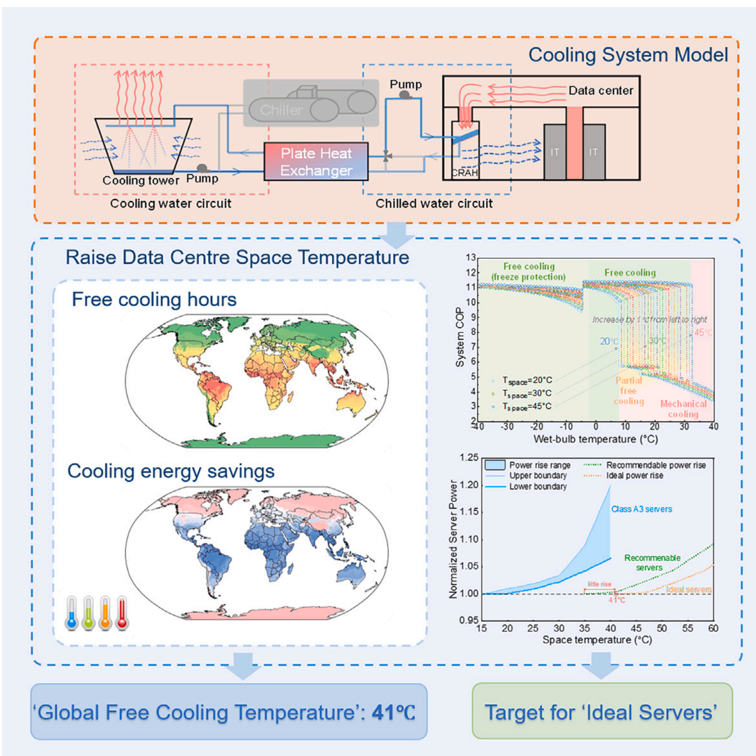
Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

The global energy impact of raising the space temperature for high-temperature data centers



Yingbo Zhang, Hangxin Li, Shengwei Wang

beswwang@polyu.edu.hk

Highlights

The global energy impacts of high-temperature data centers are quantified

The trade-off between cooling-energy savings and server power rise is analyzed

Quantitative guidance and targets are established for developing "ideal" servers

Ever-increasing energy consumption of data centers is a growing global concern. A high-temperature data center is a potential solution to reduce cooling energy by changing the cooling mechanism. Zhang et al. report that by raising the space temperature to 41°C, up to 56% of cooling energy could be saved by fully utilizing ambient air for cooling.

Zhang et al., Cell Reports Physical Science 4, 101624
October 18, 2023
https://doi.org/10.1016/j.xcrp.2023.101624

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

The global energy impact of raising the space temperature for high-temperature data centers

Yingbo Zhang,[1] Hangxin Li,[1] and Shengwei Wang[1,2,3,*]

SUMMARY

Ever-increasing energy consumption of data centers is a growing global concern. To tackle this problem, a high-temperature data center is proposed as a fundamental solution. It adopts a different cooling mechanism and makes "chiller-free" data centers possible, facilitating the transition from chiller-based cooling to completely free cooling in data centers. Here, we report the global energy impacts of adopting high-temperature data centers and critically analyze the trade-off between cooling-energy savings and server power rise using worldwide weather data and server performance data. Quantitative guidance and targets are established for developing "ideal" and "recommendable" servers, considering the server performance associated with the thermal environment. When raising the space temperature to 41°C (namely, the "global free-cooling temperature"), nearly all the land area can achieve 100% free cooling year round globally. Operating at this space temperature, up to 56% cooling-energy savings could be achieved compared with operating at a current typical space temperature of 22°C.

INTRODUCTION

As the digital backbone of our increasingly interconnected world, energy-intensive data centers pose an ever-increasing challenge to global decarbonization.[1] The energy use per square foot in data centers can be 100 times that of typical office buildings.[2] Global data center electricity use in 2021 was 220–320 TWh,[3] around 0.9%–1.3% global electricity demand.[4] It is estimated that global data center electricity use will increase to 848 TWh in 2030.[5] Notably, the cooling energy required to keep the servers in data centers from overheating is on a par with that of the servers themselves, representing 30%–40% of the total energy consumption of data centers.[6,7]

Data centers typically operate at temperatures of 20°C–25°C, following a common notion that "cold is better."[8] Engineers and operators tend to follow conventional practices according to the conservative suggestions of manufacturers and conventional wisdom. However, current servers are much more robust to high temperatures than those previously used[9] as it has been recognized by the ASHRAE (the American Society of Heating, Refrigeration and Air-conditioning Engineers) standard,[8] leading to rising interest in expanding their operating temperature ranges. In the past, some legacy data centers operated at a space temperature of 13°C,[10] forcing technicians to wear parkas while working in those spaces. In 2004, ASHRAE published the first recommended environmental temperature range for IT equipment of 20°C–25°C,[11]

[1]Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

[2]Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

[3]Lead contact

*Correspondence: beswwang@polyu.edu.hk

https://doi.org/10.1016/j.xcrp.2023.101624

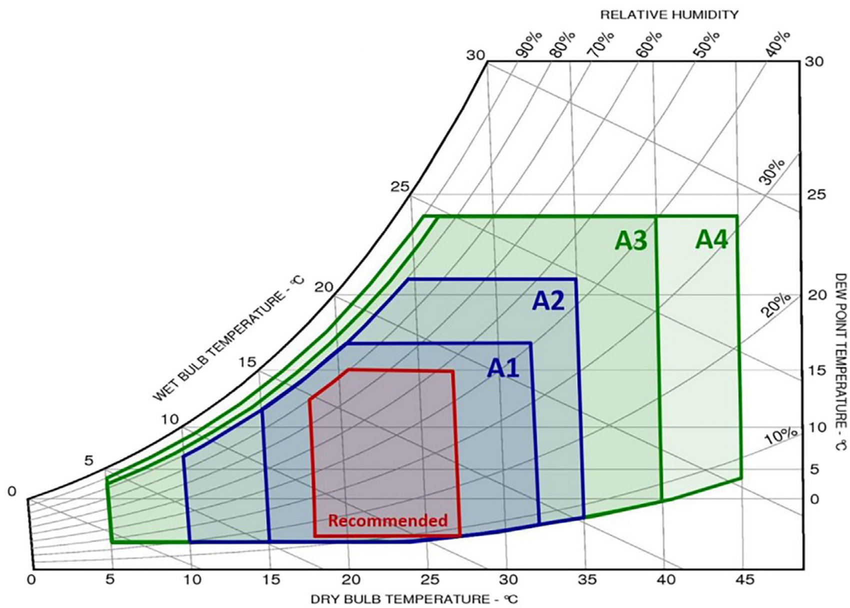


Figure 1. Environmental temperature and humidity ranges for IT equipment specified by ASHRAE

