooling efficiency and simple installation. However, this airflow layout still has the problem of hot air recirculation and cold air bypass. Uneven air distribution will cause problems such as early server failures and increased downtime. All of these will lead to a substantial increase in the operating cost of the data center. For this reason, it is necessary to improve the layout and airflow structure of the data center.

The ceiling height is a factor that affects the heat distribution of the airflow. The ceiling height should not be too high or too low. If the ceiling height is too high, more hot air will be recirculated, which will affect the inlet temperature of the rack. At lower ceiling heights, there is not enough space for the hot air to move, which leads to heat buildup. Nagarathinam et al. [57] compared the ceiling height of the

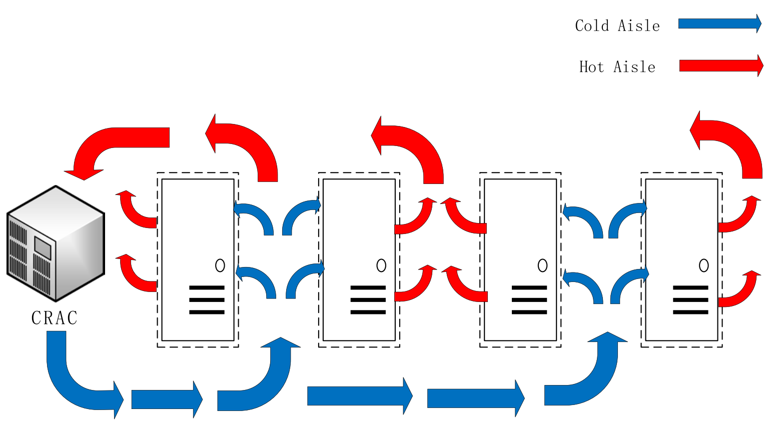

Fig. 5 Schematic diagram of downward air supply

data center with other parameters and found that the optimal ceiling height of the data center is 2.52 m. Mulay et al. [58] modified the rack and optimized the location of the server in the rack to isolate the cold air flow from the hot air flow to avoid the emergence of server hotspots. At the same time, the use of partitions to isolate the hot and cold aisle can also effectively reduce the generation of hot air recirculation and cold air bypass. In addition, in the data center, the flow rate of the perforated bricks may need to be adjusted to meet the airflow requirements of the racks. There are many ways to balance the airflow of porous bricks. Among them, variable open area porous bricks and under-floor partition walls are the most effective solutions [59, 60].

3.2 Thermal-aware scheduling technology

Different from energy-aware scheduling, thermal-aware scheduling focuses on thermal issues in data centers. It considers minimize the heat from the active servers, so the chance of hot spots and cooling load is also reduced. The thermal-aware scheduler can not only reduce the *PUE* of the data center, but also save energy and protect the environment. Zhuravlev et al. [61] investigated the thermal-aware scheduling problem and divided the scheduling algorithm into reactive, active and hybrid. Reactive scheduling refers to the scheduling method adopted after a task has a thermal anomaly, while active scheduling needs to predict the future thermal conditions to avoid thermal anomalies in the system. The efficiency of active forecasting depends on the forecasting method. Reactive scheduling is simple and takes a short time to execute scheduling tasks, but its cost and energy saving are not as good as active scheduling.

Wang et al. [62] designed an active method based on task thermal analysis. The method schedules the hottest tasks on the coldest server, in other words, puts the hottest tasks on the coldest server for execution, and follows the microprocessor-based task scheduling principle, that means the hot tasks must be scheduled before cold tasks. In addition, the authors also used the RC model to calculate the current temperature of the server, and used the temperature distribution of each job to predict the temperature of the node after the job is scheduled. Obviously, this kind of scheduling can prevent the server from running at the highest temperature, thereby reducing cooling energy consumption. However, the method proposed by the authors was based on homogeneous servers. They did not consider that the servers in most data centers in reality are heterogeneous. Since the temperature of the server is related to time and space to a certain extent. In space, the entrance temperature of the server is related to the layout of the computer room. In time, the workload of the server at a certain moment is related to historical load information and future load. For this reason, Sun et al. [63] focused on the topic of space–time thermal-aware at data centers. Firstly, they designed a spatio-temporal analysis model, and based on this, described the entire data center thermal-aware scheduling problem. Subsequently, under the condition of thermal constraints, an online scheduling heuristic algorithm is designed to solve the workload and thermal management problem. The authors used the proposed heuristic algorithm to schedule tasks to balance the server load. In particular, in thermal management, the algorithm combines DVFS technology which dynamically adjusts the operating frequency of the device to avoid the temperature exceeding the threshold. Compared with other thermal-aware scheduling methods, the proposed method has better performance and energy benefits due to the consideration of temporal and spatial correlation and DVFS technology.

MirhoseiniNejad et al. [64] presented a unified thermal perception method to control IT system and cooling systems. The authors proposed a thermal model for the overall data center. The model considered the thermal interaction between IT and the cooling unit, which saved power by studying the synergy between the workload scheduler and the cooling unit's operating parameters. The author describes the optimization problem as a constraint problem, as follows:

$$\underset{T_{inlet}^{water}, Q_{air}^{rmcu}, \overline{u}}{Minimize} \quad P^{fan} + P^{chiller} \tag{6}$$

$$subject \;\; to \; \sum\nolimits_{i=1}^{n} ui = d \tag{7}$$

$$0 \le ui \le u^{\max}, i = 1, \ldots, n \tag{8}$$

$$\max\left(T_{inlet}^{server}\right) \le T_{red}^{server} \tag{9}$$

among them, the decision variables are T_{inlet}^{water}, Q_{air}^{rmcu}, $\overline{u}$. The T_{inlet}^{water} is the temperature of cool water provided to the RMCU by the chiller. Q_{air}^{rmcu} is the air flow rate of the cooling unit. $\overline{u}$ is the utilization vector of length n. The author modeled the cooling energy consumption as the sum of the energy consumption of the fan and the energy consumption of the chiller. The goal is to find the optimal value of the variable to minimize energy consumption while keeping the server's inlet temperature below the red line temperature T_{red} (Eq. 9). Equation 8 guarantees system availability. To solve this optimization problem, a model is used to correlate the operating parameters and the assigned workload with the temperature distribution. Experimental results show that compared with other workload distribution algorithms, the joint optimization proposed by the authors can save a lot of cooling energy.

3.3 Other energy-saving technologies

3.3.1 Liquid cooling

Today, new data center optimization technologies are changing the way that the data center managers deploy cooling systems and the control platform. The liquid cooling method is an efficient cooling method developed to support high-density and large-scale components. Liquid cooling includes two parts: direct cooling technology and indirect cooling technology. Direct liquid cooling refers to the direct contact between electronic components and liquid coolant, allowing the coolant to take away the heat of the electronic components. Direct liquid cooling includes immersion and spray. Direct immersion liquid cooling is a new type of heat dissipation technology that has attracted the attention of the industry in recent years. At the Global Supercomputing Conference during these years, many server companies have demonstrated their products on immersion liquid cooling, which has greatly improved the attention to liquid cooling in the industry. Submerged liquid cooling has higher heat dissipation efficiency because the heating element is in direct contact with the coolant, and there is no maintenance of hardware such as sealed enclosures and fluid connectors. Indirect liquid cooling is different from direct liquid cooling, the cooling liquid and electronic devices are not in direct contact. Instead, a liquid cooling radiator or evaporator is used as an intermediate heat dissipation component to realize heat exchange. Indirect liquid cooling can use a coolant with high heat transfer, thereby reducing power consumption, and the cooling liquid and electronic devices contact indirectly makes the system safer. However, the addition of the sealed enclosure and the server-grade piping required to conduct the liquid makes its deployment and maintenance more difficult than direct liquid cooling. Indirect liquid cooling is divided into single-phase indirect liquid cooling and two-phase indirect liquid cooling according to the presence or absence of phase change.

