Journal of Systems Architecture 119 (2021) 102253

Contents lists available at ScienceDirect

Journal of Systems Architecture

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

A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization

Qingxia Zhang [a], Zihao Meng [b], Xianwen Hong [c], Yuhao Zhan [c], Jia Liu [d], Jiabao Dong [e], Tian Bai [b], Junyu Niu [a], M. Jamal Deen [f,*]

[a] School of Computer Science and Technology, Fudan University, Shanghai, 201203, China
[b] DataScience Group, Gridsum, Beijing, 100083, China
[c] Hefei Data Center, Postal Savings Bank of China, Hefei, Anhui, 230000, China
[d] School of Civil and Resource Engineering, University of Science and Technology Beijing, Beijing 100083, China
[e] School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[f] Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada

ARTICLE INFO

Keywords:
CPSS
Cooling system
Data center
Power consumption management
Optimization strategy

ABSTRACT

Data center is a fundamental infrastructure of computers and networking equipment to collect, store, process, and distribute huge amounts of data for a variety of applications such as Cyber–Physical–Social Systems, business enterprises and social networking. As the demands of remote data services keep increasing, both the workload of the data center and its power consumption are rapidly rising. An indispensable part of a data center is the cooling system which provides a suitable operation environment, and accounts for around 30% of the power consumption of the data center. Therefore, optimized energy management of data center's cooling system is a highly profitable research area. Generally, a cooling system is made up of a mechanical refrigeration sub-system and a terminal cooling sub-system. Heat generated during operation of the data center will be absorbed by the latter one, and transferred into the outdoor environment via the former one. Depending on the cooling principle, current cooling solutions can be classified into air-cooling, liquid-cooling or free cooling technology. Although air-cooling is widely used in most existing data centers, the other two solutions have attracted more interests due to their excellent cooling effectiveness and higher energy efficiencies. Among the different cooling equipment, the chillers and fans are the major power consumers of the entire cooling system. Therefore, modeling of their power consumption is important for energy management of the cooling system, which can be classified into mechanism-based methods and data-driven methods. Based on the aforementioned models, optimization strategies for the operation management of cooling equipment are proposed to reduce the power consumption of the cooling system, which mainly includes the model predictive control-based methods and reinforcement learning-based methods. This paper is an overview of the data center's cooling system, which mainly includes the mainstream cooling solutions, the power consumption modeling methods and the optimization control strategies. In addition, several current challenges and future work in the data center's cooling system are described.

1. Introduction

Nowadays, data center (DC) has become a fundamental infrastructure of people's daily lives, providing remote data services for business, entertainment and many other human's requirements. With centralized computation and storage resources (CCSRs), DCs can resolve the issue of large-scale data processing and storage simultaneously for massive users [1–3]. Generally, the physical components of a DC include information and communications technology (ICT) devices and auxiliary equipment. The ICT devices are in charge of data center's pivotal functions such as data transmission and processing, including the servers, the network switches and the routers. The auxiliary equipment, such as the cooling system and the power supply system, are used to ensure the stable operation of the ICT devices. For the ICT devices, the operation temperature is an important factor that can greatly affect the stability of their performance. In the era of big data, the booming demands for cloud servers greatly facilitate the expanding scales of DCs [1,3], which also keeps the DC's workloads increasing. However, more ICT devices

* Corresponding author.
E-mail address: jamal@mcmaster.ca (M. Jamal Deen).

https://doi.org/10.1016/j.sysarc.2021.102253
Received 13 April 2021; Received in revised form 6 July 2021; Accepted 24 July 2021
Available online 31 July 2021

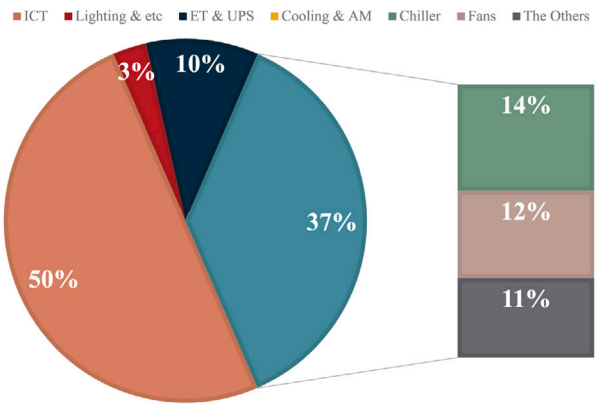


Fig. 1. The power consumption breakdown of the entire data center and the cooling system. ICT (information and communications technology) denotes ICT equipment. The ET & UPS denote the electricity transformer and uninterruptible power supply. The AM denotes the fans for air movement. The chiller and the fans are major power consumers while the other equipment includes the cooling tower, pumps and so on.

under high computational loads will generate a great amount of heat, which can seriously affect their stable operation. To provide a suitable environment for them, the cooling system has become an indispensable part of a DC, which also brings high power consumption. Therefore, the heat removal by the cooling system is one of the most prominent challenges in the maintenance of DCs [2,4,5].

In general, a typical DC cooling system is made up of a mechanical refrigeration sub-system (MRSS) and a terminal cooling sub-system (TCSS) [6]. The MRSS, usually comprised of equipment such as chillers, pumps and cooling towers, provides cold supply for the heat removed in the TCSS. The TCSS transfers heat from the indoor environment to the outdoor environment via the MRSS, the main solutions of which are air-cooling, liquid-cooling and free-cooling techniques. During DC's operation, the cooling system accounts for around 30% of the total power consumption. According to Fig. 1, the main power consumers of the cooling system are the fans of the computer room air conditioner (CRAC) and the chiller. To reduce their power consumption, adaptive control strategies using different algorithms have been proposed to adjust their set-points by real-time conditions such as workloads and operation temperatures. As two main paradigms to obtain the optimal control strategy, model-predictive control (MPC) [7–9] based on system dynamics modeling of cooling equipment, or reinforcement learning control (RLC) [10–13] analyzing the system feedbacks of decisions, can be used. Given the escalating workloads of DC, research on the operation management of the cooling system should not only focus on the cooling technique itself, but also consider the control strategies with the characteristics of the scenarios and the users in the DC data services.

In recent years, the powerful CCSRs provided by DCs have facilitated the development of Cyber–Physical–Social Systems (CPSSs) [14]. A CPSS, representing a holistic architecture with data collection devices, data transmission networks and system actuators, is applied to establish interactions among cyberspace, the physical environment and human society such as in smart manufacturing [15], smart city [16], wearable sensors [17–19], smart homes [20–22], and social dispersed computing [23]. In CPSS applications, multi-sources heterogeneous data is acquired from human, equipment objects and sensors in the social and physical layers, which becomes continuous dataflow at the edge of Internet-of-Things (IoT) networks [24]. The DC's central cloud continuously receives a huge amount of data, processing and analyzing it to provide information cognition and feedback execution to the edge data sources. Recently, under the impact of COVID-19, the lockdown significantly increased the business demands of cloud-based CPSS deployment for its stable remote access [25], which also bring challenges to DC's auxiliary equipment especially the cooling system.

As IoT technology is introduced in a DC's maintenance, the cooling system is equipped with a cyber layer for status monitoring and remote control, thus becoming a cyber–physical system. Since the Human-in-the-Loop (HitL) also exists in a DC cooling system, it will be beneficial for its power consumption efficiency enhancement to take the social factors into consideration. The benefits can be summarized by the following three points, which is illustrated in Fig. 2. **(1) *Considering Human's activity impacts on a DC cooling system.*** In a specific application scenario, the workloads of ICT devices in different time periods are unbalanced, which makes the required cooling capacity in each period different. For example, the COVID-19 outbreak has made online meetings common in business scenarios. Higher workloads of a DC by the transmission and processing of heterogeneous data such as text and video information, which is increased significantly during people's working hours. The cooling system should also be deployed with higher cooling power in those time periods, to keep the ICT devices stable operations under heavy workloads and reduce the power consumption cost in other free periods. **(2) *Effective Human–Computer interactions.*** It makes the DC monitoring system more efficient if it can be embedded with reliable human–computer interactions. Technicians are able to remotely acquire the real-time operating status such operation power of different devices such as cooling equipment. They can also carry out real-time control to different equipment, and receive successful messages when their commands take effect. **(3) *Intelligent Human Strategies Modeling.*** With different ICT workloads and specific environmental conditions, the corresponding adjustments made by technicians to cooling equipment can be valuable expert data. Artificial intelligence methods can be used to learn the control experience from the analysis of such data, which can build the intelligent control models to adaptively regulate the system under different conditions to reduce the power consumption.

To improve the power consumption efficiency of DC cooling systems, hardware optimization, such as the layout design and cooling equipment upgrade of TCSS and MRSS, is a basic aspect. Given the impacts by CPSS characteristics, it is also important to give the cooling system the capabilities of adaptive workload regulation and intelligent control scheduling. Therefore, this survey mainly discusses the classical solutions and recent development of the software and hardware components in DC cooling systems. The contributions can be mainly summarized into the following points. **(1) *Summarizing Thermal Efficiency Enhancement Technologies of DC Cooling hardware.*** The configurations of typical cooling systems in DCs, including air-cooling, liquid-cooling or free-cooling, will be introduced in Sections 2–4. For each of them, their unique hardware designs and the characteristics of their MRSS and TCSS are discussed in detail. By analyzing their advantages and limitations, the suitable scenarios of these three techniques are also summarized. **(2) *Analyzing State-of-the-art Software Development on DC Cooling Power Modeling and Optimization Control.*** The power consumption modeling of the two major power consumers, the chiller and the CRAC fans, are presented in Section 5.1. The mechanism-based modeling methods of the CRAC fans and the latest data-driven methods for the chillers are mainly introduced. Also, their merits and drawbacks are discussed. The MPC and RLC-based methods, as two mainstream adaptive control strategies, are introduced in Section 5.2 from the latest publications. The principles of the two algorithms and how they are used for DC's cooling control are presented. Moreover, the similarities and differences between MPC and RLC are discussed. **(3) *Summarizing Promising Directions for Current and Future Development of DC Cooling.*** The challenges which currently affect the DC cooling performances are summarized with their possible solutions, in Section 6.1. How emerging technologies can be used to improve DC cooling operation in future are also included in Section 6.2.

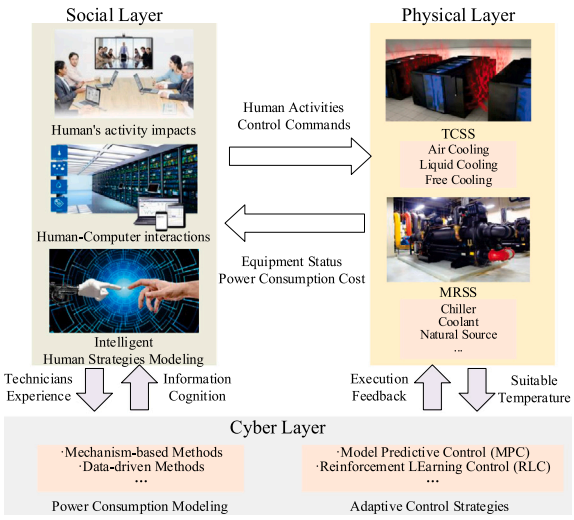


Fig. 2. How the social factors of CPSS benefit the DC cooling system.

Table 1
Summary of air cooling.

Level	Subtype [Refs]	Advantages	Main challenges
Room level	Raised-floor design [2,26,27]	Easy for implementation and maintenance.	Low heat dissipation efficiency and air intermixing.
Row level	Inter-row cooling [28–33]	Flexible distribution of cold air.	Spatial confinement.
	Overhead cooling [34–36]	Higher space utilization efficiency.	High possibility of cold air bypass.
Rack level	Internal enclosure [37,38]	Relatively highest power density.	High installation cost.

2. Air-cooling technology

Air-cooling technology is the most conventional solution that is widely applied in large scale data centers. Its main advantages are simple maintenance and acceptable operation cost [2,39]. The basic mechanism of the air cooling technology is illustrated in Fig. 3. It is accomplished by the airflow cycle in TCSS, whose cold supply is provided by the computer room air conditioner (CRAC). In the next two subsections, we will discuss terminal cooling and mechanical refrigerating air cooling technologies.

2.1. Terminal cooling of air cooling

Depending on the equipment densities in DCs, implementations of the air-cooling technology can be further classified into several branches, each of which is designed to improve power consumption efficiency from a certain level of cooling terminals [4]. Among these branches, the most sophisticated implementations is chip-oriented cooling. However, optimizations implemented inside the rack, such as the server-level schemes or chip-level schemes, are not very practical for some data centers since they greatly increase the expenses for implementation and maintenance [2]. Therefore, the approaches introduced below are room-level, row-level and rack-level air cooling technology, whose advantages and challenges are summarized in Table 1.