The recommended region is based on an analysis of the total cost of ownership in 1996.[2] The green region is the newly added environmental temperature and humidity range of IT equipment in 2011.[13] The upper-temperature limits of classes A3 and A4 are 40°C and 45°C, respectively.[8]

which is considered an authoritative standard. In 2008, ASHRAE expanded the recommended environmental temperature range to 18°C–27°C.[12] In 2011, the range of allowable environmental temperatures was further expanded and two new classes (A3 and A4) were adopted based on updated critical information (IT design and failure data) provided by IT manufacturers,[13] as shown in Figure 1. The class A3 environment has an upper-temperature bound of 40°C, while the class A4 environment has an even higher upper-temperature bound of 45°C. In 2016, a comprehensive discussion of the workflow for raising the space temperature was published.[14,15]

Some operators mistakenly consider the use of ASHRAE-recommended temperature ranges to be mandatory. However, ASHRAE itself emphasizes that "although the recommended ranges are highlighted as a separate red region, it was never intended that the recommended temperature range would be the absolute limits of inlet air temperature for IT equipment."[13] Notably, this recommended region is based on an analysis concerning the total cost of ownership in 1996.[2] With the technological evolution of IT equipment, increasingly advanced products that use superior materials have come to the market,[16] with next-generation functional components and improved chip-cooling technologies. Most servers installed around 2010 could fall into what ASHRAE categorizes as class A1 or A2 servers, whereas, some servers today are designed to operate at a maximum ambient temperature of 40°C or 45°C, corresponding to class A3 and class A4 of the ASHRAE thermal guideline[17] (see the examples under Table S3). Class A4 servers are expected to eventually become widely used in data centers. Thus, with servers withstanding higher-temperature operation and the ground-breaking extension of their environmental temperature range, owners and operators are being given the confidence to build and operate data centers at significantly higher temperatures than those used today.

Data center cooling technology has progressed significantly over time.[18] Improvements in air-handling design, particularly for data centers, were the first major steps for increasing cooling efficiency.[19] The proposal of "aisle containment" has changed the layout of data centers, greatly improving cooling efficiency by reducing the mixing of the hot exhaust air from a server with the cold intake air from the central cooling supply.[20] A similarly improved coordination control strategy between the server and the cooling sides can achieve a better match of cooling supply and demand.[21] Furthermore, recently developed artificial intelligence (AI)-based real-time transient temperature predictions of server CPUs and the cold chamber help to better manage airflow in data centers.[22] Although these technologies can improve cooling efficiency, the resulting energy savings are limited, as they only improve the efficiency of the "cold distribution." In addition, some of the new technologies proposed to tackle high-cooling-energy consumption problems are considerably more efficient, such as liquid cooling.[5] Unfortunately, these options are often hampered by the conservative minds of designers and operators and by technological and economic challenges, such as maintenance and investment.[8]

Developments in cooling technology have improved the efficiency of data center cooling systems largely. At the same time, raising the space temperature in data centers is another effective means to improve cooling system efficiency by reducing the temperature differential in the vapor compression cycle. Furthermore, raising the space temperature to a certain level could eliminate thoroughly the use of chillers and could maximize "free-cooling" hours. Free cooling is the process of using external ambient air for cooling, which is commonly adopted today in air-cooled data centers and is considered the single greatest opportunity for cooling-energy savings.[16] High-temperature data centers change the cooling mechanism fundamentally, enabling and promoting the transition from chiller-based cooling to completely free cooling—this could become a breakthrough for the data center industry.

Several previous studies[23–25] have considered the impact of the raising space temperature by several degrees on data center cooling energy. However, these studies only addressed this impact in some specific scenarios. In this study, we attempt to quantify the global cooling-energy savings of raising the space temperature in a wide range by selecting 3 representative cities in each climate zone, totaling 57 cities in 19 climate zones[26] (see Table S4) and using their weather data from typical years. We also analyze the impact of raising the space temperature on the free-cooling hours in data centers worldwide. Moreover, a critical analysis of the mutual effects and trade-off between server power rise and cooling-energy savings is performed when raising space temperature. The further development target of server energy performance associated with the working thermal environment is established. To conduct our quantitative evaluation accurately and reliably, we develop a basic cooling system model based on fundamental mathematical formulas and the test data of cooling equipment from manufacturers. The results clarify the global energy impact of high-temperature data centers, providing quantitative guidance for IT and cooling professionals to make informed decisions on the future development of the data center industry.

RESULTS

Performance of data center cooling systems under different ambient air and space temperatures

A basic cooling system model is developed based on fundamental principles and mathematical formulas combined with the test data of cooling equipment

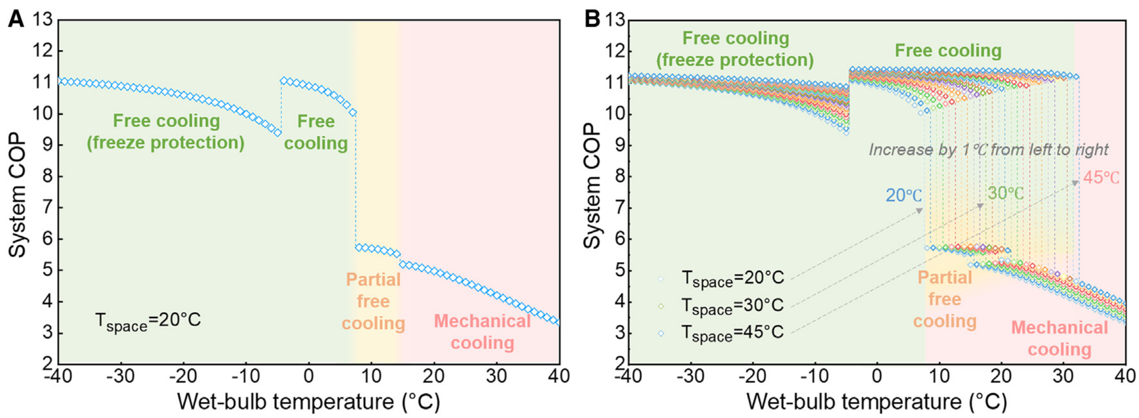


Figure 2. Cooling system COP versus wet-bulb temperature
(A) At a space temperature of 20°C.
(B) At space temperatures ranging from 20°C to 45°C.

performance. The cooling system comprises cooling distribution equipment, cooling equipment, and heat rejection equipment. In general, the three typical operation modes of the cooling system are the mechanical-cooling mode, the partial-free-cooling mode, and the free-cooling mode (see Figures S2–S4). Using the weather data for typical years, we simulated the cooling system under the three operation modes corresponding to actual outdoor conditions and compared their resulting coefficients of performance (COPs). The mode that satisfies the cooling load and consumes the least cooling energy was selected for every outdoor condition.