Liquid cooling technology has received extensive attention from the industry. 3 M company [65] developed an implementation scheme of two-phase immersion cooling in an open tank of a server. Yan et al. [66] proposed a direct spray cooling method that used inclined gas-assisted nozzles to induce multiple spray cooling covering the entire server board area. However, the main disadvantage of this method is the high-maintenance heat exchange operation. Indirect liquid cooling has been used in IBM's server cooling scheme [67]. The heat transfer process of indirect liquid cooling occurs in a special cold plate. The focus of existing research is on the design and experiment of various cold plates such as porous media and microchannel heat sinks to study chip-level thermal management solutions [68, 69]. Liquid cooling technology has high cooling efficiency and is suitable for high-computing cluster systems. Compared with the conventional water cooling system, the liquid cooling system eliminates the need for chilled water systems such as chillers and refrigeration pumps, thus saving a lot of room space. However, the current liquid cooling technology lacks corresponding standards. The granularity of liquid cooling is at the cabinet (box) level, server level or even chip level. The low granularity, complex systems and relatively difficult maintenance make it difficult to promote at present. Therefore, liquid cooling technology is a huge opportunity for the existing data center cooling architecture, but also a huge challenge.

3.3.2 Waste heat recovery

The IT equipment in the data center generates a lot of heat during operation, and most of the electrical energy entering the server is emitted in the form of heat, however, cooling the heat consumes more resources. Therefore, people have begun to think about the problem of waste heat recovery in data centers in recent years. Recycling the waste heat generated by the data center can not only reduce carbon emissions but also save data center operating costs. The waste heat recovery of the data center is different from the conventional industrial waste heat recovery system. Firstly, although the heat of the data center is sufficient, the quality of the waste heat is generally poor. Secondly, the temperature is limited by the temperature of the data center electronic devices, which is generally limited to 85 °C. These limitations make it difficult to reuse it through conventional thermodynamic cycles. For this reason, Ebrahimi et al. [70] compared the waste heat source and waste heat flow of different types of data centers. They discussed the applicability of eight waste heat recovery technologies to different data center designs and determined the most promising waste heat recovery solutions. Through comparative analysis, it is concluded that absorption refrigeration and Organic Rankine Cycle (ORC) are the two most effective solutions for waste heat recovery in data centers. Absorption refrigeration provides additional cooling water for the cooling load, thereby reducing the load of CRAC, while ORC directly uses waste heat flow to provide on-site power generation.

Some large cloud computing companies recycle the waste heat generated by the data center. Facebook used the waste heat of the data center for office heating to save a lot of air-conditioning costs. In [71], the opportunity and challenge brought by the waste heat of the data center as the main heating source of office buildings and apartment buildings were studied. The authors proved that using the waste heat of the data center to directly heat the home can save a lot of energy. Marcinichen et al. [72] studied the role of data center waste heat recovery in the water preheating process of coal-fired power plants. They modeled the data center as a cooling cycle and the power utility as a thermal cycle. In this model, the waste heat of the data center is between the condenser and the circulation system before the feed water heating, which makes the waste heat generated by the data center to be well utilized. Through feasibility analysis of the system, the authors believed that using the waste heat of the data center in the water waste heat process of coal-fired power plants is environmentally friendly and energy-saving.

3.3.3 Natural cold source utilization

The utilization of natural cold source is usually divided into two aspects. On the one side, it refers to the use of outdoor cold air to take away the heat of the data center in areas with low geographical environment temperature, which greatly shortens the operating time of air conditioners. On the other side, it refers to the use of outdoor natural water resources to cool the data center in areas close to rivers and seas. Using these natural environmental advantages can greatly reduce the energy consumption of data centers, thereby saving electricity costs. For this reason, during these years, researchers have been exploring how to effectively use natural cold source to optimize energy consumption of data centers. The utilization of natural cold air can be divided into two categories: direct utilization and indirect utilization. Direct utilization refers to a technology that directly introduces outdoor cold air into the data center for cooling when the outdoor temperature meets certain conditions. Since the data center has strict requirements on humidity and cleanliness, it is necessary to humidify and filter the outdoor air before it enters the data center. Indirect utilization of natural cold air, using heat exchange equipment to achieve cold and heat exchange between outdoor air and indoor air. It can reduce the indoor circulating air temperature while ensuring indoor humidity and cleanliness. The difficulty of this method is how to effectively realize the heat exchange of the air.

Direct air cooling is a method of introducing outdoor air directly into the data center to cool down the equipment. The typical deployment of a direct air-free cooling system is shown in Fig. 6. When the outdoor air temperature is low, the outdoor cold air is directly introduced into the data center through the filter to cool the equipment. In some large data centers, a natural cooling energy-saving mechanism is used, and it has been proved that using this mechanism can save a lot of energy costs. Intel has used this system in a 10mw data center for about 10 months, and they concluded that the use of wind-side free cooling schemes can save $2.87 million in energy every year [73]. Microsoft and Google also took advantage of free natural air cooling opportunities and built fresh air natural cooling data centers in Europe [74, 75]. Yin et al. [76] conducted a feasibility analysis on whether five cities in China can use natural cold source. Experiments have shown that this cooling method is not suitable due to the large amount of dustfall in Harbin and Beijing. However, Shanghai, Kunming, and Guangzhou are suitable for natural cooling due to climate and geographical location. It can clearly be seen that direct natural air cooling technology depends on local environmental conditions.

To ensure continuous and efficient operation of high-density computing systems and air-conditioning systems in a data center, the temperature and humidity of the data center environment must be strictly controlled. The direct introduction of cold outdoor air into the data center will affect its internal environment, thereby affecting the availability of the data center to a certain extent. Therefore, the indirect use of outdoor cold air technology is introduced. The natural cooling of the indirect air side requires an air-to-air heat exchanger to achieve. The typical deployment is designed in Fig. 7. The system uses a heat pipe or a heat wheel to realize the heat exchange of the air. The heat pipe or the heat wheel transfers sensible heat through the evaporation of the working fluid near the hot return air and the condensation near the outside cold air. Researchers are currently exploring

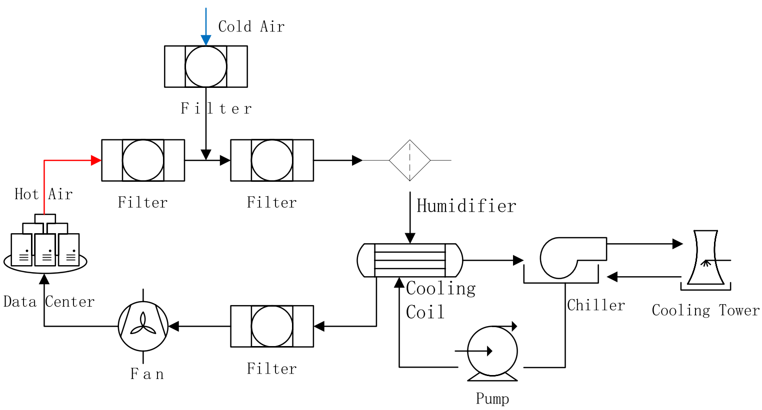

Fig. 6 Direct air-side economizer

the impact of using different heat exchangers on data center energy consumption. Bao et al. [77] used plate heat exchangers in the heat exchange of communication base stations. The experimental results show that this air heat exchanger reduced the total power consumption of the corresponding electronic stations by 29%.