2.1.1. Room-level cooling

Room-level cooling is aiming for the total heat dissipation of the room, which usually uses the layout of raised floor for cold air supply [2,26]. Generally, the computer rooms of air-cooled DCs are set up with plenum under the raised floor and ceiling vent on the roof, which is illustrated in Fig. 4(a). The cold airflow is continuously provided by the CRAC, which will be delivered into the indoor environment through the perforated tiles on the raised floor plenum. In the computer rooms, different aisles separate the racks housing rows of servers. Cold airflow is ejected to each row through its cold aisle on one side, driven by the pressure difference between the plenum and the computer room. Through the hot aisle on the other side, warmed airflow will circulate to the ceiling vent after absorbing heat generated by the servers, which closes the airflow cycle.

In the aforementioned airflow cycle, the main efficiency failure is from the intermixing between the warmed return air and the cold inlet air, which will cause harmful effects such as the hot air recirculation (HAR) and the cold air bypass (CAB) [4,40]. In HAR, a portion of exhausted air fails in circulating to the hot aisle and intermingles with the supplied air in the cold aisle usually due to cold inlet air paucity. As the hot air recirculates over the ICT devices, the internal temperature of the racks will rise. Worse still, hot spots will occur inside the racks, and they can severely reduce the server performance. Conversely, CAB refers to under-utilization of cold air caused by excessive supply or leakage. The cold air fails in entering the racks along the cold aisle, but flows to the upper space of the room, which results in maldistribution of the cooling air. Extra cooling capacity has to be used to solve these problems, which brings huge power waste.

Different airflow management techniques have been proposed to alleviate the HAR and CAB effects. Among them, the air containment system (ACS), which blocks the airflow intermixing by retrofitting physical barriers, is one of the most effective schemes [27]. For example, the hot aisle containment system (HACS) is a physical component developed to significantly reduce the HAR effect. A typical layout of HACS

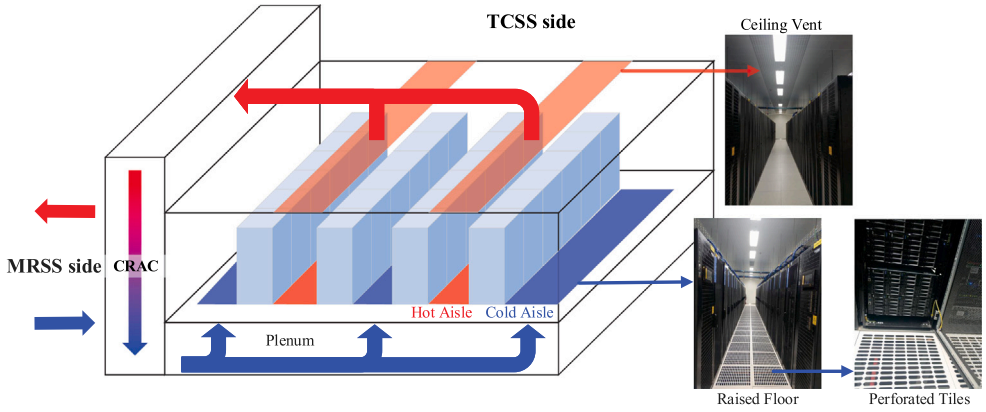


Fig. 3. The basic mechanism of the room-level air-cooling technology under the raised-floor design.

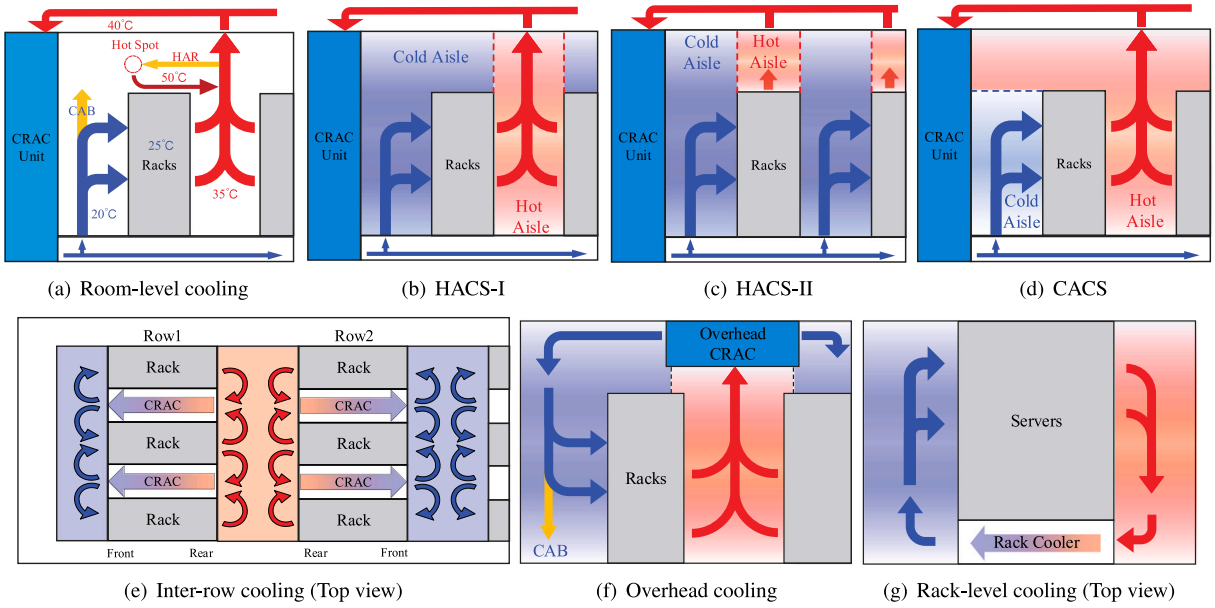


Fig. 4. Different branches of air-cooling technology. Among them, (a) depicts the room-level cooling without any containment which will cause CAB or HAR with hot spot. To solve these problems, the hot air containment system, shown in (b) and (c), and the cold air containment system, shown in (d), have been designed. The row-level cooling includes inter-row cooling and overhead cooling, shown in (e) and (f) respectively. The rack-level cooling is to mount rack cooler within the rack, which is shown in (g). It is noted that (e) is a top view of the computer room while (g) is the top view inside a rack with an enclosure.

is to seal the hot aisle by installing enclosures between the rear of racks and the ceiling on the roof, which is illustrated in Fig. 4(b). However, the hot aisle in this layout must be wide enough to avoid excessive pressure rise [4,41]. For data centers with insufficient space for the enclosure solution the chimney barrier, shown in Fig. 4(c), achieves the isolation by compact vertical ducts on top of the racks [4,42]. The high pressure in the narrow space of the vertical ducts demands that the duct should be strong enough. The cold aisle containment system (CACS), shown in Fig. 4(d) is designed to lessen the CAB effect [43–45]. In contrast to the typical layout of the HACS, the CACS prevents intermixing by enclosing the cold aisle with partitions between the perforated tiles and the airflow inlet on the racks. In this way, all other indoor spaces except CACs will become the hot aisles, which is an unacceptable environment for technicians to work in for a long time [27].

The main advantage of room-level cooling is that its cooling distribution can be flexibly adjusted by reconfiguration of the perforated tiles, which makes the room-level cooling more suitable for data centers with low equipment density [2]. Also, during the implementation of room-level cooling, the CRAC is usually located outside the computer room, which makes it easier to maintain the cooling equipment with negligible impact on the operation of ICT devices [38]. However, some important cooling components on this level, such as the plenum, are statically yoked to the building structure, making their modifications expensive [2]. Moreover, extra effort on preventing air intermixing is still a limitation of the room-level cooling, since the schemes such as ACS partially solves the problem [27]. Also, there is considerable uncertainty on the airflow distribution within the room, even when some containment is implemented.

2.1.2. Row-level cooling

Compared to room-level cooling, the main advantage of row-level cooling is to place the CRAC units near the ICT devices to shorten the airflow paths. There are two alternatives for the CRAC configurations on this level—inter-row cooling and overhead cooling. Their mechanisms are shown in Fig. 4(e) and (f).

In the configuration of inter-row cooling, the CRAC units are installed between adjacent racks so that the airflow is a rear-to-front distribution design. In this way, the warmed air exhausted from the rear of the racks can be uniformly absorbed into the CRAC units through the fans near the rear. The cold air, discharged from the fans on the front, will circulate into the racks again with a much shorter path than that of the room-level cooling. With this layout, the cooling capacity can be more flexibly controlled so that the CRAC units near the servers with higher IT load can have a higher power. However, the footprint of each CRAC unit for row-level cooling is similar with that of a rack, which will bring challenges on their spatial management. Due to space limitations, overhead cooling can be a better alternative solution in small-sized data centers [34–36]. The special CRAC units for overhead cooling are mounted in the upper middle of the two rows, which transforms the hot aisle into an upflow design. Nevertheless, the CAB effect is a main problem in the overhead cooling because cold air is more likely to collect at the bottom in rooms with vertical temperature differences.

In general, the row-level cooling can be regarded as an enhancement of room-level cooling, which gets higher power utilization efficiency [38]. The shorter airflow path makes predictions of the airflow distribution more accurate, which further helps to adjust the placement of racks for achieving better cooling performance [28]. Meanwhile, since cooling equipment is attached to each row of racks, it will be convenient to install additional cooling equipment during the expansion of the data center. However, compared with the room-level cooling, row-level cooling for a computer room with larger footprint results in higher installation cost [4,28]. Also, as cooling equipment is installed inside the computer room, the maintenance of the row-level cooling requires technicians to work near the ICT devices which may cause operation interruptions.

2.1.3. Rack-level cooling

In rack-level cooling, the rack coolers are mounted inside the racks, which further shortens the path of the airflow cycle [2,4,46]. As shown in Fig. 4(g), an extra enclosure housed with rack cooler is usually attached with the rack. Using a small partition, the internal space of the rack is divided into hot and cold aisles [37,38]. Within such a tight space, the exhausted warmed air and the cold supply air can follow the corresponding aisles to finish the cycle. During this process, the power usage of the fans driving the cold air circulation is significantly reduced.

Since different type of ICT devices have their own cooling demands, the flexibility of rack-level cooling on different ICT devices is a key to

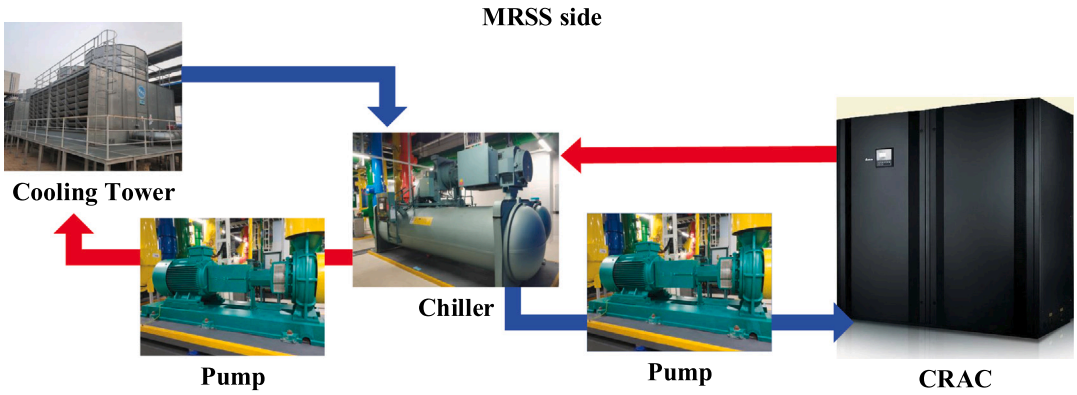


Fig. 5. The structure of a typical mechanical refrigeration subsystem (MRSS), which usually includes cooling towers, pumps, chillers and computer room air conditioner (CRAC).

improve the power usage efficiency [2,4,38]. For instance, the one-unit (1U) server requires higher airflow supply compared with the two-units (2U) server. Also, blade servers have higher demands of airflow than the communication enclosures. With rack-level cooling, the distributed cooling capacity can be adjusted according to the actual demand of the specific racks, which is beneficial in power usage saving of the fans.

Among the three levels discussed in this section, the rack-level cooling allocates highest power to each single rack housing ICT devices. In the relatively closed internal space of the rack, the intermixing of exhausted air and supply air can be greatly avoided, which gives the shorter airflow path higher controllability [47,48]. In terms of space constraints, another advantage of the rack-level cooling is that modular in-rack coolers can be conveniently deployed into different racks [38]. Since dynamical right-sizing techniques are gradually adopted in data center management, the active server redistribution, based on the real-time workload demands, requires the cooling solution to be flexible, which is one of the excellent features of rack-level cooling [4,49,50]. In comparison, power wastage is unavoidable for room-level or row-level cooling when the devices in some racks are set in the sleep-mode [4]. As a tradeoff, the cost of rack-level cooling installation is relatively high, possibly because of the over-design of the cooling capacity [38]. Another tradeoff is that with a cooler in each rack, the maintenance of rack-level cooling requires a considerable amount of work.

2.2. Mechanical refrigeration of air cooling

The heat absorbed by the TCSS is dissipated to the outdoor environment through dual refrigeration cycles (DRCs) inside the MRSS, as shown in Fig. 5. The DRCs are comprised of a heat transfer cycle and a heat rejection cycle, with the chiller as the connector.