Figure 2 shows the cooling system COP at wet-bulb temperatures ranging from −40°C to 40°C with space temperatures between 20°C and 45°C. Notably, the COP of the free-cooling mode that uses dry coolers for freeze protection at very low ambient temperature is lower than the COP of the free-cooling mode that uses open cooling towers. This can be attributed to the heat rejection efficiency of dry coolers being lower than that of open cooling towers.[27] The open cooling towers expose water directly to the cold atmosphere, thereby transferring the heat directly to the air via sensible and evaporative cooling. By comparison, a dry cooler involves indirect contact between the heated fluid and ambient air,[28] with this heat transfer mechanism providing an inferior cooling efficiency.[27]

The cooling system COP is then dramatically reduced once the chiller is operating. In both partial-free-cooling and mechanical-cooling modes (which require chillers), the cooling system COP is around 3–5, whereas in the free-cooling mode, the COP is around 11 (without chillers operating). These results demonstrate a significant decrease in the cooling system COP when switching from free cooling to chiller-based cooling. Additionally, the cooling system COP in partial free cooling only has a slight improvement compared with the mechanical-cooling mode. This is primarily attributed to the adopted control strategies and the cooling system concerned. The chillers have certain minimum condensation temperature limits, which constrain the outlet temperature of the cooling water from the cooling towers, as illustrated by Equation 9. The limitation hinders the complete utilization of free cooling in the partial-free-cooling mode.

Figure 2A shows that for a given space temperature of 20°C, the cooling system COP generally decreases as the ambient air wet-bulb temperature increases, except

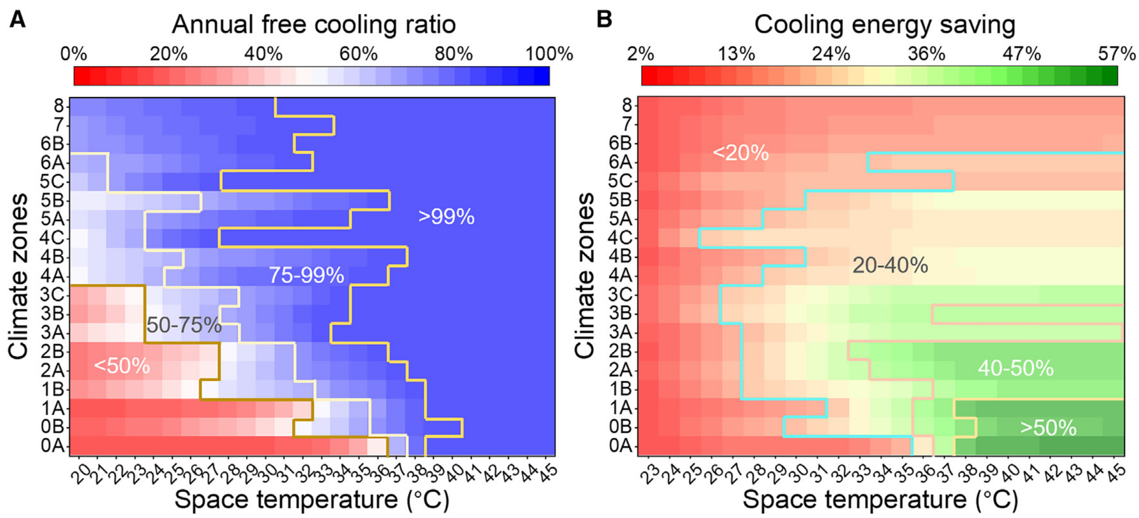


Figure 3. Annual free-cooling ratio and cooling-energy savings in 19 climate zones

(A) Annual free-cooling ratio when raising data center space temperatures from 20°C to 45°C in 19 climate zones.

(B) Annual cooling-energy savings compared with the baseline temperature (22°C) in 19 climate zones.

when freeze protection is switched off during free-cooling modes. In free-cooling modes, the higher ambient air wet-bulb temperature requires cooling tower fans to operate at higher speeds to maintain the chilled water supply temperature at the outlet of the heat exchanger at a certain setpoint, resulting in higher fan energy use. Additionally, in the partial-free-cooling and mechanical-cooling modes, higher ambient air wet-bulb temperatures will require the chillers to operate at higher condensing temperatures, resulting in lower COPs for the chillers.

As the space temperature increases, data centers can achieve higher cooling system COPs, as shown in Figure 2B. This is primarily because data centers can operate in free-cooling mode at higher space temperatures, and thus a high-cooling-system COP is achieved over an extended range of ambient air wet-bulb temperatures. Secondarily, higher space temperatures allow higher chilled water supply temperatures, and thus higher chiller COPs for given air-side system designs or temperature differentials in partial-free-cooling and mechanical-cooling modes. In free-cooling mode, higher space temperatures allow the cooling tower fans to operate at lower speeds, and therefore they consume less energy to maintain the same chilled water supply temperature at the outlet of the heat exchanger. Clearly, as the space temperature increases from 20°C to 30°C and then 45°C, the free-cooling range is extended by 10 K (i.e., wet-bulb temperature from 7°C to 17°C) and then 25 K (i.e., from 7°C to 32°C), respectively. The increase in cooling-system COP is in the range 0.8%–1.7% for every 1 K rise in data center space temperature. In other words, PUE (power usage effectiveness) can decrease in the range of 0.009–0.027 for a 1 K rise in data center space temperature depending on the climate condition (assume the baseline PUE of 1.55, the average in 2022[29]).