In addition to using natural cold air, the utilization of natural cold source can also use natural water resources, which is related to the location of data centers. An effective geographic location can save a lot of energy consumption for data center's operations and maintenance in the future. Some large Internet operators attach great importance to the location of data centers. For example, water-cooled data centers are usually built in areas with abundant water resources, such as large rivers and seaside. Google invested 200 million euros to build a data center in Northern Europe, where the natural climate brings huge opportunities for data center cooling. Microsoft built the data center on a seabed in California, using sea water to achieve the goal of naturally reducing energy consumption.

Based on data center energy modeling, Jinkyun et al. [78] evaluated the impact on data center energy usage by describing the climate differences in several climate zones and the differences in the corresponding data center cooling systems. For several different cooling types, representative climatic conditions of the four regions are applied to the data center energy model, and suggestions for energy optimization are given for each climatic zone. Through the study of 17 data center cases, the trend and application technology of green data center energy efficiency are obtained. Simulation experiments show that the energy demand of the data center exhibits different patterns in different climate regions. In hot and humid areas, the outdoor temperature is high throughout the year, and it does not have the advantages of air cooling and water cooling. In hot and dry areas, the use of direct air cooling systems with evaporative cooling can reduce energy consumption by 38%. In addition, the outdoor temperature in cold regions is low

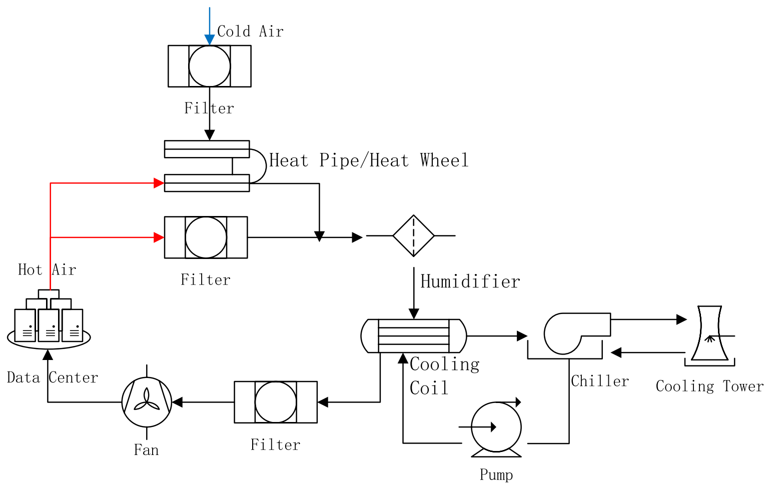

Fig. 7 Indirect air-side economizer

throughout the year, so the indirect air-cooled energy-saving system is effective and can reduce energy consumption by up to 80%.

3.4 Comparison of energy-saving methods for cooling systems

This section mainly focuses on a comprehensive comparison of energy-saving methods for cooling systems. These strategies can be roughly divided into three aspects: (1) location layout and airflow organization of computer rooms; (2) thermal-aware scheduling technology; (3) other optimization methods include liquid cooling, waste heat recovery technology and natural cold source utilization. Table 2 shows a comprehensive comparison of these methods.

4 Joint optimization of computing system and cooling system

Previous studies have focused on unilateral energy optimization of data center computing system or cooling system. However, the management of the computing system will affect the requirements of cooling system. From one side, server consolidation without understanding the cooling system may result in too many server hot spots in the data center, therefore increase the demand for cooling capacity and reduce cooling efficiency. From the other side, minimizing cooling energy consumption requires calculating the system balance workload. However, this causes more active servers running and develops more computing power. Therefore, it is necessary to find a reasonable scheduling solution for the joint computing system and cooling system. By developing and evaluating fine-grained models of CRAC, room temperature, humidity, and servers, and integrating these models into the overall scheduling algorithm is an effective solution to minimize total energy consumption. However, the complex relationship between the computing system and the cooling system makes joint optimization challenging. For the joint optimization method, the key is to strike a balance between computing energy and cooling energy to achieve better overall energy efficiency.

In terms of joint optimization calculation of energy consumption and cooling energy consumption, most of the works is done by establishing a total power consumption model and developing an algorithm to minimize the total power consumption. Xiang et al. [79] established a thermal model by comprehensively analyzing the airflow and the temperature distribution of the server CPU. The authors modeled the total energy consumption of the data center as follows: $E_{total} = E_{computing}(S) + E_{cooling}(S)$. Based on this, they proposed an integrated virtual machine scheduling algorithm GRANITE to minimize the total energy consumption of the data center. The decision of scheduling algorithm is based on offline recognition model, including workload model, server model and cooling model. This algorithm considers the initial placement and dynamic migration of VMs. In order to improve cooling efficiency, the algorithm dynamically adjusts the capacity of CRAC. The article established a constraint model under the constraints of server capacity, SLA, and CPU critical temperature, so as to find the optimal scheduling

Table 2 Comparison for the energy saving of cooling systems

Literature	Specific strategies and methods	Optimization technology	Implementation effect
59	Modifying the location of racks and servers to isolate the hot and cold aisles	Airflow organization	Experiments show that this method can increase the heat dissipation effect by 50%
63, 64	Scheduling tasks or servers based on thermal-aware, tasks/servers with higher temperatures are prioritized	Thermal-aware scheduling	Experiments show that it can effectively reduce the total energy consumption and TCO
65	A unified thermal perception method to control the IT system and cooling system and a thermal model for the overall data center	Thermal-aware scheduling	Experimental results show that compared with other workload distribution algorithms, the joint optimization can save a lot of cooling energy consumption
67	A multiple spray cooling method covering the entire server board area	Liquid cooling	This method has good cooling effect, but its heat exchange operation is complicated
69, 70	Studying the design of various cold plates such as porous media and microchannel radiators	Liquid cooling	High cooling efficiency and difficult maintenance
72	Using the waste heat generated by the data center for residential heating	Waste heat recovery	Experimental results show that this solution can save a lot of energy waste
73	Using the waste heat generated by the data center in the water preheating process of coal-fired power plants	Waste heat recovery	High economic efficiency, saving a lot of energy waste
74, 75, 76	Using outdoor natural cold air to save energy	Natural cold source utilization	It can save a lot of money for enterprises

scheme and air supply temperature, so that the total energy consumption is minimized. Feng et al. [80] proposed a global-energy-aware virtual machine placement (VMP) strategy to reduce the total energy consumption of data centers from multiple aspects. The authors modeled the total energy consumption as a combination of server power consumption, cooling energy consumption, and network energy consumption. A two-step SAG algorithm is designed to lower the energy consumption of cloud data centers where multiple VMs are deployed. In order to solve the complexity of thermal behavior and the uncertainty of workload, the scheduling plan of the data center needs to dynamically integrate thermal factors and energy consumption factors. To this end, Shashikant Ilager et al. [81] designed an energy and thermal-aware scheduling (ETAS) algorithm that dynamically consolidates VMs to minimize the overall energy consumption while proactively preventing hotspots. The authors established an accurate power model and a thermal model, they used ETAS to reduce the total energy consumption of the data center without any hotspot creation. In a similar way, thermal-aware HPC job scheduling has been investigated by Sun et al. [63], where the primary focus of the work is to reduce the makespan. These methods reduce the total energy consumption to a certain extent and are simple to implement, but they either do not have a complete power consumption model or mostly confined to HPC workloads.