In the heat transfer cycle, cooled water will be continuously produced by the compressor of the chiller through vapor–liquid phase transition. Then, it is pumped into the cooling coils of the CRAC. Circulating into the CRAC from the ceiling vent, warmed airflow will be cooled through the cooling coils. Then, the obtained cold airflow will return to the computer room under the drive of fans. Meanwhile, warmed water flowing out of the cooling coils will return to the chiller, and the cycle is repeated after being re-cooled.

The heat rejection cycle starts by delivering the warmed water, generated in the vapor–liquid phase transition, into the cooling tower. Absorbed heat carried by the warmed water will be dissipated into the ambient environment by the cooling tower. Finally, the obtained cooled water will flow back to the chiller as the cold supply of its compressor, which completes this cycle.

To reduce the power consumption of the MRSS, modeling and optimization methods for cooling equipment are proposed which will be discussed in Section 5. Utilizing ambient environment sources as cold supply, free-cooling technology is another solution, that is described in Section 4.

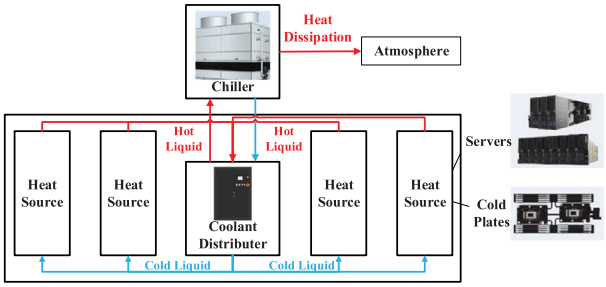


Fig. 6. Schematic of indirect liquid-cooling with coolant distributor.

3. Liquid-cooling technology

Air-cooling is the most common cooling method used in data centers. However, it is an inefficient cooling method due to the low density and heat dissipation capacity of air. In contrast, liquid-cooling is one of the most effective methods, saving total energy consumed compared to air-cooling system. According to how liquid coolant contacts heat source, liquid-cooling methods can be mainly divided into indirect and direct liquid-cooling methods. Similar with air-cooling methods, liquid-cooling methods can also be discussed in terms of MRSS and TCSS. Usually, indirect liquid-cooling consists of MRSS, while direct liquid-cooling consists of TCSS. Different branches of liquid-cooling are summarized in Table 2.

3.1. Mechanical refrigeration of indirect liquid-cooling

Indirect liquid-cooling is a heat dissipation process where the heat sources and liquid coolants contact indirectly. In place of air-cooling radiators, the whole process requires the evaporators or other liquid-cooling radiators [57].

In typical indirect liquid-cooling methods illustrated in Fig. 6, there is a coolant distributor (CD), which is a MRSS shown in Fig. 7, offering a loop to connect the cooling source and the heat source [58]. In other words, coolant from an external cooling source flows into the CD to be delivered to heat sources. Afterwards, the hot liquid stream flows back to the CD and is delivered outside. Meanwhile, the coolant is usually delivered to the processors with the most heat dissipation, while the rest of the heat sources are air-cooled. Combining the characteristics of MRSS, three types of indirect liquid-cooling methods will be discussed separately in the following subsections.

3.1.1. Mechanical refrigeration of single-phase liquid-cooling

Single-phase cooling is a heat transfer process without phase change of the circulating coolant. However, this method has a risk of liquid

Table 2
Summary of liquid-cooling methods.

Mechanism	Subtype [Refs]	Advantages	Main challenges
Indirect cooling	Single-phase cooling [51]	More efficient in thermal conduction than air cooling methods.	Risk of liquid leakage.
	Two-phase cooling [52]	Temperature distribution on heat surface is more uniform and the efficient is higher than single-phase cooling methods.	Coolant should be selected more carefully with lower boiling point.
	Heat-pipe cooling [53]	Reduced risk of liquid leakage and no need for chip-level pump.	Limited heat transfer performance over long distance.
Direct cooling	Pool-boiling cooling [54]	More adaptive with low-cost in maintenance for server components.	Immersion risk.
	Spray-boiling cooling [55,56]	More efficient in heat transfer.	High-cost maintenance for servers.

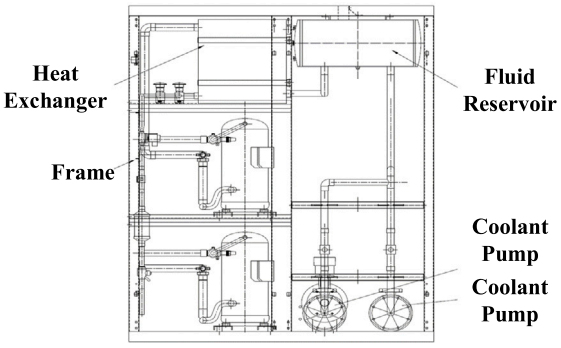


Fig. 7. Structure of coolant distributor.

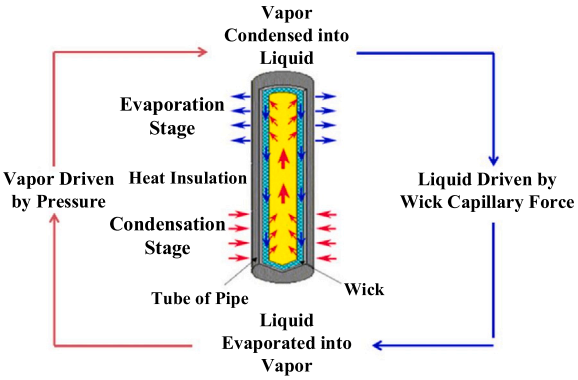


Fig. 8. Working mechanism of heat-pipe cooling.

leakage. Therefore, the mechanical refrigeration sub-system is significant in single-phase liquid-cooling.

In MRSS of single-phase liquid-cooling, the cold plate is a metal plate with high thermal conductivity. The heating device is usually mounted on top of the cold plate, and the liquid coolant circulates in the coolant passages within the cold plate or through the tubes attached to the cold plate [52]. The heat generated by devices is conducted to the coolant channels and then transferred out of the channels by single-phase convection. Based on the MRSS with a cold plate, the risk of liquid leakage can be significantly reduced in a system with multiple processors in series.

The performance of single-phase liquid-cooling methods largely depends on multiple design parameters of cold plates. These parameters include porous media, microchannel radiators, and heat sink pressure [51,59,60].

3.1.2. Mechanical refrigeration of two-phase liquid-cooling

Two-phase cooling is a heat transfer process with phase change of the circulating coolant in the system. The mechanical refrigeration process of two-phase cold plates is similar to the one of single-phase cold plates, with the only difference being that the liquid flows through cold plates and absorbs latent heat when the liquid evaporates. The two-phase cold plate usually needs larger latent heat for phase change of the coolant, a higher heat transfer coefficient and a smaller temperature variation. Also, in two-phase cooling, the temperature gradient on the heat surface is lower and the heat transfer rate is higher, which helps in reducing the flow of the liquid-cooling system while producing a more uniform temperature distribution.

The main focus of improving the two-phase cooling effect is designing efficient porous media and micro-channel radiators at the chip level [52]. However, there are still main disadvantages of the two-phase cooling methods related to flow instability like liquid reversal, and temperature and pressure fluctuation that may eventually cause overheating and burnout of surface [61].

3.1.3. Mechanical refrigeration of heat-pipe cooling

Heat-pipe cooling shown in Fig. 8 is a passive two-phase cooling method driven by the temperature difference between the heat source and radiator. In MRSS of heat-pipe cooling, the steam generated at the end of the evaporator is efficiently transported to the end of condenser through pressure-driven advection. This process forms a condensed fluid that transfers heat generated by devices out of the MRSS. The condensate is then returned to the end of the evaporator by gravity at the thermal siphon or capillary of the heat pipe [53].

One advantage of the heat-pipe cooling method is the reduced risk of liquid leakage inside the server. At the processor level, the sealing device is free from fluid connectors in the MRSS of heat-pipe cooling. In addition, the gravity-driven mechanism of heat-pipe cooling method eliminates the need of a chip-level pump [62], so the entire system can be more stable without problems associated with having too many complicated actuators in MRSS.

The thermal flow characteristics of these cooling methods depend on the choice of various heat pipes and condenser coolants. In terms of processor cooling, loop heat pipes with flat evaporators are preferred to conventional heat pipes because of their geometric compatibility with the environment and better heat transfer capabilities over long distances [63]. The heat pipe coolant can be water, ammonia, methanol or acetone, while the condenser coolant can be liquid or air.

3.2. Terminal cooling of direct liquid-cooling

In the direct liquid-cooling methods, the liquid coolant contacts the electronic devices directly, and the dielectric fluid offers electrical insulation [64]. Because there is no requirement of sealing enclosures or pipes to deliver liquid flow in the server layer, one of the main advantages of direct liquid-cooling is its adaptability and convenience [65]. Thus, the main part of direct liquid-cooling is TCSS. In addition, without pipes or enclosures, the liquid is capable of absorbing heat and cooling all other server components in the whole space of system [64].

Despite the latent heat of phase change and the direct contact between coolant and heat source in TCSS, the thermo physical properties of the fluid medium are significantly lower than water. Therefore, direct cooling method is generally not regarded as a good heat removal technique compared to indirect liquid-cooling methods.

3.2.1. Terminal cooling of pool-boiling cooling

Pool-boiling is a passive all-liquid-cooling method, which means that the electronic plate is completely immersed in liquid coolant [64]. When the temperature of the heat source surface increases beyond the saturation temperature of the medium, the liquid in the cooling tank boils. In TCSS of pool-boiling cooling, there is latent heat transfer, gravity-driven two-phase advection and bubble-induced mixed flow [54]. During pool boiling, steam bubbles formed around the surface of heat source rise to the cooling tank, where the dielectric coolant condenses through a water-cooled heat exchanger to cool the system. The main advantage of a pool-boiling cooling system is the removal of server-level sealed pipes, enclosures and fluid connectors, making this method an adaptable solution with low-cost maintenance for server components.

3.2.2. Terminal cooling of spray-boiling cooling

In the TCSS of spray-boiling method, due to the differential pressure produced by the nozzle plate, the liquid coolant is atomized and then dispersed into small droplets before reaching the heat source surface [55,56]. Spray cooling can be direct or indirect to effectively cool the server components. For the direct method, its disadvantage is the need for high-cost maintenance for servers which are usually sealed in steam chambers containing two-phase coolant flow. For the indirect method, heat transfer occurs in the cold plate. However, due to the similarity between indirect spray-cooling and indirect liquid-cooling method, there are similar disadvantages in practice as mentioned in Sections 3.1.1 to 3.1.3.

4. Free cooling technology

By leveraging natural cold source, free cooling can reduce the overall energy consumption of data centers. As summarized in Refs. [66–68], free cooling technologies can be commonly classified as air-side, water-side and heat pipe based systems. As the names suggest, this classification method derives from the classical technologies described in Sections 2 and 3 . To explicitly present the optimization strategies, we summarize free cooling technologies in the following three aspects: natural cold sources, cooling carriers and heat transfers mechanisms. The overall heat transfer paradigm between natural cold sources and heat sources in a data center is shown in Fig. 9.

4.1. Natural cold sources and cooling carriers

Data centers using free cooling systems are usually located where cold air or water can be provided [69]. Apart from these direct cold sources, energy such as solar energy [70], waste heat of data centers [71], motion energy of seawater [72] and other heat sources can be collected as power supply for mechanical units such as chillers, pumps, cooling towers with CRACs and heat exchangers.

Cooling carriers refer to carriers for directly exchanging heat with heat exchangers or the data center equipment. Natural cold resources may not directly participate in the heat transfer process with internal heat sources. For example, some water-side free cooling systems acquire cold water from auxiliary air cooled or cooling tower systems [73] which are used for cooling the water circulating to internal CRACs working in the economizer mode. Air cooling systems are integrated with dry coolers to directly realize heat exchange between chilled water from the data center and ambient cold air [74]. Cooling carriers can be combined flexibly with different unstable natural cold sources to produce unified cooling resources for efficient and indirect heat transfer schemes.

4.2. Heat transfers mechanisms

Based on whether the cold sources have physical interaction with the data center internal heat sources or not, the heat transfer mechanisms can be classified into direct and indirect categories.

4.2.1. Direct free cooling

Direct air-side free cooling drains the cold fresh air source directly into the data center through the air circulation system [75]. The energy saving effect of this scheme depends directly on the short-term and real-time indoor and outdoor temperature difference [76], as well as the climate environment at the data center location [69]. Contaminants or high humidity air can increase potential risks of equipment aging and damage [77]. Therefore, higher requirements for environmental conditions lead to the restriction on the data center site choices and additional costs for dehumidification and filtration sub-systems.

4.2.2. Indirect free cooling

With indirect free cooling techniques and economizers, data center facilities are shielded from the unstable environmental conditions. Heat exchangers are now responsible for the heat transfer process between internal and external heat sources. Influences and risks driven by extreme weather and polluted air are reduced, resulting in higher stability and longer service life for IT systems [6,78]. In Table 3, a summary of classic heat exchangers is given.