Impact of higher space temperatures on free-cooling hours in data centers worldwide and "global free-cooling temperature"

As the space temperature increases, more free-cooling hours become possible. Using the weather data[30] of 57 representative cities from 19 climate zones[26] worldwide, we assessed the potential free-cooling hours at different space temperatures for these cities. Figure 3A shows the mean annual free-cooling ratio (i.e., the

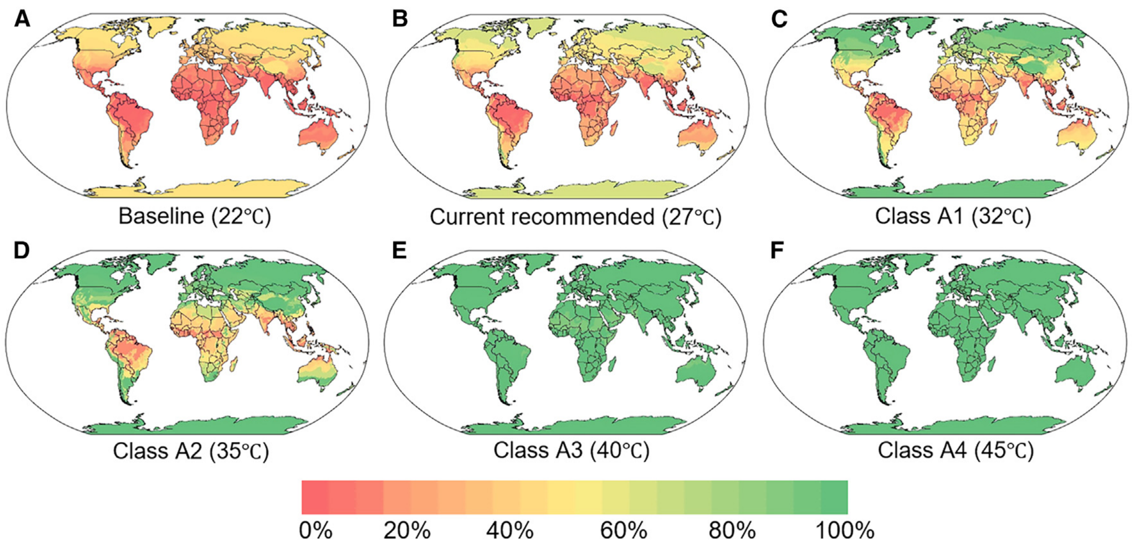


Figure 4. Global maps of annual free-cooling ratio at different space temperatures

(A) At a baseline space temperature of 22°C.
(B) At 27°C (upper limit of current recommendation).
(C) At 32°C (upper limit of class A1).
(D) At 35°C (upper limit of class A2).
(E) At 40°C (upper limit of class A3).
(F) At 45°C (upper limit of class A4).

percentage of free-cooling hours over a year) in 19 climate zones as the space temperature increases from 20°C to 45°C. Here, the climate zones listed on the y axis vary gradually from extremely hot (zone 0) to extremely cold (zone 8). At a space temperature of 20°C, data centers in cold climate zones show high annual free-cooling ratios, but those in hot climate zones show much lower ratios. For instance, the annual free-cooling ratio is 58% in the cold-humid zone 5A. By contrast, nearly no free cooling is possible in hot climate zones, such as 0% in the extremely hot-humid zone 0A. However, if the space temperature is raised to 35°C, an over 99% annual free-cooling ratio can be achieved in zone 5A, while it can be achieved even in zone 0A if the space temperature is further raised to 38°C. Importantly, all climate zones can achieve nearly 100% free cooling year round if the space temperature is raised to 41°C.

Therefore, a new concept, the global free-cooling temperature, is defined as the data center space temperature at or above which the central cooling system can work in free-cooling mode year round in almost all climate zones worldwide. Similarly, the free-cooling temperature of a particular city is defined as the data center space temperature at or above which the central cooling system can work in free-cooling mode year round in the city. According to the above analysis, the value of the global free-cooling temperature is 41°C.

Figure 4 shows the global map of the annual free-cooling ratio of data centers operating at different space temperatures. Regions closer to the equator have lower annual free-cooling ratios than other regions, which is why data centers are often recommended to be built in cold regions. To analyze this further, we consider the percentages of global land area (1.391 billion km^2) that would allow 50%, 75%, or 100% annual free cooling for a given space temperature: for a typical space temperature of 22°C (see Figure 4A), these percentages are 58.7%, 49.9%, and 0%, respectively; for

CellPress
OPEN ACCESS

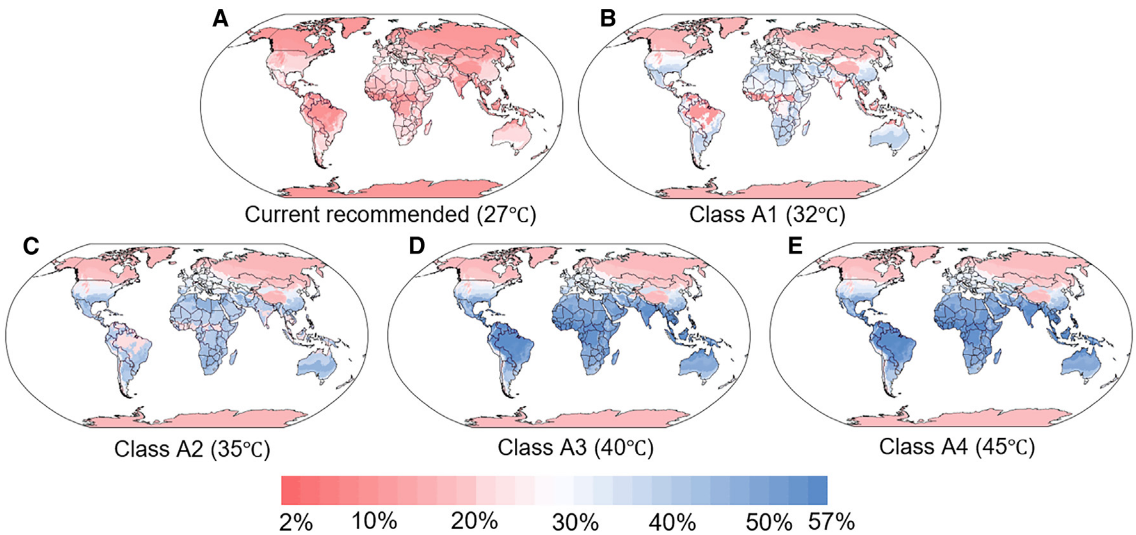


Figure 5. Global maps of cooling-energy savings with reference to the baseline space temperature (22°C)
(A) At 27°C (upper limit of current recommendation).
(B) At 32°C (upper limit of class A1).
(C) At 35°C (upper limit of class A2).
(D) At 40°C (upper limit of class A3).
(E) At 45°C (upper limit of class A4).

a 27°C space temperature (see Figure 4B), they become 69.7%, 58.7%, and 0%, respectively; for a 32°C space temperature (see Figure 4C), they become 87.0%, 76.5%, and 36.5%, respectively; for a 35°C space temperature (see Figure 4D), they become 93.9%, 80.8%, and 60.2%, respectively; for a 40°C space temperature (see Figure 4E), they become 99%, 99%, and 93.7%, respectively; and for a 45°C space temperature (see Figure 4F), they become 99%, 99%, and 99%, respectively. Clearly, increasing the space temperature can dramatically increase the area of land with higher annual free-cooling ratios.