In this regard, intelligent algorithms can play a role. Arroba et al. [82] proposed a new power perception model and a meta-heuristic optimization strategy, which relies on simulated annealing (SA) algorithm to realize the joint optimization of computing energy consumption and cooling energy consumption. With regard to the optimization of the cooling system, a cooling strategy based on the temperature of the equipment in the system is proposed in order to find the maximum cooling setting value of CRAC unit. To avoid thermal problems, the authors defined a maxCooling Set Point for each host to ensure that the CPU temperature is within a safe range. In the end, the cooling set point is defined as the minimum value of all servers within the maximum cooling set point. The study found that the total energy efficiency was increased to 21.74% by using a combination of meta-heuristic algorithm and best-fit declivity algorithm.

Other works managed all the resources of the cloud data center by establishing an overall energy consumption framework in the data center. In order to provide an overall cloud platform to manage all resources in the cloud data center, Gill Et al. [83] discussed an energy consumption aware resource scheduling technology (CRUZE) based on Cuckoo algorithm, which is used for the overall management of cloud computing resources, including servers, network, storage and cooling systems. The authors presented a system model for all components of the cloud data center (Fig. 8). The system model is divided into three layers, namely (1) SaaS: This layer handles incoming workloads and forwards them to the workload manager. (2) PaaS: This layer consists of a controller to control different resources of the entire system. The controller has five submodules: workload manager, VM/ resource manager, fault manager, cooling manager, and energy manager. (3) IaaS: The IaaS layer includes information related to the cloud infrastructure. The authors modeled the comprehensive energy consumption of cloud data center as: $E = E_{processor} + E_{storage} + E_{memory} + E_{network} + E_{cooling} + E_{extra}$

. This model covers the various components of the cloud data center. The article established various models and a fitness function based on energy consumption and reliability and then used the Cuckoo algorithm to solve the model to perform overall scheduling of heterogeneous workloads. The experimental results show that the proposed technology can reduce energy consumption by 20.1% and increase reliability and CPU utilization by 17.1% and 15.7%, respectively. Similarly, a power optimization framework (PowerNetS) for coordinating servers, data center network and cooling systems was proposed in [84]. PowerNetS uses workload correlation analysis to save more energy during server and traffic consolidation. More importantly, PowerNetS attempts to change the DCN topology during the server consolidation process in order to obtain more intra-server traffic and shorter traffic through fewer switches. Wan et al. [85] also studied a cross-layer optimization framework to minimize the holistic energy consumption in data center. The advantage of the framework is that it dynamically coordinates the CPS and CLS via decision variables across various layers to obtain a global optimal control strategy.

The above researches show that workload scheduling with a fine combination of server state and data center thermal characteristics is a novel and effective strategy to reduce total energy consumption. However, due to the complex relationship between workload and thermal characteristics, scheduling is challenging. This complexity is that the workload of the data center changes all the time, and the thermal management of the data center is also complicated, involving airflow

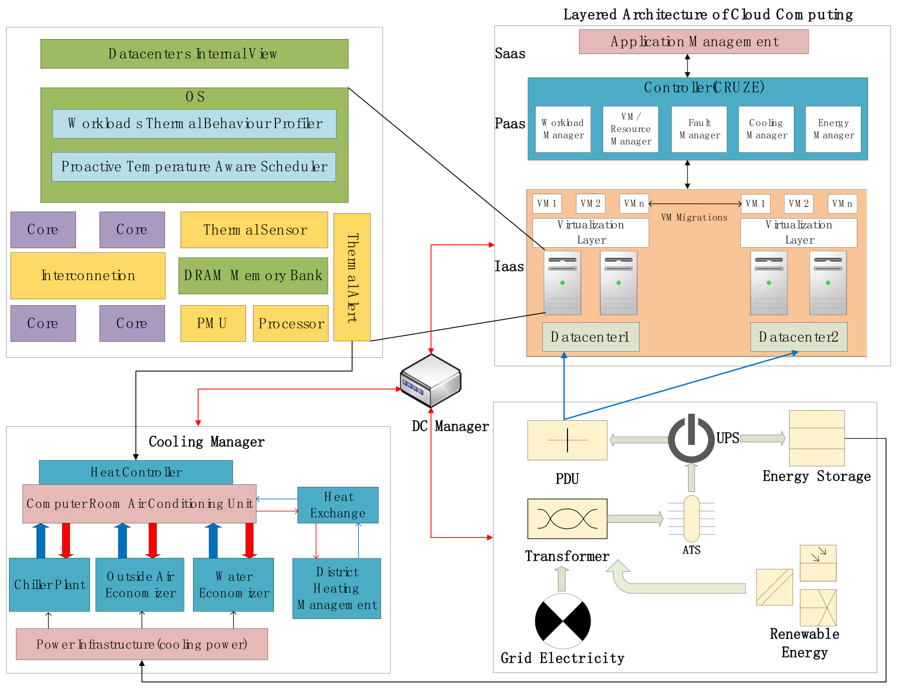

Fig. 8 Cloud resource scheduling system model based on Cuckoo optimization

organization, server location, and so on. For these reasons, researchers need to formulate dynamic and globally optimized scheduling strategies in the future.

5 Analysis and prospect of future research

Researchers have conducted a lot of research on the energy-saving technology of cloud data center, given many energy-saving algorithms and technologies, and achieved good results. However, most of the existing research attempts to predict and analyze the optimized energy consumption of a single component, and few studies the energy-saving optimization of multiple components in the entire data center. Especially due to the rise of big data and cloud computing, cloud data centers have been transformed into a new generation of data centers. Their large scale, high power consumption density, and highly dense internal components make the construction of a new generation of green data centers face more new problems and challenges. Through the analysis and summary of existing research, we point out that the problems and directions of cloud data center energy-saving technology that need further research and improvement are: integral optimization of energy consumption jointing multiple components, energy saving using artificial intelligence methods, energy saving based on novel hardware equipment, hybrid cooling energy saving and comprehensive energy conservation with various energy technologies.

5.1 Integral optimization of energy consumption jointing multiple components

Internet service prividers (ISPs) rely heavily on data centers to support the growing computing demand for their services, which is leading to a growing problem with data center energy consumption. Current efforts to address this issue have focused on reducing the number of active servers through server consolidation, thereby increasing resource utilization in the data center. However, this approach usually does not take into account other core components of data center (cooling systems, network, and storage), which account for a significant portion of the total energy consumption of cloud data centers. As for the energy saving of cooling systems, the existing research focuses on the optimization of hardware equipment and thermal management, few researches combine the computing system to carry out the overall optimization. Therefore, the optimization scheduling of energy consumption jointing refrigeration equipment and IT equipment in the data center will become a development trend. For example, in [79], by establishing a thermal model and comprehensively analyzing factors such as airflow and temperature, a scheduling algorithm was designed to minimize data center energy consumption under the constraints of these thermodynamic conditions. Zheng et al. [84] proposed a power optimization framework for coordinating server, data center's network and cooling systems to minimize the power consumption of data centers.