Wheel heat exchangers. For air-side free cooling, wheel heat exchangers such as the Kyoto wheels system realize heat and humidity exchange simultaneously with the rotating wheels [79]. Heat wheels can achieve high cooling efficiency, but they rely on efficient and sufficient exchange surface to reach the cooling requirements for the data centers [80]. Moreover, heat and humidity are exchanged physically on wheels, leading to potential internal air contamination risks.

Plate heat exchangers. Plate heat exchangers can be used to bypass chillers when the outside ambient temperature reaches set points [81]. The outdoor and indoor air is isolated from each other, making plate heat exchangers suitable for situations with polluted air [82]. Nevertheless, the maintenance costs and constrained operating temperature limits the deployment of this kind of exchangers.

Evaporative heat exchangers. Evaporative free cooling systems transfer heat from hot air to condensate water with direct or indirect evaporative strategies [83]. They are appropriate for areas with dry climate and effectively extend the free cooling operation periods. The evaporation process is influenced by multiple factors such as ambient humidity and temperature, air flow velocity and humidity, and the form of contact surface. As a result, it is difficult to model and control the heat transfer process precisely, which can lead to unstable cooling performance.

Heat pipe exchangers. Heat pipes can achieve competitive thermal conductivity, which is an ideal feature for natural cold source exploitation [84]. Heat pipe based free cooling systems are commonly classified into three categories: independent system [85], integrated system [69] and cold storage system [86]. The heat pipe exchangers can be affected by temperature fluctuation when exchanging heat with cold sources. Hence, effective monitoring of the cooling effects of heat pipe based economizers and auxiliary cooling systems should be provided simultaneously to offer more stable cooling performance.

Using natural cold sources and adjusting measures to local conditions, free cooling technologies have the potential to greatly reduce the energy consumption in data centers. However, instability characteristics of the environments, such as temperature and humidity changes, and air pollution, should be taken care of. In brief, both heat transfer efficiency with the natural cold source and sustainable cooling performance under complicated and volatile ambient situations are top technical concerns for free cooling technologies (see Table 4).

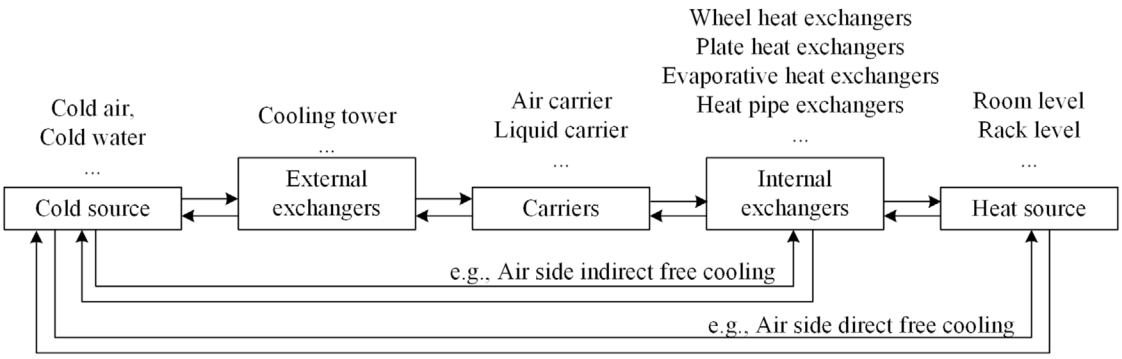


Fig. 9. Free cooling paradigm and examples.

Table 3
Summary of free cooling heat transfer mechanisms.

Mechanism [Refs]	Subtype [Refs]	Advantages	Main challenges
Direct free cooling [75–77]	–	Natural cold source comprehensively blend with internal heat source with low investment costs.	Direct contact brings potential damages to facilities.
Indirect free cooling	Wheel heat exchangers [79,80]	Shield internal space from external unstable cold sources while keeping high heat transfer efficiency.	High maintenance cost. Contamination on centralized wheels can lead to performance degradation.
	Plate heat exchangers [81,82]	Isolation from internal and external environment with scalable heat transfer capacity.	Limited operating temperature and pressure use.
	Evaporative heat exchangers [83]	Isolation from internal and external environment and expansion on free cooling operation time.	Relies on environmental characteristic like temperature and humidity.
	Heat pipe exchangers [69,84–86]	High heat transfer efficiency without external disturbance.	Backup and complementary cooling technologies and devices are needed for failures.

Table 4
Practice for adopting different cooling strategies.

Technology	Advantage	Suitable scenarios
Air cooling	Easy for maintenance and acceptable operation cost.	DC with underfloor space and relatively low heat load.
Liquid cooling	Higher efficiency on thermal conduction with lower environmental impact.	Liquid cooling should be utilized in scenarios with large heat load, where the air cooling cannot meet cooling demand.
Free cooling	Lower overall energy consumption by leveraging natural cold sources.	DC is located where ambient conditions, such as temperature, humidity and air quality are appropriate and can produce sustainable cold sources for cooling processes.

5. Power consumption modeling and optimization methods

In the MRSS of a data center, much power will be consumed for the operations of cooling equipment. Therefore, optimization methods are needed to provide set-points for cooling devices to reduce the power consumption of their operations. This will be introduced in Section 5.1. Power consumption modeling using relations between the cooling equipment measurable variables and the power consumption for the optimization methods will be introduced in Section 5.2.

5.1. Power consumption modeling of cooling equipment

Power consumption modeling takes the measurable variables as inputs to estimate the power consumption of the cooling equipment. As illustrated by Fig. 1, the chiller and the fans of the CRAC are the main power consumers of the cooling system, so their power consumption model will be reviewed here.

Depending on the modeling method, different power consumption models can be mainly divided into mechanism-based methods and data-driven methods. The former means that explicit relations between the input variables and the power consumption are constructed with domain knowledge. This is more suitable for the modeling of relatively simple processes. In contrast, data-driven methods establish gray-box or black-box models, taking the readily measured environmental variables as inputs. The complex nonlinear relations between the inputs and the target can be better captured through historical data analysis, which makes data-driven models more powerful to deal with sophisticated equipment such as the chiller.

5.1.1. Power consumption modeling of the fan

In a data center, the power consumption of the fans can reach more than half of the CPU (central processing unit) power consumption [87,88]. Since the operation of fans is less affected by other equipment, the power consumption can be basically modeled by Eq. (1) [87,89,90]).

$$P = ks^3. \tag{1}$$

Here, s denotes the speed of the fan while k is a coefficient determined by the actual operating condition. In related publications [87,91], k is determined by the temperature difference between the warmed exhausted air and the outside air. When different workloads are considered, then Eq. (1) is extended into Eq. (2) [92].

$$P(t) = P_{base}\left(\frac{R(t)}{R_{base}}\right)^3. \tag{2}$$

Here, $P(t)$ denotes the power consumption of a fan at time t while P_{base} denotes the power consumption when the system is in idle state. The $R(t)$ and the R_{base} denote the revolutions per minute of the fan at time t and in idle state, respectively.

For fans that are mounted in a subsystem [93], the power consumption can be modeled by Eq. (3).

$$P = \frac{V\Delta P}{\prod_i \eta_i}. \tag{3}$$

Here, V and ΔP denote the airflow rate and the pressure head of the fan subsystem, respectively. The η_i denotes the efficiency of a component of the subsystem such as the fan or the motor.

The aforementioned equations are classified as mechanism-based methods. There also exist data-driven methods using data mining techniques to model the power consumption of fans. For example, in [94], neural networks are used in an ensemble way which takes the environmental variables such as the ICT workload, the temperature of the supply air and chilled water as inputs. When evaluated on the test dataset, it achieved 7.21% on mean absolute percentage error (MAPE) which is defined by Eq. (4). The MAPE is defined by

$$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i-\hat{y}_i}{y_i}|. \tag{4}$$

Here, y and $\hat{y}$ denote the real values and estimated ones, while y_i and $\hat{y}_i$ denote the corresponding values on the ith sample, respectively.

5.1.2. Power consumption modeling of the chiller

Given both the internal complex thermodynamic process of the chiller and the external influences from other equipment, it is relatively more difficult to model its power consumption. There exist few publications constructing mechanism-based models to address this issue [95–97]. For example, in [96], mechanism-based models were established to estimate the power consumption of a vapor compression chiller. The first and second laws of thermodynamics were utilized to calculate the generated entropy of the chiller. However, nearly 30 equations were established to describe the thermodynamic processes of the chiller's different components such as the evaporator, compressor and condenser. The complexity of mechanism-based models make them relatively unsuitable for the chiller's power consumption modeling.

A more efficient approach is to use data-driven methods which can be machine learning (ML)-based and empirical based. In the ML-based approach, black-box models are used to estimate the power consumption without requiring prior knowledge. Data of the environmental variables that are related to the power consumption can be collected by sensors, which will be the inputs of the estimation model. The complex relation between the inputs and the estimated target can be obtained from the ML-based algorithms through data regression. Typically, artificial neuron network (ANN)-based models are more generally constructed for power estimation of the chiller [98–101]. For example, in [98], in addition to the ambient condition and controlling setpoints, the season of operation was also taken as the input variables, which effectively reduced the average coefficient of variance of the root mean squared error (CvRMSE) from 29% to 22.8% on the test dataset. The CvRMSE is defined by Eq. (5). The CvRMSE is defined by

$$CvRMSE(y,\hat{y}) = 100\% \times \frac{1}{\bar{y}}\left(\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}\right)^{\frac{1}{2}}. \tag{5}$$

Here, n denotes the number of samples while p denotes the number of variables. The $\bar{y}$ denotes the mean of the label values while other notations are the same as in Eq. (4).

In a recent publication [102], an improved long short-term memory (LSTM) was proposed as the estimation model. Industrial data of the chiller is usually in high complexity, which is hard to be modeled by vanilla LSTM. Therefore, an extra dense layer was placed before the LSTM cell in the improved LSTM model to pre-extract quality features of the input variables. Another dense layer with a dropout layer is also embedded in this model to prevent the over fitting problem. After preprocessing, 7 variables having strong correlation with the power consumption were selected as the inputs, including parameters such as environmental temperatures, pressures and ICT workload. Different from other methods, the proposed LSTM model could output 3 estimated statuses which are the power of the chiller's cold source, the power of refrigeration secondary pump and the Power Usage Effectiveness (PUE) defined by Eq. (6).

$$PUE = \frac{E_{DC}}{E_{ICT}}. \tag{6}$$

Here, E_{DC} and E_{ICT} are the energy consumption of the data center and the ICT equipment, respectively. Compared with other approaches, ML-based methods have the following advantages. The designers only need to determine the input variables and the type of regression model to use. The nonlinear relations of the complex thermodynamic process can be automatically captured by data regression on powerful models. However, a limitation of the ML-based methods is that the training stage requires a sufficient amount of training data for model convergence. Also, data acquisition is sometimes inconvenient when the seasonal operation of the cooling system is considered.

Compared with the black box of the ML-based methods, the empirical branch constructs a gray box which is an equation of the input variables with coefficients to be determined. The equation is initially built with basic domain knowledge as well as practical experience such as its monotonicity. The coefficients will be determined through data regression in which only a small amount of data is required. Some classical empirical models were proposed in early studies [103,104]. For example, in [103], the empirical model is given by Eq. (7).

$$E_a = E_r[a_0 + a_1 PLR + a_2 PLR^2 + a_3\frac{\Delta T_{in}}{\Delta T_{in,des}} + a_4\left(\frac{\Delta T_{in}}{\Delta T_{in,des}}\right)^2 + a_5 PLR\frac{\Delta T_{in}}{\Delta T_{in,des}}]. \tag{7}$$

$$PLR = \frac{L}{L_r}. \tag{8}$$

Here, E_a and E_r denote the actual power and rated power of the chiller. The PLR is the partial load ratio, which denotes the ratio of the actual cooling load L to the chiller's rated capacity L_r. The $\frac{\Delta T_{in}}{\Delta T_{in,des}}$ denotes the ratio of water inlet temperature difference to the designed value. The coefficients $\{a_0, a_1, a_2, a_3, a_4, a_5\}$ were determined through least-squared fitting. However, this equation with only one set of the coefficients cannot track well the actual power consumption when the chiller is operated under different conditions [105]. Therefore, an improved equation based on Eq. (7) was proposed in [105], which is given by Eq. (9).

$$E_{ch} = b_0 + b_1 PLR + b_2 PLR^2, \tag{9}$$

$$with \quad b_0 = a_0 + a_3\frac{\Delta T_{in}}{\Delta T_{in,des}} + a_4\left(\frac{\Delta T_{in}}{\Delta T_{in,des}}\right)^2, \tag{10}$$

$$and \quad b_1 = a_1 + a_5 PLR\frac{\Delta T_{in}}{\Delta T_{in,des}}. \tag{11}$$

Here, the b_2 is the same as the a_2 of Eq. (7). In Eq. (9), while the chiller is working in different condition, the ΔT_{in} will be in a different small range. Thus, for each condition, the ΔT_{in} can be regarded as constant, which makes the b_0 and b_1 constant as well. In this way, the coefficients to determine are $\{b_0, b_1, b_2\}$ under different conditions. Evaluated on the test-set collected from a chiller plant system, Eq. (9) has reached a value of 0.935 on R^2 index which is defined by Eq. (12). The R^2 index is defined by

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}. \tag{12}$$

Here, the $y, \hat{y}, \bar{y}$ denote the label values, the estimated values, and the mean of the label values, respectively.