Impact of raising space temperature on data center cooling-energy savings worldwide

We further quantify the global cooling-energy savings achievable by raising data center space temperature worldwide, based on the estimated cooling-system COP at different wet-bulb temperatures (see Figure 2). Figure 3B shows the worldwide mean annual cooling-energy savings when the space temperature is raised from 22°C (the baseline space temperature) to up to 45°C in 19 climate zones. For instance, 13%–56% cooling energy could be saved if the space temperature was raised to 41°C in commonly used air-cooled data centers. A maximum cooling-energy savings of up to 57% could be achieved in zone 0A if the space temperature was raised to 45°C. In addition, every degree (K) of space temperature increase could result in 2%–6% cooling-energy savings, depending on the climate conditions.

Figure 5 shows the global map of cooling-energy savings when raising the space temperature from the baseline 22°C–27°C, 32°C, 35°C, 40°C, and 45°C, respectively. Clearly, when the space temperature is raised, the regions near the equator (with hot weather throughout the year) show higher potential cooling-energy savings than the cold regions, as the hot regions have no free cooling at the baseline space temperature and therefore have high energy-consumption baselines. To analyze this further, we consider the percentages of global land area that could allow 20%, 40%, and 50% annual cooling-energy savings for a given space temperature: for a 27°C space temperature (see Figure 5A), these percentages are 2.9%, 0%, and 0%, respectively; for a 32°C space temperature (see Figure 5B), they become 44.0%, 0%, and 0%, respectively; for a 35°C space temperature (see Figure 5C), they become 47.4%, 11.1%, and 0%, respectively; for a 40°C space temperature (see Figure 5D), they become 53.7%, 36.9%, and 19.1%, respectively; and for a 45°C space temperature (see Figure 5E), they become 53.7%, 36.9%, and 19.1%, respectively. Clearly, increasing the space temperature can significantly increase the potential cooling-energy savings of data centers worldwide. It is worth noticing that, due to the global temperature increase, some critical temperatures, such as the global free-cooling temperature, would need to be raised accordingly.

Impact of raising space temperature on data center energy performance and "ideal servers"

The question remains why most owners have yet to adopt higher space temperatures in their data centers. Critics of high-temperature data centers might argue that raising the space temperature is accompanied by an increase in server power and a decrease in server reliability. The long-term high-temperature operation might be accompanied by accelerating component degradation,[31] shortening lifespan, increasing failure rate,[32] and, accordingly, maintenance and downtime.[33] However, the reliability of servers under high-temperature operation would and should be largely enhanced with the improvement of the thermal design of servers,[34] the development of chip-cooling technologies,[35] and advanced material properties. This is therefore an important issue that needs to be investigated considering the future development of servers.

Current servers can be categorized into four (thermal) environment classes: classes A1, A2, A3, and A4. Servers of the latest generations (i.e., classes A3 and A4), which support wider environmental temperature ranges, are still uncommon in the market and are usually associated with higher prices because of their enhanced heat removal mechanisms and more robust components.[13] Notably, class A3 and A4 servers can work at higher space temperatures and demonstrate less server power rise when working at the same high space temperatures (server inlet temperature) as class A1 or A2 servers.[8] Generally, the main contribution to server power rise comes from the server fan, as the fan runs faster to prevent the server components from overheating; a relatively minor part of the power rise is due to server components catering to the same computing load at higher temperatures.[10] When only considering the server performance associated with the thermal environment and data center cooling energy, it is optimal for servers to operate at the highest space temperatures possible. However, manufacturing servers that can sustain higher temperatures means greater technical challenges and higher costs. Then, what is the expected server performance associated with the thermal environments of ideal or preferred servers? In this section, we analyze the impacts of raising space temperature on the data center energy performance, considering the server power rise of current and future servers.

Figures 6A–6C show the normalized total power and the annual free-cooling ratio when raising the space temperature in three cities from three representative climates, i.e., Beijing, Kunming, and Hong Kong. Figure 6D shows the power rise range of class A3 servers[8] and the power rise trend of ideal servers (dotted lines; note, the power rise of current class A3 servers at a given space temperature is in a range, depending on the server models and manufacturers[8]). The normalized total power of adopting class A3 servers shown in Figures 6A–6C corresponds to the lower boundary of the class A3 server power rise in Figure 6D.