5.2 Energy saving using artificial intelligence methods

In order to improve the utilization of system resources, most of the existing server energy-saving methods use virtualization technology to dynamically allocate resources to realize load balancing [37, 42, 43]. Reasonable scheduling can shorten the time, reduce server hotspots and thus reduce energy consumption. However, too many VMs migration and change will affect system performance. Therefore, the migrations of virtual machine should be minimized while ensuring system resource utilization. By using ML to accurately predict the load of future applications, unnecessary migration can be reduced and energy consumption can be reduced. In [34], the regression model was used to predict CPU utilization. In this method, virtual machine with the shortest migration time was selected for migration according to the predicted value, and the virtual machine managed migration with the physical machine which has the minimum CPU utilization was selected. At the same time, a predictive monitoring system can be set to monitor the changes of environmental parameters (such as temperature, humidity, etc.) and operating parameters (such as CPU utilization, load, etc.) of the data center in real time which can predict the changes of parameters at the next moment. For example, a power monitoring system established by Facebook's data center collects equipment power information every 3 s, which determines whether to limit the power supply of the equipment according to the power threshold, stigma, and cabinet to achieve the purpose of controlling the power of the data center. It can accommodate more equipment to reduce energy consumption [86].

In addition, modeling based on ML and NN has become a more practical solution to optimize data center energy consumption. Through the training of a large amount of data, the optimized control method of server and refrigeration equipment was obtained. For example, back propagation (BP) neural network or artificial neural network (ANN) was used to model and optimize the *PUE* value of data centers, and LSTM network model was used to model and optimize the efficiency value of refrigeration equipment. Google has used NN to model its data center and optimize energy consumption [87]. Google used sensors to collect data center's data of infrastructure operation and power usage information, they used a multilayer NN to study the data and built models to predict the *PUE* value of the data center. The results show that ML method can effectively predict data center energy demand, so energy efficiency can be further improved by adjusting parameter settings. Except the above methods, some of the most representative computational intelligence algorithms can be used to solve data center energy-saving problems, such as monarch butterfly optimization (MBO), earthworm optimization algorithm (EWA), elephant herding optimization (EHO), moth search (MS) algorithm, etc. They are suitable for solving optimization problems, and the energy consumption problem of the data center can be considered as a combined optimization problem. These algorithms can help find the global optimal solution. Therefore, they can be used to solve the problem in the future.

5.3 Energy saving based on novel hardware equipment

For the purpose of improving the power efficiency of the server, it is necessary to improve the technology on the hardware equipment. Compared with the old servers, the new low-power servers will make great progress in the optimization of energy and efficiency. So far, some technologies have made some progress, such as porting highly parallel applications to general-purpose GPUs, servers using ARM architecture with lower power consumption [88]. Baidu designed and deployed an energy-saving rack server called "Beiji" in 2011. "Beiji" rack server system is a server system specially developed for energy saving. It adopts centralized refrigeration, centralized power supply design, and has a rack management control module (RMC) which is responsible for cooling, power monitoring and other management functions [89].

Similarly, on the refrigeration side, the use of new refrigeration equipment can significantly improve the efficiency of refrigeration. At present, common air-conditioning systems in data centers are air-cooled direct expansion air-conditioning system and cold water room air-conditioning system, in which the cold water room air-conditioning system also includes air-cooled cold water type and water-cooled cold water type. In recent years, in order to explore more efficient refrigeration equipment, a number of new refrigeration system models with flexible cooling methods and high refrigeration efficiency have emerged. For example, double cooling source system, heat pipe composite system. Double cooling source system is divided into air cooling/freezing water system and water cooling/freezing water system. Therefore, the trend of energy saving in the future is to develop more efficient energy-saving hardware on top of the current infrastructure in order to achieve higher energy efficiency, which is a huge challenge for hardware equipment manufacturers and relevant researchers.

5.4 Hybrid cooling energy saving

Faced with the huge energy demand brought by the construction and operation of data centers, the energy-saving strategy of data centers has developed from the initial ventilation cooling system to liquid cooling and natural cooling source utilization. Liquid cooling technology has the advantages of high cooling efficiency, low energy consumption and environmental protection. Liquid cooling technology was first applied to the cooling of super or large computer chips and is expected to be widely applied to the cooling of servers in higher density data centers in the future. In 2018, at the I/O Developer Conference, Google announced that it would be using liquid cooling technology in its data centers for the first time, and said that in the future it would be switching to liquid cooling in its data centers. The utilization of natural cold source refers to the use of outdoor cold air or natural cold source water to cool the data center. This technology has a good application prospect for areas with generally low outdoor temperature or close to natural water cooling source. In 2018, Microsoft built its data center on the seabed, using natural seawater to cool the data center and reduce the use of air conditioning to save energy. Hybrid cooling is the application of different cooling technologies (such as liquid cooling, natural cooling, air cooling, water cooling, etc.) in the same data center to achieve comprehensive energy saving. The European Nuclear Research Council has established a hybrid cooling data center, nearly 9% of the servers in the data center used a liquid cooling solution that includes free cooling [90]. By using hybrid cooling, local liquid cooling can be provided in any hot spot, and other cooling schemes can be selectively used to cool other non-hot spots. Therefore, hybrid cooling can effectively save cooling power, thus greatly reducing cooling costs.

5.5 Comprehensive energy conservation with various energy technologies

In the face of huge energy consumption, the main power source of cloud data center still comes from fossil fuels. However, the combustion of fossil fuels produces a large amount of carbon emissions, which is not conducive to environmental protection. The use of clean and renewable energy in data centers has become an industry trend. Greenpeace has promoted and encouraged the use of new energy in data centers by rating the clean energy of data centers of global IT companies. At the same time, governments around the world have also developed energy-saving regulations and policies. New energy refers to various forms of energy, such as wind and solar energy, that are different from traditional energy. IT companies are applying new energy in their data centers. For example, Google's wind farm in Kenya and other several wind farms in Texas. Facebook also built its solar data center in Oregon. Although these renewable new energies can improve the environmental problems of data centers, their unstable and time-varying characteristics make it impossible for data centers to fully use new energies at present. Some researches solve this problem by modeling new energies and using prediction techniques and algorithms to predict the amount of new energies available in the future. But it is difficult to meet the large-scale energy consumption demand of data centers through new energy source. Traditional power grid is still the most important power supply method. Therefore, how to optimize the new energy structure and cooperate with new energy and traditional energy to achieve more efficient and environmentally friendly power supply is the future research direction of data center energy conservation.

6 Conclusion

In this paper, we summarize and classify the existing energy-saving methods of cloud data center, focuses on the existing energy-saving technologies based on IT equipment and cooling systems, and describes the advantages and limitations of each method. The joint optimization scheme formed on this basis can effectively reduce the energy consumption of the cloud data center, which is also the current research focus. In the end, the development direction and trend in the future are proposed. From a global perspective, different energy-saving methods have different contributions to the energy saving of data centers. In the initial construction of the data center, the consumption of air conditioning in the room can be reduced by about 30% through scientific and reasonable site selection and rational use of natural energy. In the optimization of the server, the power adjustment technology can effectively reduce the energy consumption by about 20–40%. By optimizing the resource scheduling algorithm and management strategy, the total energy consumption can be saved by about 10–25%. In addition, in the process of data center construction, reasonable optimization of network system, storage system and other systems can also effectively reduce the energy consumption of the data center.