Another recent study [106] proposed a similar empirical equation which is defined by Eq. (13).

$$E_a = a_0 + a_1 PLR + a_2 PLR^2 + a_3\Delta T_{rs} + a_4{\Delta T_{rs}}^2 + a_5 PLR\Delta T_{rs}. \tag{13}$$

Different from Eq. (7) and (9), the temperature difference between return water and supply water is used for modeling. The authors claimed that the Eq. (13) has reached a value over 0.99 on R^2 index when evaluated on their test-set [106].

Compared with the ML-based methods, the advantage of the empirical methods is that fewer data is required for model fitting. Also, it gets better interpretability and faster computation speed. However, the inputs of the empirical equations are usually internal variables such as the PLR. Thus, the empirical equations cannot incorporate well external influences from other equipment of the cooling system.

5.2. Optimization strategy of cooling equipment

Originally, the operation of the data center cooling system is guided by the technicians with practical experience. With the expansion of equipment and the increase on ICT workload, many optimization strategies are proposed to automatically control the operation of cooling equipment. Given the ICT workload and a cooling target which is usually a temperature range, these strategies provide optimal set-points for the cooling equipment to enhance their power consumption efficiency.

Among the related literature, the MPC-based methods and the RL-based methods are two main approaches [107]. Traditionally, MPC is a major method for the optimal control of dynamic systems while reinforcement learning is mostly used in the computational intelligence applications community. MPC is a model-based method that has mature stability, feasibility and robustness theories and constraint handling. Also, MPC typically has high on-line and low off-line complexities and is not adaptive in a DC cooling system. In contrast, RL is a model-free method that has immature stability, feasibility and robustness theories and constraint handling. Also, RL typically has low on-line and high off-line complexities and is adaptive in a DC cooling system. Although their properties and principles are different, MPC and RL can be jointly used for the same optimization problem to achieve superior results. For example, a MPC controller on the basis of RL model with knowledge enhanced value function and policy, can effectively reduce the execution complexity while preserving stability and feasibility. Therefore, MPC-based and RL-based methods will be introduced in this section, and their optimization strategies are summarized in Table 5.

5.2.1. MPC-based optimization strategy

With the increasing computing power and the maturity of sensing technologies, MPC is becoming a more promising and practical automatic controlling method for modern industrial applications [108]. With a predefined or learnt physical model on the basis of prior knowledge or training data, MPC is suitable for multi-input and multi-output system controlling by translating the optimal control problems into linear or nonlinear optimization problems with certain constraints [109]. Based on a system dynamics model, MPC can generate the optimal control action at current states by optimizing a given future objective on prediction functions.

MPC Framework

The MPC architecture schematic and key interactions are shown in Fig. 10. As depicted, we summarize MPC components into three categories: strategies, models and algorithms. Strategies are the criteria for the optimization process to locate reasonable results in the hypothesis space. There are different kinds of strategies-the reference trajectory for the targeted variables, the constraints derived from environmental variables, and the safety requirements for the objective functions. For MPC applications, both target system model and the environmental disturbance model are needed for robust trajectory tracking. Using the objective function and constraints, optimization algorithms are used to give the best solutions.

The MPC controller regularly collects states from target systems with sensors, and iteratively generates action sequence sets and infers the optimal one by the optimizer. Then, the next predictive action in the sequence will be imposed on the target system. With the feedback loop, MPC can further forecast future system states and at the same time, make timely control adjustments according to the feedbacks.

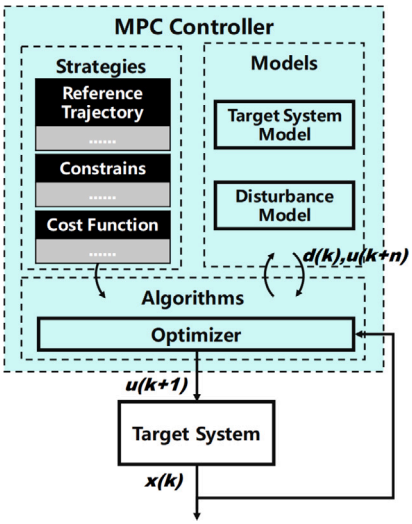


Fig. 10. MPC architecture.

MPC For DC Cooling

MPC-based methods were evaluated in various fields of automatic HVAC system control [7–9]. Depending on the application scenarios, the designs of MPC controllers differ in strategies, models and algorithms. As cooling subsystems are significant energy consumers in DC infrastructures, optimization of the control strategies imposed on targeted MRSS and TCSS components, such as airflow volume [123], fan speeds and valve opening rates [110] is a critical task.

MPC-based methods are used to achieve energy efficient control on multiple cooling applications to minimize economic costs while meeting cooling requirements [110–115]. Linear time-invariant physical models for CPU temperature prediction were created by prediction error and subspace identification methods [111]. These physical models were simulated and compared. Then, a random stepwise disturbance model was considered, and finally, the MPC outputs was converted into a quadratic optimization problem. Similar approaches were evaluated on both direct and indirect air cooling systems and achieved promising results [112]. Furthermore, a physical holistic model which combined the cold and hot air mixing process and cool air flow dynamics to construct the rack inlet temperature model locally and zonally was proposed [113]. In the economic MPC controller [114], a rack-level thermal state model was identified through a data-driven method based on a state–space model and the linear programming problem was defined at the same time. A more fine-grained MPC controller at room and floor level and with constant disturbances was proposed [110]. The DC dynamic model was based on a linear auto-regressive model with reinforcement learning. A quadratic optimization solver was implemented with TensorFlow computation graph. It is difficult for the single setpoint MPC schemes to settle the non-uniform cooling problem for different regions in a data center. To overcome this limitation, the required temperature distribution matrix based multi-setpoint MPC method was introduced and learnt from historical data [115]. As a result, zones with different temperature constraints can make adaptive and specific adjustments according to MPC decisions.

MPC can also act as coordinated controller for both IT systems and cooling systems, where QoS (quality-of-service) performance and energy efficient objectives are jointly considered [116,117]. Models concerning computation tasks, power usage and thermal dynamics are stacked to predict states driven by layered controller actions [116]. Constraints for both environmental temperature and QoS of servers are set for the sequential quadratic programming optimization process.

As summarized above, MPC is a representative model based method, well-suited for energy saving while meeting the cooling requirements.

Table 5
Summary of power modeling methods and optimization methods for cooling equipment operation.

Type	Methods [Refs]	Advantages	Main challenges
Power modeling method	Mechanism-based methods [89–93]	Interpretable models based on priori knowledge.	Hard to model complex operation process.
	Data-driven methods [98–101,103–106]	More powerful models that can handle complex process.	Hard to solve the problems of overfitting and data paucity.
Optimization algorithm	Model predictive control [110–117]	Achieve optimal step by step control actions through optimization methods under safe boundary conditions.	Hard to model physical dynamics and computationally intensive.
	Reinforcement learning [10,118,119]	More real-time and dynamic than traditional control models.	Slow convergence and faults on early stages.
	Deep reinforcement learning [120–122]	More capable of extracting features with deep learning modules.	Requirements of a large amount of operational data which is difficult to collect in real system.

With reasonable constraints, cooling performance stability can be guaranteed with optimal power consumption and without unexpected function failure.

5.2.2. RL-based optimization strategy

Reinforcement learning [118] is a trial-and-error learning algorithm made of a policy, a reward function and an decision model of the environment. In order to learn the optimal policy evaluated by the reward function, an agent of RL interacts with the environment through a sequence of actions then gets corresponding reward from the state of the environment.

Based on RL, DRL replaces the decision function in RL with a deep learning neural network, which can better map actions to rewards in a more complicated environment. This replacement mainly brings two advantages. First, the representation capacity of the agent is greatly improved. Second, the control of the agent is upgraded into an end-to-end way, which helps in achieving better performance. The most common DRL method utilized deep Q-network (DQN) [120] as the function approximation of the control policy for the agent [122]. Other typical DRL methods mainly include Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor Critic (A3C) [124] and Unsupervised Reinforcement and Auxiliary Learning (UNREAL) [125].

Application of RL and DRL in Cooling Systems

The recent advances of AI have allowed for RL [118] and DRL [120] to be widely adopted in the control optimization of a data center cooling infrastructure and other HVAC systems. Because of the slow convergence on early learning stages of RL, a multi-grid method for Q-learning was proposed for HVAC control systems in buildings [10]. Simulation results indicated the effectiveness of energy conservation and faster convergence of RL. Spatially aware workload and thermal management system (SPAWM) for data centers [126] used RL to optimize the data center's workload, thermal distribution and cooling facilities, resulting in a reduction of the max inlet temperature of 2 to 3°C. Combined with free cooling methods which mixes the rack outlet air and outdoor air to rack inlet, the DRL method in [13] was used to control the temperature and humility of the data center in an appropriate range.

For DRL methods applied to cooling systems, deep learning module by [11] was improved to select the corresponding optimal air flow rates for different zones in the whole building. To minimize energy consumption and improve tenants' comfort, a DDPG approach for energy optimization and thermal comfort control was proposed in [12]. In addition to DDPG, DQN was adopted in [119] to realize end-to-end control for a radiant heating system. Compared to rule-based control, experiment shows that the DQN agent reduced by 18.2% the cooling demand. Recently, performance among DQN, Branching Dueling Q-Network (BDQ), and DDPG for data center cooling management were compared together using the Active Ventilation Tiles (AVTs) control problem as case study [127], in which results indicated that DQN provided the best performance.

Evaluation of RL and DRL in Cooling System

In cooling methods, online RL often suffers from slow convergence and faults in the early stages. Also, the trial-and-error nature of RL sometimes results in unwise actions and unsatisfying online performance, which may lead to thermal comfort problems [10]. Therefore, agents in most RL-based models are first trained offline with a HVAC simulator or historical data and then fine-tuned online [119].

A DRL agent called Gnu-RL [121] was pre-trained offline on historical data from imitation learning to a proportional–integral–derivative (PID) controller. Afterwards, Gnu-RL was improved by online training while interacting with the environment. According to experiments in a real conference room, Gnu-RLs end-to-end control policy focusing on the amount of air flow supplied could save 16.7% of cooling demand, compared to conventional rule-based control. However, it is challenging for RL to deal with complex thermal and psychrometric dynamics, due to the limitation of RL's features extraction. DRL addresses well the problem with better capability from deep learning module.

In consideration of practical use, HVAC has some constraints on parameters of cooling methods, such as temperature and humidity of cooling air for data centers. To minimize cooling energy without violation to these constraints, a DRL approach [122] was proposed to adaptively learn optimal penalty weights based on Lagrangian primal–dual policy optimization. The penalty weights helped the parameters of optimal policy converge effectively with satisfying their constraints. These experiments demonstrated that their constrained DRL approach achieved less cooling energy consumption and less violations of constraints, compared to unconstrained DRL methods.

6. Current challenges and future work

Based on the discussions of technology, power consumption modeling and control strategy optimization of the cooling system of a data center, several challenges are identified. Some current and future challenges are shown in Fig. 11.To focus our discussion, we select three important current challenges and present possible solutions. Also, to build a cooling system with higher energy efficiency for the CPSS's data center, some additional considerations should be investigated and implemented. The performance characteristics and the state of progress of these challenges are summarized in Table 6.

6.1. Current challenges

6.1.1. Cooling rearrangement for changes on network topologies

Recently, researchers are working on designing novel network topologies to bring more connectivity between services and reduce latencies in complex business scenarios [128]. The network topology changes will lead to changes on related device environments, such as spatial arrangements of servers, hardware densities and pipeline wiring. In consequence, the performance and adaptability of the cooling systems will meet challenges in several aspects, like arrangements of cold and hot aisles.

Table 6
Summary of current challenge and future work.

Types	Issues	Performance characteristics	State of progress
Current challenges	Cooling rearrangement for changes on network topologies.	Novel network topologies can lead to changes to device environments and further bring adaptability challenges to cooling performance.	Miniaturization of TCSS and adjusting cold and hot aisles automatically to local computer room conditions.
	Waste heat reutilization for district heating.	High quality heat is required to satisfy the district cooling demand.	Two-phase cooling system.
	Power consumption modeling of the entire system.	Individual modeling cannot resolve the problem of coupling effects among different cooling equipment.	Deep learning models with feature selection techniques.
	Optimization for edge DC cooling.	Provide sufficient cooling power in smaller rooms and handle unstable edge computing cooling demands.	Leveraging cooling technologies such as liquid cooling and free cooling, and automated monitoring and controlling methods.
Future work	Designing better cooling systems using natural sources.	Leveraging natural cold source with free cooling is promising to further reduce energy consumption.	Integrated free cooling technologies with fine-grained controlling.
	Proposing control models with higher accuracy.	The precision and timeliness of the model construction directly affect the design of cooling subsystem and the effect after deployment.	Modeling the thermal behavior of different cold sources, such as air and liquid, and at different levels such as room level, rack level and so on.
	Reducing DC cooling operation costs with digital twin and analytics.	Real-time data sensing and modeling techniques.	Incorporating with computational fluid dynamics analysis.
	Capturing and using data created in DCs.	Construction of DCs with enough sensors and electrical infrastructure.	AI technologies allow finding working patterns of DCs for further optimization.
	Achieving human-free management in DCs.	Improve automation and reduce maintenance cost.	Adaptive optimization strategy & Intelligent patrol-robots.