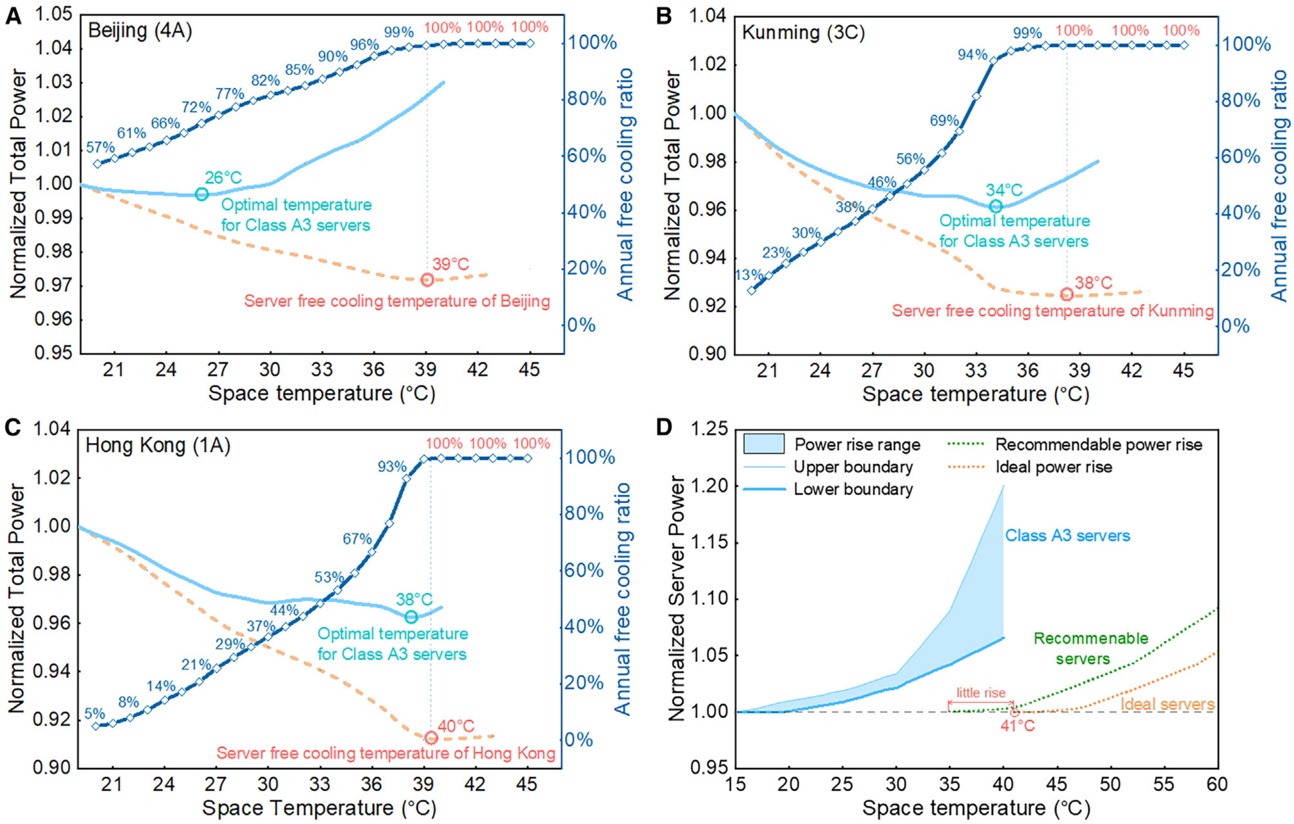


Figure 6. Impact of raising space temperature on the annual free-cooling ratio and normalized server power and total power in data centers in different cities

Normalized total power is calculated against the baseline power at 19°C. Normalized server power is calculated against the baseline power at 15°C.[8]

(A–C) Normalized total power in (A) Beijing, (B) Kunming, and (C) Hong Kong.

(D) Normalized power rise range of class A3 servers and the estimated power rise of ideal servers (mote: we have chosen the lower boundary of the class A3 server power curve for the energy performance assessment, considering the future trend of technology development).

Thus, the optimal space temperature (that corresponds to the minimum total power) is observed to differ for each city. The optimal space temperatures for class A3 servers in Beijing, Kunming, and Hong Kong are 26°C, 34°C, and 38°C, respectively. This is caused by the different annual wet-bulb temperature distributions in each city. In addition, cities in cold climates usually show lower optimal space temperatures than cities in hot climates. For example, at the space temperature of 20°C, the annual free-cooling percentages in Beijing and Hong Kong are 57% and 5%, respectively. When raising the space temperature, Beijing approaches its optimal space temperature at 26°C, whereas Hong Kong approaches its optimal space temperature at 38°C. This is because Beijing has a higher free-cooling baseline (57%) and less room for improvement. Therefore, the cooling-energy saving due to increased free-cooling hours cannot offset the server power rise in Beijing when the space temperature is over 27°C.

The dashed line in Figures 6A–6C shows the normalized total power for ideal servers. It is expected that the servers should work reliably in each city, without the server power rise, when the temperature is not higher than the city's free-cooling temperature, i.e., the temperature corresponding to 100% free-cooling year round in the city. Adopting these servers, data centers could operate in free-cooling mode almost year round in any city. According to Figures 6A–6C, the free-cooling temperatures of Beijing, Kunming, and Hong Kong are 39°C, 38°C, and 40°C, respectively.

To facilitate free cooling of data centers worldwide, an ideal server is defined as a perfect server from the perspective of cooling-energy performance. This server is ideal, as its working temperature is high enough to allow 100% free cooling globally when located in any land region worldwide. The working temperature of this ideal server is also the critical temperature for improving server design with the lowest effort/cost to achieve 100% free cooling worldwide. Therefore, ideal servers should be able to work reliably without increasing the server power when the space temperature is raised to the global free-cooling temperature (i.e., 41°C). This means that the power of ideal servers should not increase with temperature when the space temperature is raised to 41°C (see Figure 6D). As the power of servers often only increases slightly over the first few degrees of temperature rise before increasing significantly, servers that can work reliably without significant server power increase when raising space temperature to the global free-cooling temperature would be highly recommendable. We therefore name these servers "recommendable servers."

DISCUSSION

This study presents a pioneer comprehensive worldwide investigation on the cooling-energy savings of high-temperature data centers and critically analyzes the trade-off between cooling-energy savings and server power rise. The results show that every 1 K increase in the space temperature could enhance the cooling-system COP by 0.8%–1.7% and reduce cooling-energy consumption by 2%–6%, depending on the climate conditions. In other words, PUE can decrease in the range of 0.009–0.027 for a 1 K rise in data center space temperature. If the space temperature in data centers was raised to 41°C, almost all land regions across the world could achieve nearly 100% free cooling throughout the year, and the PUE could be reduced by up to 0.252. Meanwhile, 13%–56% cooling energy could be saved compared with the baseline space temperature setting of 22°C, with practically no additional cooling-energy savings for further raising of the space temperature.

Currently, the space temperature settings in most data centers remain conservative, at typically 20°C–25°C, while ASHRAE recommends a temperature range of 18°C–27°C. However, we found that the optimal space temperature in each city depends on the type of servers. For example, the optimal space temperatures for class A3 servers in Beijing, Kunming, and Hong Kong are 26°C, 34°C, and 38°C, respectively. Therefore, it is important to consider the actual climate conditions in a particular city and the server performance associated with the thermal environment of the chosen servers (refer to the specifications of servers) to determine the optimal space temperature settings.

As a basic recommendation and target for server development associated with the thermal environment, we found that a global free-cooling temperature of 41°C is the minimum space temperature that would allow all climate zones to achieve nearly 100% free cooling year round. Considering the server performance associated with their thermal environment, ideal servers should be able to work reliably without server power rise as the space temperature increases up to 41°C. Considering the manufacturing challenges and costs, recommendable servers should work reliably without significant server power rise for space temperatures up to 41°C.

The comfortable temperature range of the human body is narrow and determined by nature, whereas servers could be designed to work in wider temperature ranges, such as by using enhanced printed circuit board materials[13] and third-generation semiconductors.[36] The main concerns regarding the high-temperature operation of data centers are server reliability and data-processing performance. Therefore, the key to implementing high-temperature data centers is to develop and widely deploy servers and IT equipment that enable high-temperature operation. Servers designed for class A3 and A4 environments, with upper-temperature limits of 40°C and 45°C, respectively, are equipped with optimized heat removal mechanisms and more robust components. Unfortunately, the manufacturing costs for these servers can greatly exceed those of conventional servers.[13] From the perspective of practical implementation, the cost premiums of class A3 and A4 servers, or later-generation servers, need more careful evaluation before their wide deployment. This evaluation must consider the cooling-energy savings from improved chiller energy efficiency and the increased free-cooling hours, which lower operating expenses, and compare these against the initial investment costs of adopting next-generation servers while potentially allowing chiller-free cooling.