Acknowledgements The authors are grateful to the anonymous reviewers for their valuable comments and suggestions. This work is supported by National Natural Science Foundation of China (Grant Nos. 62072187, 61872084), Major Program and of Guangdong Basic and Applied Research (2019B030302002), Guangzhou Science and Technology Program key projects (Grant Nos. 202007040002, 201907010001).

References

1. White paper (2020) IDC:2025 China will have the world's largest data circle. http://www.d1net.com/uploadfile/2019/0214/20190214023650515.pdf. Accessed 28 June 2020.
2. Belkhir L, Elmeligi A (2018) Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J Clean Prod 177:448–463. https://doi.org/10.1016/j.jclepro.2017.12.239
3. Data Center Cooling Working Group of Chinese Refrigeration Society (2018) China Data Center Annual Research Report on Cooling Technology Development. China Construction Industry
4. Ren C, Wang D, Urgaonkar B, Sivasubramaniam A (2012). Carbon-aware energy capacity planning for datacenters. In: 2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, pp 391–400. doi: https://doi.org/10.1109/MASCOTS.2012.51
5. Johnson P, Marker T (2009) Data centre energy efficiency product profile. Equipment energy efficiency committee (E3) of the Australian Government Department of the Environment, Water, Heritage and the Arts (DEWHA), Tech. Rep, 212.
6. Orgerie AC, Assuncao MD, Lefevre L (2014) A survey on techniques for improving the energy efficiency of large-scale distributed systems. ACM Comput Surv (CSUR) 46(4):1–31. https://doi.org/10.1145/2532637
7. Beloglazov A, Buyya R, Lee YC, Zomaya A (2010) A taxonomy and survey of energy-efficient data centers and cloud computing systems. Adv Comput 82:47–111. https://doi.org/10.1016/B978-0-12-385512-1.00003-7
8. Salimian L, Safi F (2013) Survey of energy efficient data centers in cloud computing In: Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing. pp 369–374. doi: https://doi.org/10.1109/UCC.2013.76
9. Kheirabadi AC, Groulx D (2016) Cooling of server electronics: A design review of existing technology. Appl Therm Eng. https://doi.org/10.1016/j.applthermaleng.2016.03.056
10. Ni J, Bai X (2017) A review of air conditioning energy performance in data centers. Renew Sustain Energy Rev 67:625–640. https://doi.org/10.1016/j.rser.2016.09.050
11. Wan J, Gui X, Kasahara S, Zhang Y, Zhang R (2018) Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey. IEEE Access 6:48867–48901. https://doi.org/10.1109/ACCESS.2018.2866840
12. Li X, Jiang XH, Wu CH, Ke-Jiang YE (2015) Research on Thermal Management Methods for Green Data Centers. J Computer, (10):72–92. doi: https://doi.org/10.11897/SP.J.1016.2015.01976
13. Nadjahi C, Louahlia H, Lemasson S (2018) A review of thermal management and innovative cooling strategies for data center. Sustain Comput: Inform Syst 19:14–28. https://doi.org/10.1016/j.suscom.2018.05.002
14. Mastelic T, Oleksiak A, Claussen H, Brandic I, Pierson JM, Vasilakos AV (2015) Cloud computing: survey on energy efficiency. ACM Comput Surv 47(2):36. https://doi.org/10.1145/2656204

15. Malla S, Christensen K (2019) A Survey on Power Management Techniques for Oversubscription of Multi-Tenant Data Centers. ACM Comput Surv 52(1):1–31. https://doi.org/10.1145/3291049
16. Zhang W, Wen Y, Wong YW, Toh KC, Chen CH (2016) Towards Joint Optimization Over ICT and Cooling Systems in Data Centre: A Survey. IEEE Commun Surv Tutorials 18(3):1596–1616. https://doi.org/10.1109/COMST.2016.2545109
17. Kim KH, Beloglazov A, Buyya R (2011) Power-aware provisioning of virtual machines for real-time Cloud services. Concurr Comput: Practice Exp 23(13):1491–1505. https://doi.org/10.1002/cpe.1712
18. Wu CM, Chang RS, Chan HY (2014) A green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. Futur Gener Comput Syst 37:141–147. https://doi.org/10.1016/j.future.2013.06.009
19. Gu L, Zeng D, Barnawi A, Guo S, Stojmenovic I (2014) Optimal task placement with QoS constraints in geo-distributed data centers using DVFS. IEEE Trans Comput 64(7):2049–2059. https://doi.org/10.1109/TC.2014.2349510
20. Tang Z, Qi L, Cheng Z, Li K, Khan SU, Li K (2016) An energy-efficient task scheduling algorithm in DVFS-enabled cloud environment. J Grid Comput 14(1):55–74. https://doi.org/10.1007/s10723-015-9334-y
21. Ge R, Feng X, Feng W, Cameron K W (2007) Cpu miser: A performance-directed, run-time system for power-aware clusters In: 2007 International Conference on Parallel Processing (ICPP 2007). IEEE, pp 18–18. doi: https://doi.org/10.1109/ICPP.2007.29
22. Shuja J, Gani A, Shamshirband S, Ahmad RW, Bilal K (2016) Sustainable Cloud Data Centers: A survey of enabling techniques and technologies. Renew Sustain Energy Rev. https://doi.org/10.1016/j.rser.2016.04.034
23. Berral J L, Goiri Í, Nou R, Julià F, Guitart J, Gavaldà R, Torres J (2010) Towards energy-aware scheduling in data centers using machine learning. In: Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking. pp 215–224. doi: https://doi.org/10.1145/1791314.1791349
24. Duy TVT, Sato Y, Inoguchi Y (2010) Performance evaluation of a green scheduling algorithm for energy savings in cloud computing. In: 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). IEEE, 1-8. doi: https://doi.org/10.1109/IPDPSW.2010.5470908
25. Beloglazov A, Abawajy J, Buyya R (2012) Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Futur Gener Comput Syst 28(5):755–768. https://doi.org/10.1016/j.future.2011.04.017
26. Meisner D, Gold BT, Wenisch TF (2009) PowerNap: eliminating server idle power. ACM SIGARCH Comput Architec News 37(1):205–216. https://doi.org/10.1145/2528521.1508269
27. Rong H, Zhang H, Xiao S, Li C, Hu C (2016) Optimizing energy consumption for data centers. Renew Sustain Energy Rev 58:674–691. https://doi.org/10.1016/j.rser.2015.12.283
28. Zhang L, Zhuang Y, Zhu W (2013) Constraint programming based virtual cloud resources allocation model. Int J Hybrid Inform Technol 6(6):333–344. https://doi.org/10.14257/ijhit.2013.6.6.30
29. Jiankang D, Hongbo W, Shiduan C (2015) Energy-performance tradeoffs in IaaS cloud with virtual machine scheduling. China Commun 12(002):155–166. https://doi.org/10.1109/CC.2015.7084410
30. Song W, Xiao Z, Chen Q, Luo H (2014) Adaptive resource provisioning for the cloud using online bin packing. IEEE Trans Comput 63(11):2647–2660. https://doi.org/10.1109/TC.2013.148
31. Chen M, Zhang H, Su Y Y, Wang X, Yoshihira K (2011) Effective VM sizing in virtualized data centers. In: 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops. IEEE, pp 594-601. doi:https://doi.org/10.1109/INM.2011.5990564
32. Xu M, Tian W, Buyya R (2017) A survey on load balancing algorithms for virtual machines placement in cloud computing. Concurr Comput: Practice Exper 29(12):e4123. https://doi.org/10.1002/cpe.4123
33. Ding Y, Qin X, Liang L, Wang T (2015) Energy efficient scheduling of virtual machines in cloud with deadline constraint. Futur Gener Comput Syst 50:62–74. https://doi.org/10.1016/j.future.2015.02.001
34. Barthwal V, Rauthan M, Verma R (2019) Virtual Machines Placement Using Predicted Utilization of Physical Machine in Cloud Datacenter. In: International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India.