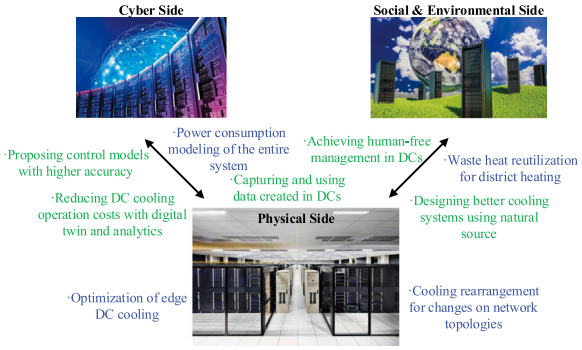


Fig. 11. The current (blue) and future challenge (green) on data center cooling system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To embrace these technology changes, miniaturization of the TCSS can offer opportunities for convenient assembling and maintaining. Moreover, adjusting cold and hot aisles automatically to local computer room conditions, can effectively eliminate intermixing between cold and hot aisles and improve overall power efficiency.

6.1.2. Waste heat reutilization for district heating

After being absorbed by the coolant, the waste heat generated by the ICT devices can be reutilized for district heating, if there is heating demand near the data center's location [129]. However, an important challenge is due to the poor quality of the waste heat, which means that its temperature is relatively low. A possible solution is to use liquid cooling such as spray cooling, in which the temperature of the captured heat can reach to 60 °C. Additionally, using a two-phase cooling system can further increase the temperature of the captured heat to 75 °C, which better satisfies the district heating demands [130]. However, the cost brought by the extra heat reutilization equipment should be carefully studied to ensure that the reutilization is a revenue stream and not a burden for the data center.

6.1.3. Power consumption modeling of the entire system

Traditionally, to obtain the total power consumption of the cooling system, different parts of the cooling system were modeled individually. The sum of these models' outputs will be the target to reduce. However, there exists coupling effects in the MRSS which means that the operation of some equipment will also influence other adjacent equipment. This makes the above scheme difficult to provide accurate estimations for the power consumption of different cooling equipment. A possible solution is to establish one holistic model to estimate the total power consumption of the entire system, which transforms it into an end-to-end problem. Given various system parameters as inputs, a powerful model that is capable of tracing the complex relations among the system's components is required. In some recent studies, machine learning models such as LSTM are used for modeling of the entire cooling loop in a HPC (high-performance computer) data center [131]. However, apart from the model establishment, the selection of the input variables also requires further studies, but this could cause overfitting problems.

6.1.4. Optimization of edge data center cooling

Current trends are for more edge computing scenarios to satisfy the application needs to lower the transfer latency and network overhead, as well as to enhance privacy and data protection [132]. With smaller space and more high-density hardware in edge computing, more challenges are brought to the data center's cooling system. First, integrated cooling units and architecture should provide sufficient cooling power in smaller physical spaces. Second, loads on edge computing units are more unpredictable and unstable, and computing request bursts have higher requirements for more fine-grained cooling methods. Third, compared to centralized data centers, the maintenance costs on cooling equipment and controlling strategies for widely distributed edge data centers are considerable.

To adapt to the edge-cloud computing paradigm trends, cooling technologies and corresponding cooling optimization methods need to evolve. Liquid cooling methods are promising cooling techniques with higher heat dissipation capacity for edge DCs compared to air cooling, especially in high-density computer rooms. Free cooling technologies can also be considered as supplements for cooling, while the volume of the equipment and complex surrounding environment should be considered to provide stable cooling power. Moreover, automation, such as continuous monitoring with accident prediction and unified cooling strategy management, can greatly reduce maintenance costs which is very beneficial for distributed edge DCs.

6.2. Future work

The increasing demands for more computing resources in larger and larger data centers presents many opportunities for future research in their cooling technologies. In this section, we present five areas for future work on cooling systems for data centers. They include the use natural cooling sources, more accurate control models, application of digital twin technology, collecting and using data related to thermal management for optimized operation, and autonomous management.

6.2.1. Designing better cooling systems using natural sources

In terms of cooling technologies, future work should design and implement readily available, highly reliable and safe hybrid cooling systems. Such systems should make use of natural cooling sources according to local conditions, so as to reduce the energy consumption and ensure the continuity and stability of cooling systems [133].

Free cooling methods using a natural cooling source is promising, but there are few cooling subsystems only relying on free cooling because of the instabilities of a natural cold source due to factors such as temperature, humidity and pollution. Therefore, an integrated system combining free cooling and other traditional cooling technologies can be one of the primary future works. This work involves improving control policies to choose between free cooling and other cooling methods to prevent the failure of any subsystem and ensure the overall stability of the cooling system.

6.2.2. Proposing control models with higher accuracy

In terms of control models, building a real-time and dynamic thermal behavior model will help to realize the accurate prediction and management of energy consumption in multiple levels and dimensions. A system with a more advanced thermal behavior model can maintain low energy consumption while meeting the computing needs of data centers with a heterogeneous complex cooling system.

The thermal behavior model should consider the dynamic thermal characteristics of air, liquid and other different cold sources at room level, rack level and other levels. Then, the cooling system can realize accurate prediction of energy consumption and fine-grained optimization of model parameters and states. Thus, the system can deal with the problem of load fluctuation caused by the sudden change of business flow or computing needs in CPSS.

6.2.3. Reducing data center cooling operation costs with digital twin and analytics

Digital Twin [134] represents both a physical and a virtual world in which physical entities or products have dynamic digital representations. The data center industry is quickly evolving to meet the enormous rise in big data processing needs for many applications including CPSS. Therefore, for data centers, digital twins can be a key technology to investigate these evolving data processing needs, especially in optimizing cooling strategies. Digital twins can comprehensively perceive several interacting entities, such as cooling facilities, DC operators and computing tasks at the same time. Therefore, accurate monitoring, simulation and prediction on dynamic cooling behaviors and outcomes can be analytically achieved in a real-time manner. Key properties including computation, control, and communication can also be analytically described by digital twin [135]. With Artificial Intelligence for IT operations, when the data center expands its scale, it can dynamically adjust the layout of the facilities in its cooling system based on the simulation of digital twins. Moreover, it is easier to identify cooling performance anomalies and the corresponding root causes accurately with less human inspection based on the digital twin simulations. That is, the digital twin enables more accurate manipulation and less maintenance of cooling systems, thus contributing to better cloud services in data centers for important applications such as CPSS services, internet-of-things, 3D graphics or gaming.

6.2.4. Capturing and using data created in data centers

Currently, data centers are generating significant volumes of data about their power usage, heat emission and cooling demands. Capturing the data generated by data centers themselves is a complex process. However, capturing and utilizing such data is valuable for optimization of the data center's operation and thermal management. Constructing data centers with enough sensors and electrical infrastructure will allow for the optimized thermal management and use of data centers, and provide useful insights for the construction of new data centers. With such data collected, more real-time online control can be achieved to avoid overloading caused by excessively high usage. Besides, by having the data collected from data centers, AI technologies can then allow us find working patterns of data centers for further optimized operation. For instance, data-driven models can optimize cooling pattern by enabling dynamic adaptive cooling control and load balancing.

6.2.5. Achieving human-free management in data centers

The lockdown brought by COVID-19 has highlighted the importance of the remote management of data centers. Current and future data centers should be capable of withstanding higher workloads during emergency, with fewer direct manual regulation. To ensure the stable operation in such cases, two techniques are required.

The first technique is the remote monitoring and control of data center equipment. For the ICT devices, both their workloads and their physical properties such as working temperature should be sensed and transmitted to the monitoring system. For maintenance equipment, their operation status should be flexibly adjusted remotely so that higher cooling power can be deployed to the cooling system in time.

The second technique is the self-regulation capability of data centers. An example of the self-regulation is the adjustment of cooling equipment set-points introduced in Section 5.2. Also, for the ICT devices, the automatic job assignment and load balancing are valuable research avenues that can greatly benefit the data center operation in different job and loading scenarios.

7. Conclusions

A data center is the fundamental infrastructure in the era of big data. The cooling system of a data center is an indispensable component to provide a suitable operation temperature for the information and communications technology devices. This paper presents an survey of the technology, power consumption modeling and control strategy optimization of a data center's cooling system. Current cooling solutions are classified into air-cooling, liquid-cooling and free cooling technology. The mechanisms of their mechanical refrigeration part and terminal cooling part were discussed, and their advantages and limitations are summarized. Also, the proposed approaches to estimate the power consumption of the major energy consumers of the cooling system, are classified into mechanism-based methods and data-driven methods. Moreover, the optimization strategies proposed to regulate the operation of the cooling devices are summarized. These strategies include the MPC-based methods and the RL-based methods. With increasing scale, data centers have to face challenges on cooling system management which are summarized in this paper with the possible solutions. Finally, some emerging techniques are also discussed which can improve the data center cooling performances in future.

Declaration of competing interest

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

References

[1] M. Dayarathna, Y. Wen, R. Fan, Data center energy consumption modeling: A survey, IEEE Commun. Surv. Tutor. 18 (1) (2015) 732–794.

[2] A.H. Khalaj, S.K. Halgamuge, A review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system, Appl. Energy 205 (2017) 1165–1188.

[3] B. Whitehead, D. Andrews, A. Shah, G. Maidment, Assessing the environmental impact of data centres part 1: Background, energy use and metrics, Build. Environ. 82 (2014) 151–159.

[4] J. Wan, X. Gui, S. Kasahara, Y. Zhang, R. Zhang, Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey, IEEE Access 6 (2018) 48867–48901.

[5] S. Alkharabsheh, J. Fernandes, B. Gebrehiwot, D. Agonafer, K. Ghose, A. Ortega, Y. Joshi, B. Sammakia, A brief overview of recent developments in thermal management in data centers, J. Electron. Packag. 137 (4) (2015).

[6] A. Capozzoli, G. Primiceri, Cooling systems in data centers: state of art and emerging technologies, Energy Procedia 83 (2015) 484–493.

[7] G. Serale, M. Fiorentini, A. Capozzoli, D. Bernardini, A. Bemporad, Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: Problem formulation, applications and opportunities, Energies 11 (3) (2018) 631.

[8] A. Afram, F. Janabi-Sharifi, A.S. Fung, K. Raahemifar, Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system, Energy Build. 141 (2017) 96–113.

[9] A. Schirrer, M. Brandstetter, I. Leobner, S. Hauer, M. Kozek, Nonlinear model predictive control for a heating and cooling system of a low-energy office building, Energy Build. 125 (2016) 86–98.

[10] B. Li, L. Xia, A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings, in: 2015 IEEE International Conference on Automation Science and Engineering (CASE), IEEE, 2015, pp. 444–449.

[11] T. Wei, Y. Wang, Q. Zhu, Deep reinforcement learning for building HVAC control, in: Proceedings of the 54th Annual Design Automation Conference 2017, 2017, pp. 1–6.

[12] G. Gao, J. Li, Y. Wen, Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning, 2019, arxiv preprint arXiv:1901.04693.

[13] D. Van Le, Y. Liu, R. Wang, R. Tan, Y.-W. Wong, Y. Wen, Control of air free-cooled data centers in tropics via deep reinforcement learning, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 306–315.

[14] Q. Zhang, C. Tang, T. Bai, Z. Meng, Y. Zhan, J. Niu, M.J. Deen, A two-layer optimal scheduling framework for energy savings in a data center for cyber–physical–social systems, J. Syst. Archit. 116 (2021) 102050.

[15] W. Jiang, L. Wen, J. Zhan, K. Jiang, Design optimization of confidentiality-critical cyber physical systems with fault detection, J. Syst. Archit. 107 (2020) 101739.

[16] X. Wang, L.T. Yang, X. Chen, M.J. Deen, J. Jin, Improved multi-order distributed HOSVD with its incremental computing for smart city services, IEEE Trans. Sustain. Comput. PP (2018) 1.

[17] S. Majumder, T. Mondal, M. Deen, Wearable sensors for remote health monitoring, Sensors 17 (2017) http://dx.doi.org/10.3390/s17010130.

[18] S. Majumder, L. Chen, O. Marinov, C. Chen, T. Mondal, M.J. Deen, Noncontact wearable wireless ECG systems for long-term monitoring, IEEE Rev. Biomed. Eng. 11 (2018) 306–321, http://dx.doi.org/10.1109/RBME.2018.2840336.

[19] E. Nemati, M.J. Deen, T. Mondal, A wireless wearable ECG sensor for long-term applications, IEEE Commun. Mag. 50 (1) (2012) 36–43, http://dx.doi.org/10.1109/MCOM.2012.6122530.

[20] M. Deen, Information and communications technologies for elderly ubiquitous healthcare in a smart home, Pers. Ubiquitous Comput. 19 (2015) 573–599, http://dx.doi.org/10.1007/s00779-015-0856-x.