The high-temperature data center is a major development direction for next-generation data centers in addition to the development of servers themselves. High-temperature data centers facilitate and promote the transition from chiller-based to chiller-free cooling, offering a tremendous opportunity for the data center industry to maximize cooling-energy savings. This approach could break new ground, leading to a cliff-like drop in the cooling-energy consumption of data centers. From the perspective of cooling, the optimization of cooling systems in data centers at room, rack, and server levels is essential for high-temperature operation. From the perspective of IT equipment, robustness in higher-temperature working environments and innovative chip-cooling technology are keys to implementing high-temperature data centers. The development of third-generation semiconductors allows chips to operate at higher temperatures fundamentally. The improvement of chip-level cooling technologies could increase effectively heat dissipation of heat sources to enhance the high-temperature tolerance of servers. These advancements could eliminate the concerns about high-temperature server operation. Our findings provide quantitative guidance for IT and server professionals to further develop IT equipment and servers that take the data center cooling energy into account. The results should also be valuable for other decision-makers and stakeholders in developing standards and guidelines for next-generation data centers.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information questions should be directed to the lead contact, Shengwei Wang (beswwang@polyu.edu.hk).

Materials availability

No materials were used in this study.

Data and code availability

All data are made available with sources specified in the supplemental information.

Modeling approach

In this study, a commonly used cooling system is selected. Its cooling load is assumed as 4,000 kW. The specifications of corresponding components are selected according to the cooling load (Table S1). The cooling system consists of chillers, heat exchangers, cooling towers, chilled water pumps, cooling water pumps, and computer room air handlers. The cooling system can be divided into two subsystems, the cooling water loop and the chilled water loop (Figure S1). In addition, a parallel configuration is used on the cooling water loop, and a serial configuration is used on the chilled water loop according to the common engineering practice[37,38] and comparative assessment.[39] In this cooling system, the approach temperature between the chilled water supply temperature and the supply air temperature is set at 7°C.[2] We also assume that the server inlet temperature equals the supply air temperature, that the server inlet temperature is even, and that there are no hot-spot problems.

The cooling system model

Chiller models and the assumption for heat exchangers

Braun's method is used to model chillers,[40] in which two variables are involved, i.e., the load and the temperature difference between the leaving condenser and chilled water flows. Equation 1 is the correlation of the chiller power and variables. In addition, the testing data of the chiller used in this study, which are for high-temperature applications, are from a major manufacturer (Trane), allowing a chilled water supply temperature up to 20°C.

$$\frac{P_{ch}}{P_{des}} = a_0 + a_1 X + a_2 X^2 + a_3 Y + a_4 Y^2 + a_5 XY, \quad \text{(Equation 1)}$$

$$X = \frac{\dot{Q}_e}{\dot{Q}_{des}}, \text{and} \quad \text{(Equation 2)}$$

$$Y = \frac{(T_{cwr} - T_{chws})}{\Delta T_{des}}, \quad \text{(Equation 3)}$$

where X is the ratio of the chiller load to the design load, Y is the leaving water temperature difference divided by a design value, P_{ch} is the chiller power consumption, and P_{des} is the power consumption at the design condition. The empirical coefficients in Equation 1 (a_0, a_1, a_2, a_3, a_4, and a_5) are determined using linear least-squares curve fitting based on the chiller performance data from the manufacturer. Detailed empirical coefficients and parameters of the chiller model can be found in Table S2. An example of a chiller COP at the temperature of leaving condenser water of 28°C is shown in Figure S5. In addition, a 1.5°C temperature approach for heat exchangers is assumed.[39,41]

Cooling tower models

In this study, heat rejection equipment includes open cooling towers and dry coolers.[28] The cooling water loop would switch to dry coolers when the ambient wet-bulb temperature is lower than −5°C, according to engineering practice, to avoid freeze problems.

Open cooling towers involve sensible and latent heat transfer. In this study, the ε-NTU method is used to model open cooling towers.[42] Braun[40] stated that air effectiveness can be determined using the relationships for sensible heat exchangers with modified definitions for the number of transfer units and the capacitance rate ratios, using the assumption that the Lewis number equals 1. For a counterflow cooling tower, it is described as Equation 4.

$$\varepsilon_a = \frac{1 - \exp(-NTU(1 - m^*))}{1 - m^* \exp(-NTU(1 - m^*))}, \quad \text{(Equation 4)}$$

where

$$\mathrm{NTU} = \frac{h_D A_v V_{cell}}{m_a}, \text{ and} \quad \text{(Equation 5)}$$

$$m^* = \frac{m_a C_s}{m_{w,i} C_{pw}}. \quad \text{(Equation 6)}$$

The saturation-specific heat, C_s, is defined as the average slope of the saturation enthalpy with respect to the temperature curve. It can be determined with the water inlet and outlet conditions and psychometric data using Equation 7.

$$C_s = \frac{h_{s,w,i} - h_{s,w,o}}{T_{w,i} - T_{w,o}} \quad \text{(Equation 7)}$$

The wet-bulb temperature is used as a primary input of the open cooling tower model because the enthalpy can be approximated as a formula related to wet-bulb temperature,[43] at a given atmospheric pressure. For dry coolers, the ε-NTU method is also used to obtain the performance of the dry cooler at different conditions. Dry-bulb temperature is one of the main input parameters for the dry cooler model. Detailed mathematical references of these two models can be found in TRNSYS 18 component mathematical manual.[42]

According to the desired cooling capacity for cooling towers and outdoor conditions, the airflow rate can be determined. The fan power is in cubic growth of the rotational speed ideally (k = 3), as shown in Equation 8.[44] In this study, k is selected as 1.5 based on practical *in situ* operation data. Cooling towers cannot achieve the ideal performance (k = 3) in practical operations.[44]

$$\frac{W_{ct}}{W_{ct,design}} = \left(\frac{Q_{ct}}{Q_{ct,design}}\right)^k, \quad \text{(Equation 8)}$$

where W_{ct} is the energy consumption of cooling towers, $W_{ct,design}$ is the energy consumption of cooling towers at the design condition, Q_{ct} is the airflow rate, and $Q_{ct,design}$ is the airflow rate at the design condition.