35. Khoshkholghi MA, Derahman MN, Abdullah A, Subramaniam S, Othman M (2017) Energy-efficient algorithms for dynamic virtual machine consolidation in cloud data centers. IEEE Access 5:10709–10722. https://doi.org/10.1109/ACCESS.2017.2711043
36. Shaw R, Howley E, Barrett E (2017) An advanced reinforcement learning approach for energy-aware virtual machine consolidation in cloud data centers. In: 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, pp. 61–66. doi: https://doi.org/10.23919/ICITST.2017.8356347
37. Liu Y, Sun X, Wei W, Jing W (2018) Enhancing energy-efficient and QoS dynamic virtual machine consolidation method in cloud environment. IEEE Access 6:31224–31235. https://doi.org/10.1109/ACCESS.2018.2835670
38. Richard E, Jim G (2020) DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40. Accessed 10 July 2020.
39. Yu W, Li X, Yang H, Huang B (2017) A multi-objective metaheuristics study on solving constrained relay node deployment problem in WSNS. Intell Autom Soft Comput. https://doi.org/10.1080/10798587.2017.1294873
40. Mansouri N, Zade BMH, Javidi MM (2019) Hybrid task scheduling strategy for cloud computing by modified particle swarm optimization and fuzzy theory. Comput Ind Eng 130:597–633. https://doi.org/10.1016/j.cie.2019.03.006
41. Amini Motlagh A, Movaghar A, Rahmani AM (2020) Task scheduling mechanisms in cloud computing: A systematic review. Int J Commun Syst 33(6):e4302. https://doi.org/10.1002/dac.4302
42. Ragmani A, Elomri A, Abghour N, Moussaid K, Rida M (2019) FACO: A hybrid fuzzy ant colony optimization algorithm for virtual machine scheduling in high-performance cloud computing. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01631-5
43. Kruekaew B, Kimpan W (2020) Enhancing of Artificial Bee Colony Algorithm for Virtual Machine Scheduling and Load Balancing Problem in Cloud Computing. Int J Comput Intell Syst 13(1):496–510. https://doi.org/10.2991/ijcis.d.200410.002
44. Vila S, Guirado F, Lerida JL, Cores F (2019) Energy-saving scheduling on IaaS HPC cloud environments based on a multi-objective genetic algorithm. J Supercomput 75(3):1483–1495. https://doi.org/10.1007/s11227-018-2668-z
45. Chhabra A, Singh G, Kahlon KS (2020) QoS-Aware energy-efficient task scheduling on HPC cloud infrastructures using swarm-intelligence meta-heuristics. CMC-Comput Mater Continua 64(2):813–834
46. O'Connor M, Chatterjee N, Lee D, Wilson J, Agrawal A, Keckler SW, Dally WJ (2017) Fine-grained DRAM: energy-efficient DRAM for extreme bandwidth systems. In: 2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 41-54. doi:https://doi.org/10.1145/3123939.3124545
47. Li F, Das S, Syamala M, Narasayya V R (2016) Accelerating relational databases by leveraging remote memory and RDMA. In: Proceedings of the 2016 International Conference on Management of Data. pp 355–370. doi: https://doi.org/10.1145/2882903.2882949
48. Novakovic S, Daglis A, Bugnion E, Falsafi B, Grot B (2016) The case for RackOut: Scalable data serving using rack-scale systems. In: Proceedings of the Seventh ACM Symposium on Cloud Computing. pp 182–195. doi: https://doi.org/10.1145/2987550.2987577
49. Barthels C, Loesing S, Alonso G, Kossmann D (2015) Rack-scale in-memory join processing using RDMA. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp 1463–1475. doi: https://doi.org/10.1145/2723372.2750547
50. Nitu V, Teabe B, Tchana A, Isci C, Hagimont D (2018) Welcome to zombieland: practical and energy-efficient memory disaggregation in a datacentre. In: The Thirteenth EuroSys Conference. doi: https://doi.org/10.1145/3190508.3190537
51. Mann V, Kumar A, Dutta P, Kalyanaraman S (2011) VMFlow: Leveraging VM mobility to reduce network power costs in data centers. In: International Conference on Research in Networking. Springer, Berlin, Heidelberg, pp 198–211.
52. Masdari M, Nabavi SS, Ahmadi V (2016) An overview of virtual machine placement schemes in cloud computing. J Netw Comput Appl 66:106–127. https://doi.org/10.1016/j.jnca.2016.01.011
53. Zhang, Yan, Ansari, Nirwan (2012) HERO: Hierarchical energy optimization for data center networks. In: IEEE International Conference on Communications. iEEE.
54. Zhang Y, Ansari N (2015) HERO: hierarchical energy optimization for data center networks. IEEE Syst J 9(2):406–415. https://doi.org/10.1109/JSYST.2013.2285606