[21] S. Majumder, E. Aghayi, M. Noferesti, H. Memarzadeh-Tehran, T. Mondal, Z. Pang, M. Deen, Smart homes for elderly healthcare-recent advances and research challenges, Sensors 17 (2017) 2496, http://dx.doi.org/10.3390/s17112496.

[22] Y. Liu, L. Zhang, Y. Yang, L. Zhou, L. Ren, F. Wang, R. Liu, Z. Pang, M.J. Deen, A novel cloud-based framework for the elderly healthcare services using digital twin, IEEE Access 7 (2019) 49088–49101, http://dx.doi.org/10.1109/ACCESS.2019.2909828.

[23] M. García-Valls, A. Dubey, V. Botti, Introducing the new paradigm of social dispersed computing: applications, technologies and challenges, J. Syst. Archit. 91 (2018) 83–102.

[24] M. Wazid, A.K. Das, R. Hussain, G. Succi, J.J. Rodrigues, Authentication in cloud-driven IoT-based big data environment: Survey and outlook, J. Syst. Archit. 97 (2019) 185–196, http://dx.doi.org/10.1016/j.sysarc.2018.12.005.

[25] E. Butler, The future of data centers in the post-Covid world, 2021, https://www.datacenterdynamics.com/en/opinions/future-data-centers-post-covid-world/, Accessed July 4, 2021.

[26] Y. Taniguchi, K. Suganuma, T. Deguchi, G. Hasegawa, Y. Nakamura, N. Ukita, N. Aizawa, K. Shibata, K. Matsuda, M. Matsuoka, Tandem equipment arranged architecture with exhaust heat reuse system for software-defined data center infrastructure, IEEE Trans. Cloud Comput. 5 (2017) 182–192.

[27] R. Bob Sullivan, G. Li, X. Zhang, Cold aisle or hot aisle containment - Is one better than the other? in: 2018 IEEE International Telecommunications Energy Conference (INTELEC), 2018, pp. 1–4, http://dx.doi.org/10.1109/INTLEC.2018.8612444.

[28] P. Lin, V. Avelar, How row-based data center cooling works, 2014, APC White Paper 208.

[29] T. Gao, B.G. Sammakia, J. Geer, B. Murray, R. Tipton, R. Schmidt, Comparative analysis of different in row cooler management configurations in a hybrid cooling data center, in: International Electronic Packaging Technical Conference and Exhibition, Vol. 56888, American Society of Mechanical Engineers, 2015, V001T09A011.

[30] S.K. Shrivastava, J.W. VanGilder, System and method for arranging equipment in a data center, 2012, US Patent 8, 219, 362.

[31] Y.M. Manaserh, M.I. Tradat, G. Mohsenian, B.G. Sammakia, M.J. Seymour, General guidelines for commercialization a small-scale in-row cooled data center: a case study, in: 2020 36th Semiconductor Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), IEEE, 2020, pp. 48–55.

[32] U. Chowdhury, W. Hendrix, T. Craft, W. James, D. Agonafer, Optimal design and modeling of server cabinets with in-row coolers and air conditioning units in a modular data center, in: ASME 2019 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems, 2019.

[33] S.K. Shrivastava, J.W. VanGilder, System and method for arranging equipment in a data center, 2018, US Patent 9, 996, 659.

[34] C.-H. Wang, Y.-Y. Tsui, C.-C. Wang, Airflow management on the efficiency index of a container data center having overhead air supply, J. Electron. Packag. 139 (4) (2017).

[35] V. Sorell, Raised floor versus overhead cooling in data centers, in: Data Center Handbook, John Wiley & Sons, 2014, pp. 429–439, http://dx.doi.org/10.1002/9781118937563.ch23, Ch. 23.

[36] R. Grantham, K. Lemke, Method and apparatus for installation and removal of overhead cooling equipment, 2013, US Patent 8, 405, 982.

[37] P.-C. Chen, Server and cooler moduel arrangement, 2011, US Patent 8, 045, 328.

[38] K. Dunlap, N. Rasmussen, Choosing between room, row, and rack-based cooling for data centers, 2012, URL https://download.schneider-electric.com/files?p_Doc_Ref=SPD_VAVR-6J5VYJ_EN.

[39] A.C. Kheirabadi, D. Groulx, Cooling of server electronics: A design review of existing technology, Appl. Therm. Eng. 105 (2016) 622–638.

[40] X. Zhang, T. Lindberg, N. Xiong, V. Vyatkin, A. Mousavi, Cooling energy consumption investigation of data center IT room with vertical placed server, Energy Procedia 105 (2017) 2047–2052, 8th International Conference on Applied Energy, ICAE2016, 8-11 October 2016, Beijing, China.

[41] M. Sahini, E. Kumar, T. Gao, C. Ingalz, A. Heydari, S. Xiaogang, Study of air flow energy within data center room and sizing of hot aisle containment for an active vs passive cooling design, in: 2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2016, pp. 1453–1457.

[42] K. Nemati, H.A. Alissa, B.T. Murray, B. Sammakia, Steady-state and transient comparison of cold and hot aisle containment and chimney, in: 2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2016, pp. 1435–1443, http://dx.doi.org/10.1109/ITHERM.2016.7517717.

[43] Y.U. Makwana, A.R. Calder, S.K. Shrivastava, Benefits of properly sealing a cold aisle containment system, in: Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014, pp. 793–797.

[44] W.-X. Chu, R. Wang, P.-H. Hsu, C.-C. Wang, Assessment on rack intake flowrate uniformity of data center with cold aisle containment configuration, J. Build. Eng. 30 (2020) 101331.

[45] C.-H. Wang, Y.-Y. Tsui, C.-C. Wang, On cold-aisle containment of a container datacenter, Appl. Therm. Eng. 112 (2017) 133–142.

[46] S. Chapel, W. Pachoud, Air-based cooling for data center rack, 2013, US Patent 8, 453, 471.

[47] R. Das, J.O. Kephart, J. Lenchner, H. Hamann, Utility-function-driven energy-efficient cooling in data centers, in: Proceedings of the 7th International Conference on Autonomic Computing, 2010, pp. 61–70.

[48] R. You, J. Chen, Z. Shi, W. Liu, C.-H. Lin, D. Wei, Q. Chen, Experimental and numerical study of airflow distribution in an aircraft cabin mock-up with a gasper on, J. Build. Perform. Simul. 9 (5) (2016) 555–566.

[49] M. Lin, A. Wierman, L.L. Andrew, E. Thereska, Dynamic right-sizing for power-proportional data centers, IEEE/ACM Trans. Netw. 21 (5) (2012) 1378–1391.

[50] D. Shen, J. Luo, F. Dong, X. Fei, W. Wang, G. Jin, W. Li, Stochastic modeling of dynamic right-sizing for energy-efficiency in cloud data centers, Future Gener. Comput. Syst. 48 (2015) 82–95.

[51] Y.J. Lee, P.K. Singh, P.S. Lee, Fluid flow and heat transfer investigations on enhanced microchannel heat sink using oblique fins with parametric study, Int. J. Heat Mass Transfer 81 (2015) 325–336.

[52] Z. Li, S.G. Kandlikar, Current status and future trends in data-center cooling technologies, Heat Transfer Eng. 36 (6) (2015) 523–538.

[53] B. Siedel, V. Sartre, F. Lefèvre, Literature review: Steady-state modelling of loop heat pipes, Appl. Therm. Eng. 75 (2015) 709–723.

[54] L. Qiu, S. Dubey, F.H. Choo, F. Duan, Recent developments of jet impingement nucleate boiling, Int. J. Heat Mass Transfer 89 (2015) 42–58.

[55] J. Kim, Spray cooling heat transfer: The state of the art, Int. J. Heat Fluid Flow 28 (4) (2007) 753–767.

[56] E.A. Silk, E.L. Golliher, R.P. Selvam, Spray cooling heat transfer: technology overview and assessment of future challenges for micro-gravity application, Energy Convers. Manage. 49 (3) (2008) 453–468.

[57] H. Geng, Data Center Handbook, John Wiley & Sons, 2014.

[58] M.J. Ellsworth, G.F. Goth, R.J. Zoodsma, A. Arvelo, L.A. Campbell, W.J. Anderl, An overview of the IBM power 775 supercomputer water cooling system, J. Electron. Packag. 134 (2) (2012) 35–43.

[59] R. Singh, A. Akbarzadeh, M. Mochizuki, Sintered porous heat sink for cooling of high-powered microprocessors for server applications, Int. J. Heat Mass Transfer 52 (9–10) (2009) 2289–2299.

[60] E.M. Dede, Y. Liu, Experimental and numerical investigation of a multi-pass branching microchannel heat sink, Appl. Therm. Eng. 55 (1–2) (2013) 51–60.

[61] S.T. Kadam, R. Kumar, Twenty first century cooling solution: Microchannel heat sinks, Int. J. Therm. Sci. 85 (2014) 73–92.

[62] R.J. McGlen, R. Jachuck, S. Lin, Integrated thermal management techniques for high power electronic devices, Appl. Therm. Eng. 24 (8–9) (2004) 1143–1156.

[63] Y.F. Maydanik, M.A. Chernysheva, V. Pastukhov, Loop heat pipes with flat evaporators, Appl. Therm. Eng. 67 (1–2) (2014) 294–307.

[64] A. Bar-Cohen, M. Arik, M. Ohadi, Direct liquid cooling of high flux micro and nano electronic components, Proc. IEEE 94 (8) (2006) 1549–1570.

[65] P.E. Tuma, The merits of open bath immersion cooling of datacom equipment, in: 2010 26th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), IEEE, 2010, pp. 123–131.

[66] H. Zhang, S. Shao, H. Xu, H. Zou, C. Tian, Free cooling of data centers: A review, Renew. Sustain. Energy Rev. 35 (2014) 171–182.

[67] C. Nadjahi, H. Louahlia, S. Lemasson, A review of thermal management and innovative cooling strategies for data center, Sustain. Comput.: Inform. Syst. 19 (2018) 14–28.

[68] Y. Zhang, Z. Wei, M. Zhang, Free cooling technologies for data centers: energy saving mechanism and applications, Energy Procedia 143 (2017) 410–415.

[69] K.-P. Lee, H.-L. Chen, Analysis of energy saving potential of air-side free cooling for data centers in worldwide climate zones, Energy Build. 64 (2013) 103–112.

[70] T. Malkamäki, S.J. Ovaska, Solar energy and free cooling potential in European data centers, Procedia Comput. Sci. 10 (2012) 1004–1009.

[71] Z. He, T. Ding, Y. Liu, Z. Li, Analysis of a district heating system using waste heat in a distributed cooling data center, Appl. Therm. Eng. 141 (2018) 1131–1140.

[72] J. Clidaras, D.W. Stiver, W. Hamburgen, Water-based data center, 2009, US Patent 7, 525, 207.

[73] J. Li, Z. Li, Model-based optimization of free cooling switchover temperature and cooling tower approach temperature for data center cooling system with water-side economizer, Energy Build. 227 (2020) 110407.

[74] T. Gao, M. David, J. Geer, R. Schmidt, B. Sammakia, A dynamic model of failure scenarios of the dry cooler in a liquid cooled chiller-less data center, in: 2015 31st Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), IEEE, 2015, pp. 113–119.

[75] E. Oró, V. Depoorter, N. Pflugradt, J. Salom, Overview of direct air free cooling and thermal energy storage potential energy savings in data centres, Appl. Therm. Eng. 85 (2015) 100–110.

[76] D. Vos, Reducing the data center energy costs through the implementation of short-term thermal energy storage, in: Proceedings in the 8th Int Renew EnergStorConf (IRES 2013), Berlin, 2013.

[77] V. Sorell, OA economizers for data centers, Ashrae J. 49 (12) (2007) 32–34+36–37.

[78] Y. Udagawa, S. Waragai, M. Yanagi, W. Fukumitsu, Study on free cooling systems for data centers in Japan, in: Intelec 2010, IEEE, 2010, pp. 1–5.

[79] Z. Potts, Free cooling technologies in data centre applications, in: SUDLOWS White Paper, Manchester, 2011.

[80] R. Sullivan, M. Van Dijk, M. Lodder, Introducing using the heat wheel to cool the computer room, ASHRAE Trans. 115 (2) (2009) 187–192.

[81] T.M. Abou Elmaaty, A. Kabeel, M. Mahgoub, Corrugated plate heat exchanger review, Renew. Sustain. Energy Rev. 70 (2017) 852–860.

[82] R.C. Chu, M.J. Ellsworth Jr, R.R. Schmidt, R.E. Simons, Scalable coolant conditioning unit with integral plate heat exchanger/expansion tank and method of use, 2004, US Patent 6, 714, 412.

[83] S. Shao, H. Liu, H. Zhang, C. Tian, Experimental investigation on a loop thermosyphon with evaporative condenser for free cooling of data centers, Energy 185 (2019) 829–836.

[84] T. Ding, Z. guang He, T. Hao, Z. Li, Application of separated heat pipe system in data center cooling, Appl. Therm. Eng. 109 (2016) 207–216.

[85] D.-d. Zhu, D. Yan, Z. Li, Modelling and applications of annual energy-using simulation module of separated heat pipe heat exchanger, Energy Build. 57 (2013) 26–33.

[86] B. Zalba, J.M. Marín, L.F. Cabeza, H. Mehling, Free-cooling of buildings with phase change materials, Int. J. Refrig. 27 (8) (2004) 839–849.

[87] M. Dayarathna, Y. Wen, R. Fan, Data center energy consumption modeling: A survey, IEEE Commun. Surv. Tutor. 18 (1) (2016) 732–794, http://dx.doi.org/10.1109/COMST.2015.2481183.

[88] A. Vasan, A. Sivasubramaniam, V. Shimpi, T. Sivabalan, R. Subbiah, Worth their watts?-an empirical study of datacenter servers, in: HPCA-16 2010 the Sixteenth International Symposium on High-Performance Computer Architecture, IEEE, 2010, pp. 1–10.

[89] S. Yeo, M.M. Hossain, J.-C. Huang, H.-H.S. Lee, ATAC: Ambient temperature-aware capping for power efficient datacenters, in: Proceedings of the ACM Symposium on Cloud Computing, 2014, pp. 1–14.

[90] J. Kim, M.M. Sabry, D. Atienza, K. Vaidyanathan, K. Gross, Global fan speed control considering non-ideal temperature measurements in enterprise servers, in: 2014 Design, Automation & Test in Europe Conference & Exhibition (DATE), IEEE, 2014, pp. 1–6.

[91] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, C. Hyser, Renewable and cooling aware workload management for sustainable data centers, in: Proceedings of the 12th ACM Sigmetrics/Performance Joint International Conference on Measurement and Modeling of Computer Systems, 2012, pp. 175–186.

[92] A.W. Lewis, N.-F. Tzeng, S. Ghosh, Runtime energy consumption estimation for server workloads based on chaotic time-series approximation, ACM Trans. Archit. Code Optim. (TACO) 9 (3) (2012) 1–26.

[93] H.F. Wang, Total energy consumption model of fan subsystem suitable for continuous commissioning, ASHRAE Trans. 110 (1) (2004) p.1–8.

[94] A. Kusiak, M. Li, F. Tang, Modeling and optimization of HVAC energy consumption, Appl. Energy 87 (10) (2010) 3092–3102.

[95] K.C. Ng, H. Chua, W. Ong, S. Lee, J. Gordon, Diagnostics and optimization of reciprocating chillers: theory and experiment, Appl. Therm. Eng. 17 (3) (1997) 263–276.

[96] J. Saththasivam, K. Choon Ng, Prediction of chiller power consumption: an entropy generation approach, Heat Transfer Eng. 38 (4) (2017) 389–395.

[97] T.-S. Lee, Thermodynamic modeling and experimental validation of screw liquid chillers, ASHRAE Trans. 110 (2004) 206–216.

[98] J.-H. Kim, N.-C. Seong, W. Choi, Modeling and optimizing a chiller system using a machine learning algorithm, Energies 12 (2019) 2860, http://dx.doi.org/10.3390/en12152860.

[99] H.D. Vu, K.S. Chai, B. Keating, N. Tursynbek, B. Xu, K. Yang, X. Yang, Z. Zhang, Data driven chiller plant energy optimization with domain knowledge, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1309–1317.

[100] S. Park, K.U. Ahn, S. Hwang, S. Choi, C.S. Park, Machine learning vs. hybrid machine learning model for optimal operation of a chiller, Sci. Technol. Built Environ. 25 (2) (2019) 209–220.

[101] E. Sala-Cardoso, M. Delgado-Prieto, K. Kampouropoulos, L. Romeral, Predictive chiller operation: A data-driven loading and scheduling approach, Energy Build. 208 (2020) 109639.

[102] C. Xu, K. Jia, Z. Wang, Y. Yuan, A multi-component chiller status prediction method using E-LSTM, in: International Conference on Genetic and Evolutionary Computing, Springer, 2019, pp. 416–428.

[103] J.E. Braun, Methodologies for the Design and Control of Central Cooling Plants (Ph.D. thesis), 1988.

[104] M. Hydeman, K. Jr, A. Dexter, Tools and techniques to calibrate electric chiller component models, ASHRAE Trans. 108 (2002) 733–741.

[105] Y. Wang, X. Jin, Z. Du, X. Zhu, Evaluation of operation performance of a multi-chiller system using a data-based chiller model, Energy Build. 172 (2018) 1–9.

[106] Y.C. Chang, C.Y. Chen, J.T. Lu, J.K. Lee, T.S. Jan, C.L. Chen, Verification of chiller performance promotion and energy saving, Engineering 05 (1) (2013) 141–145.

[107] D. Görges, Relations between model predictive control and reinforcement learning, IFAC-PapersOnLine 50 (1) (2017) 4920–4928.

[108] D.Q. Mayne, Model predictive control: Recent developments and future promise, Automatica 50 (12) (2014) 2967–2986.

[109] T. Koller, F. Berkenkamp, M. Turchetta, A. Krause, Learning-based model predictive control for safe exploration, in: 2018 IEEE Conference on Decision and Control (CDC), 2018, pp. 6059–6066.

[110] N. Lazic, C. Boutilier, T. Lu, E. Wong, B. Roy, M. Ryu, G. Imwalle, Data center cooling using model-predictive control, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018, URL https://proceedings.neurips.cc/paper/2018/file/059fdcd96baeb75112f09fa1dcc740cc-Paper.pdf.

[111] M. Ogawa, H. Endo, H. Fukuda, H. Kodama, T. Sugimoto, T. Horie, T. Maruyama, M. Kondo, Cooling control based on model predictive control using temperature information of IT equipment for modular data center utilizing fresh-air, in: 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), IEEE, 2013, pp. 1815–1820.

[112] M. Ogawa, H. Fukuda, H. Kodama, H. Endo, T. Sugimoto, T. Kasajima, M. Kondo, Development of a cooling control system for data centers utilizing indirect fresh air based on model predictive control, in: 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, 2015, pp. 132–137.

[113] R. Zhou, Z. Wang, C.E. Bash, A. McReynolds, C. Hoover, R. Shih, N. Kumari, R.K. Sharma, A holistic and optimal approach for data center cooling management, in: Proceedings of the 2011 American Control Conference, IEEE, 2011, pp. 1346–1351.

[114] M. Kheradmandi, D.G. Down, H. Moazamigoodarzi, Energy-efficient data-based zonal control of temperature for data centers, in: 2019 Tenth International Green and Sustainable Computing Conference (IGSC), IEEE, 2019, pp. 1–7.

[115] S. Mirhoseininejad, G. Badawy, D.G. Down, A data-driven, multi-setpoint model predictive thermal control system for data centers, J. Netw. Syst. Manage. 29 (1) (2021) 1–22.

[116] Q. Fang, J. Wang, Q. Gong, Qos-driven power management of data centers via model predictive control, IEEE Trans. Autom. Sci. Eng. 13 (4) (2016) 1557–1566.

[117] L. Parolini, B. Sinopoli, B.H. Krogh, Z. Wang, A cyber–physical systems approach to data center modeling and control for energy efficiency, Proc. IEEE 100 (1) (2011) 254–268.

[118] R.S. Sutton, A.G. Barto, Introduction to Reinforcement Learning, Vol. 135, MIT press, Cambridge, 1998.

[119] Z. Zhang, K.P. Lam, Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system, in: Proceedings of the 5th Conference on Systems for Built Environments, 2018, pp. 148–157.

[120] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.

[121] B. Chen, Z. Cai, M. Bergés, Gnu-rl: A precocial reinforcement learning solution for building hvac control using a differentiable mpc policy, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 316–325.

[122] D. Van Le, R. Wang, Y. Liu, R. Tan, Y.-W. Wong, Y. Wen, Deep reinforcement learning for tropical air free-cooled data center control, 2020, arxiv preprint arXiv:2012.06834.

[123] E. Berglund, LQR and MPC control of a simulated data center, 2017.

[124] V. Mnih, A.P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning, 2016, pp. 1928–1937.

[125] M. Jaderberg, V. Mnih, W.M. Czarnecki, T. Schaul, J.Z. Leibo, D. Silver, K. Kavukcuoglu, Reinforcement learning with unsupervised auxiliary tasks, 2016, arXiv preprint arXiv:1611.05397.

[126] H. Chen, M. Kesavan, K. Schwan, A. Gavrilovska, P. Kumar, Y. Joshi, Spatially-aware optimization of energy consumption in consolidated data center systems, in: International Electronic Packaging Technical Conference and Exhibition, Vol. 44625, 2011, pp. 461–470.

[127] T. Hua, J. Wan, Z. Rasheed, L. Li, Z. Ma, Comparison of deep reinforcement learning algorithms in data center cooling management: A case study. URL http://www.zeeshanrasheed.com/wp-content/uploads/2020/09/comparision-of-deep-RL-algorithim-in-data-center.pdf.

[128] M. Nooruzzaman, X. Fernando, Hyperscale data center networks with interconnected transparent island architecture, in: 2020 IEEE Photonics Conference (IPC), IEEE, pp. 1–2.

[129] M. Wahlroos, M. Pärssinen, S. Rinne, S. Syri, J. Manner, Future views on waste heat utilization–Case of data centers in Northern Europe, Renew. Sustain. Energy Rev. 82 (2018) 1749–1764.

[130] K. Ebrahimi, G.F. Jones, A.S. Fleischer, A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities, Renew. Sustain. Energy Rev. 31 (2014) 622–638.

[131] H. Shoukourian, T. Wilde, D. Labrenz, A. Bode, Using machine learning for data center cooling infrastructure efficiency prediction, in: 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017, pp. 954–963, http://dx.doi.org/10.1109/IPDPSW.2017.25.

[132] W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: Vision and challenges, IEEE Internet Things J. 3 (5) (2016) 637–646.

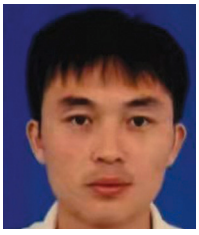
[133] K. Matthews, Six shifts for the future of data centers, https://www.vxchnge.com/blog/the-future-of-data-center-cooling.

[134] K.M. Alam, A. El Saddik, C2PS: A digital twin architecture reference model for the cloud-based cyber-physical systems, IEEE Access 5 (2017) 2050–2062, http://dx.doi.org/10.1109/ACCESS.2017.2657006.

[135] M. Grieves, Digital Twin: Manufacturing Excellence through Virtual Factory Replication, Vol. 1, White Paper, Florida Institute of Technology, 2014, pp. 1–7.

Qingxia Zhang is a Ph.D candidate of school of Computer Science, Fudan University. She received the M.S. degree in School of Computer Science & Technology from Northeastern University, Shenyang, China, in 1997. Her research interests are computer technology, big data analysis and natural language processing.

Zihao Meng received the M.S. degree from the School of Automation Science and Electrical Engineering, Beihang University, China. His research interests include industry intelligence and natural language processing.



Xianwen Hong graduated from Huazhong University of Science and Technology majoring in electrical engineering and automation in 2004. He has focused on data center infrastructure technology and development with more than a decade of experience. He works for the Postal Savings Bank of China, where he is engaged in the construction, operation and maintenance of data centers.

Yuhao Zhan is Deputy Senior Engineer, Hefei Data Center, Postal Savings Bank of China, Hefei, China. He graduated from Nanjing Tech University in 2007. He obtained bachelor's degree with a major in HVAC, He is engaged in refrigeration system construction and maintenance of data center for more than 8 Years.

Jia Liu was awarded the Ph.D. degree from the University of Science and Technology Beijing. She is a post-doctoral and lecturer in University of Science and Technology Beijing. Her research interests include smart city and safety assessment.

Jiabao Dong is a Ph.D. candidate at School of Automation Science and Electrical Engineering, Beihang University, China. His research interests include industry intelligence, robot grasping policy and semantic segmentation.

Tian Bai received the M.S. degree in School of Automation Science and Electrical Engineering, Beihang University. His research interests include deep learning, health data mining and management, service recommendation and service composition.

Junyu Niu is the professor of School of Computer Science and Technology, Fudan University . She received the Ph.D. degree in school of computer science & technology from Northeastern University, Shenyang, China, in 2001. She has engaged in the field of computer technology, big data analysis and NLP.

Dr. M. Jamal Deen is a Distinguished University Professor and Senior Canada Research Chair at McMaster University. His current research interests are nano-/opto-electronics, nanotechnology, data analytics and their emerging applications to health and environmental sciences. His research record includes more than 620 peer-reviewed articles, two textbooks and 6 awarded patents extensively used in industry. He was awarded four honorary doctorate degrees in recognition of his exceptional research and scholarly accomplishments, exemplary professionalism and valued services. He is elected to Fellow status in twelve national academies and professional societies including The Royal Society of Canada, Chinese Academy of Sciences, IEEE, APS and ECS. In 2018, he was appointed to the Order of Canada, the highest civilian honor in Canada. He served as the elected President of the Academy of Science, The Royal Society of Canada, 2015–2017.