55. Zhou L, Bhuyan LN, Ramakrishnan KK (2019) DREAM: Distributed energy-aware traffic management for data center networks. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems. pp 273–284. doi: https://doi.org/10.1145/3307772.3328291
56. Zhang H, Shao S, Xu H, Zou H, Tian C (2014) Free cooling of data centers: A review. Renew Sustain Energy Rev 35:171–182. https://doi.org/10.1016/j.rser.2014.04.017
57. Nagarathinam S, Fakhim B, Behnia M, Armfield S (2013) A comparison of parametric and multivariable optimization techniques in a raised-floor data center. J Electron Pack, 135(3). doi: https://doi.org/10.1115/1.4023214
58. Mulay V, Agonafer D, Irwin G, Patell D (2009) Effective thermal management of data centers using efficient cabinet designs. In: International Electronic Packaging Technical Conference and Exhibition. 43604: 993-999. doi: https://doi.org/10.1115/InterPACK2009-89351
59. Patankar SV, Karki KC (2004) Distribution of cooling airflow in a raised-floor data center. ASHRAE Trans 110:629–634
60. Karki KC, Patankar SV (2006) Airflow distribution through perforated tiles in raised-floor data centers. Build Environ 41(6):734–744. https://doi.org/10.1016/j.buildenv.2005.03.005
61. Zhuravlev S, Saez JC, Blagodurov S, Fedorova A, Prieto M (2012) Survey of energy-cognizant scheduling techniques. IEEE Trans Parallel Distrib Syst 24(7):1447–1464. https://doi.org/10.1109/TPDS.2012.20
62. Wang L, Khan SU, Dayal J (2012) Thermal aware workload placement with task-temperature profiles in a data center. Journal of Supercomputing 61(3):780–803. https://doi.org/10.1007/s11227-011-0635-z
63. Sun H, Stolf P, Pierson JM (2017) Spatio-temporal thermal-aware scheduling for homogeneous high-performance computing datacenters. Future Gener Comput Syst 71(jun):157–170. https://doi.org/10.1016/j.future.2017.02.005
64. MirhoseiniNejad SM, Moazamigoodarzi H, Badawy G, Down DG (2020) Joint data center cooling and workload management: A thermal-aware approach. Futur Gener Comput Syst 104:174–186. https://doi.org/10.1016/j.future.2019.10.040
65. Tuma P E (2010) The merits of open bath immersion cooling of datacom equipment. In: 2010 26th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM). IEEE, pp 123–131.https://doi.org/10.1109/STHERM.2010.5444305
66. Yan Z B, Duan F, Wong T N, et al. (2010) Large area spray cooling by inclined nozzles for electronic board. In: Electronics Packaging Technology Conference. IEEE. doi: https://doi.org/10.1109/EPTC.2010.5702609
67. Zimmermann S, Meijer I, Tiwari MK, Paredes S, Michel B, Poulikakos D (2012) Aquasar: A hot water cooled data center with direct energy reuse. Energy 43(1):237–245. https://doi.org/10.1016/j.energy.2012.04.037
68. Lee YJ, Singh PK, Lee PS (2015) Fluid flow and heat transfer investigations on enhanced microchannel heat sink using oblique fins with parametric study. Int J Heat Mass Transf 81:325–336. https://doi.org/10.1016/j.ijheatmasstransfer.2014.10.018
69. Dede EM, Liu Y (2013) Experimental and numerical investigation of a multi-pass branching microchannel heat sink. Appl Therm Eng 55(1–2):51–60. https://doi.org/10.1016/j.applthermaleng.2013.02.038
70. Ebrahimi K, Jones GF, Fleischer AS (2014) A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. Renew Sustain Energy Rev 31:622–638. https://doi.org/10.1016/j.rser.2013.12.007
71. Liu J, Goraczko M, James S, Belady C, Lu J, Whitehouse K (2011) The data furnace: Heating up with cloud computing. HotCloud.
72. Marcinichen JB, Olivier JA, Thome JR (2012) On-chip two-phase cooling of datacenters: Cooling system and energy recovery evaluation. Appl Therm Eng 41:36–51. https://doi.org/10.1016/j.applthermaleng.2011.12.008
73. Atwood D, Miner JG (2008) Reducing data center cost with an air economizer. Intel Corporation, White Paper
74. Miller R. (2009) Microsoft's chiller-less data center. Data center knowledge. https://www.datacenterknowledge.com/archives/2009/09/24/microsofts-chiller-less-data-center
75. Miller R. (2009) Google's chiller-less data center. Data center knowledge. https://www.datacenterknowledge.com/archives/2009/07/15/googles-chiller-less-data-center/
76. Yin H, Zhu Y, Wang YL, Gao Y (2011) Effects of rapamycin on cell growth and apoptosis of pancreatic carcinoma SW1990 cells. Tumor 31(1):49–52

77. Bao L, Wang J, Kang L (2012) The applied effect analysis of heat exchanger installed in a typical communication base station in Beijing of China. Energy Proc 14:620–625. https://doi.org/10.1016/j.egypro.2011.12.985
78. Cho J, Kim Y (2016) Improving energy efficiency of dedicated cooling system and its contribution towards meeting an energy-optimized data center. Appl Energy 165:967–982. https://doi.org/10.1016/j.apenergy.2015.12.099
79. Li X, Garraghan P, Jiang X, Wu Z, Xu J (2017) Holistic virtual machine scheduling in cloud datacenters towards minimizing total energy. IEEE Trans Parallel Distrib Syst 29(6):1317–1331. https://doi.org/10.1109/TPDS.2017.2688445
80. Feng H, Deng Y, Li J (2021) A global-energy-aware virtual machine placement strategy for cloud data centers. J Syst Architect 116:102048. https://doi.org/10.1016/j.sysarc.2021.102048
81. Ilager S, Ramamohanarao K, Buyya R (2019) ETAS: Energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation. Concurr Comput: Practice Exp 31(17):e5221. https://doi.org/10.1002/cpe.5221
82. Arroba P, Risco-Martín JL, Moya JM, Ayala JL (2018) Heuristics and metaheuristics for dynamic management of computing and cooling energy in cloud data centers. Softw: Practice Exp 48(10):1775–1804. https://doi.org/10.1002/spe.2603
83. Gill SS, Garraghan P, Stankovski V et al (2019) Holistic resource management for sustainable and reliable cloud computing: An innovative solution to global challenge. J Syst Softw 155:104–129. https://doi.org/10.1016/j.jss.2019.05.025
84. Zheng K, Zheng W, Li L, Wang X (2017) PowerNetS: Coordinating Data Center Network With Servers and Cooling for Power Optimization. IEEE Trans Netw Serv Manag 14(3):1–1. https://doi.org/10.1109/TNSM.2017.2711567
85. Wan J, Gui X, Zhang R, Fu L (2017) Joint cooling and server control in data centers: A cross-layer framework for holistic energy minimization. IEEE Syst J 12(3):2461–2472. https://doi.org/10.1109/JSYST.2017.2700863
86. Wu Qiang, Deng Qingyuan, Ganesh L, et al. (2016) Dynamo: Facebook's data center-wide power management system. In: Proceedings of the ACM/IEEE 43rd Annual International Symposium on Computer Architecture. Seoul, South Korea: IEEE, pp 469–480. doi: https://doi.org/10.1145/3007787.3001187
87. Gao J (2020) Machine learning applications for data center optimization. Google, http://research.google.com/pubs/pub42542.html. Accessed 26 July 2020.
88. Mastelic T, Brandic I (2015) Recent trends in energy-efficient cloud computing. IEEE Cloud Comput 2(1):40–47. https://doi.org/10.1109/MCC.2015.15
89. Pang W, Wang C, Ahuja N, et al. (2017) An advanced energy efficient rack server design, In: 2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm). IEEE. doi: https://doi.org/10.1109/ITHERM.2017.7992569
90. CERN Accelerating science, Data Centre. (2020) http://information-technology.web.cern.ch/about/computer-centre. [2020]. Accessed 20 July 2020.

Authors and Affiliations

Huiwen Cheng[1] · Bo Liu[1] · Weiwei Lin[2] · Zehua Ma[1] · Keqin Li[3] · Ching-Hsien Hsu[4,5]

✉ Weiwei Lin
linww@scut.edu.cn

Huiwen Cheng
1846199589@qq.com

Bo Liu
liugubin530@126.com

Zehua Ma
mzh.scnu@qq.com

Keqin Li
lik@newpaltz.edu

Ching-Hsien Hsu
robertchh@gmail.com

[1] School of Computer Science and Technology, South China Normal University, Guangdong, China

[2] School of Computer Science and Engineering, South China University of Technology, Guangdong, China

[3] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

[4] Department of Computer Science and Information Engineering, Asia University, Taichung City, Taiwan

[5] Department of Medical Research, China Medical University Hospital, China Medical University, Taichung City, Taiwan