In the free-cooling mode, the speed of the cooling tower fans is controlled to make the outlet water temperature of the heat exchangers reach the desired chilled water supply temperatures. At both partial-free-cooling mode and mechanical-cooling mode, the chilled water supply temperature is controlled by the chiller itself, and the speed of cooling tower fans is modulated to maintain the cooling tower water outlet temperature at a setpoint given by Equation 9, where $T_{ct,out}$ is the outlet cooling water temperature, T_{wet} is the wet-bulb temperature, and $T_{min,ct}$ is the minimum condenser water entering the temperature setpoint. The cooling tower control setting given by Equation 9 is used as a proximate or near-optimal setting.[45] Cooling tower water temperature cannot drop below the chiller minimum condenser water entering temperature setpoint, as defined by the chiller manufacturer. In addition, the cooling water return temperature is limited to 45°C to avoid calcium salt precipitation.[46]

$$T_{ct,out} = \max\left(T_{wet} + 5^\circ C, T_{min,ct}\right) \quad \text{(Equation 9)}$$

Pump models

The cooling water pumps are usually constant-speed pumps, and they are assumed to work at rated power. The chilled water pumps are variable-speed pumps. The energy consumption of chilled water pumps depends on the pressure drop, the water flow rate, and the pump efficiency, as shown in Equation 10.

$$W_{cwp} = \frac{\Delta p_{cwp} \times m_w}{\eta_{cwp}} \quad \text{(Equation 10)}$$

The pressure head of pumps (equal to the pressure drop of the chilled water loop) is set to be linear to the water flow rate in operation[44] (see Figure S6). This control strategy is practically not far from optimum for the energy-efficient control of the chilled water loop.

Computer room air handlers

The difference between supply and return air temperatures of the computer rooms is assumed to be a constant of 10 K.[2]

Limitations

The energy assessment models used in this study simulate the water-cooled chiller cooling systems, the most widely used central chiller plants for data centers today.[8,41,47] The conclusions of this study are therefore limited to the implementation of air-cooled data centers using water-cooled chiller plants. However, as the energy impacts of operating temperatures in air-cooled data centers using air-cooled chiller plants are similar, the quantitative conclusions are still valid, and the quantitative recommendations may deviate slightly. There are other types of cooling systems, such as liquid-cooled data centers, which have attracted great attention. The conclusions and quantitative recommendations of this study would not apply to the liquid-cooled data centers, while the recommendation of rising the operation temperature of server components is still beneficial to their operation. In fact, when the raised server operating temperature allows wide implementation of air-cooled data centers worldwide, the adoption of liquid-cooled data centers might not be as attractive and beneficial due to their technological and economic challenges, such as maintenance.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2023.101624.

ACKNOWLEDGMENTS

The research of this paper is financially supported by a collaborative research fund (C5018-20G) and a grant under the Hong Kong PhD Fellowship Scheme of the Research Grants Council in the Hong Kong SAR.

AUTHOR CONTRIBUTIONS

Y.Z. developed the cooling system model and wrote the draft. H.L. reviewed the draft and offered advice. S.W. revised and refined the paper, acting as the main supervisor.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 31, 2023
Revised: July 25, 2023
Accepted: September 14, 2023
Published: October 18, 2023

REFERENCES

1. Zheng, J., Chien, A.A., and Suh, S. (2020). Mitigating Curtailment and Carbon Emissions through Load Migration between Data Centers. Joule *4*, 2208–2222. https://doi.org/10.1016/j.joule.2020.08.001.
2. ASHRAE (2009). Best Practices for Datacom Facility Energy Efficiency, Second Edition.
3. Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science *367*, 984–986.
4. https://www.iea.org/reports/data-centres-and-data-transmission-networks.
5. Mytton, D., and Ashtine, M. (2022). Sources of data center energy estimates: A comprehensive review. Joule *6*, 2032–2056. https://doi.org/10.1016/j.joule.2022.07.011.
6. Mitchell-Jackson, J., Koomey, J.G., Nordman, B., and Blazek, M. (2003). Data center power requirements: measurements from Silicon Valley. Energy *28*, 837–850. https://doi.org/10.1016/S0360-5442(03)00009-4.
7. Luo, Y., Andresen, J., Clarke, H., Rajendra, M., Maroto-Valer, M., Fan, L.R., Zhao, Y.T., Han, Z.G., He, W.Y., Meng, G., et al. (2019). A decision support system for waste heat recovery and energy efficiency improvement in data centres. Appl. Energy *250*, 1217–1224. https://doi.org/10.1016/j.apenergy.2019.05.029.
8. ASHRAE (2021). Thermal Guidelines for Data Processing Environments, Fifth Edition.
9. Seaton, I. (2017). Airflow Management Considerations for a New Data Center – Part 2: Server Performance versus Inlet Temperature. https://www.upsite.com/blog/server-performance-versus-inlet-temperature/.
10. Zhang, Y., Shan, K., Li, X., Li, H., and Wang, S. (2023). Research and Technologies for next-generation high-temperature data centers – State-of-the-arts and future perspectives. Renew. Sustain. Energy Rev. *171*, 112991. https://doi.org/10.1016/j.rser.2022.112991.
11. ASHRAE (2004). Thermal Guidelines for Data Processing Environments, First Edition.
12. ASHRAE (2008). Thermal Guidelines for Data Processing Environments, Second Edition.
13. ASHRAE (2011). Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance.
14. Beaty, D.L., and Quirk, D. (2016). De-risking data center temperature increases, Part 1. ASHRAE J. *58*, 74–82.
15. Beaty, D.L., and Quirk, D. (2016). De-risking data center temperature increases, Part 2. ASHRAE J. *58*, 70–75.
16. Fitch, J. (2012). Dell's Next Generation Servers: Pushing the Limits of Data Center Cooling Cost Savings (DELL Enterprise Reliability Engineering). https://studylib.net/doc/18559703/pushing-the-limits-of-data-center-cooling-cost-savings.
17. Seaton, I. (2017). Airflow Management Considerations for a New Data Center – Part 1: Server Power versus Inlet Temperature. https://www.upsite.com/blog/server-power-versus-inlet-temperature/.
18. 2022 Best Practice Guidelines for the EU Code of Conduct on Data Centre Energy Efficiency. https://e3p.jrc.ec.europa.eu/publications/2022-best-practice-guidelines-eu-code-conduct-data-centre-energy-efficiency.
19. Fleischer, A.S. (2020). Cooling our insatiable demand for data. Science *370*, 783–784. https://doi.org/10.1126/science.abe5318.
20. Niemann, J., Brown, K., and Avelar, V. (2011). Impact of Hot and Cold Aisle Containment on Data Center Temperature and Efficiency, *135* (Schneider Electric Data Center Science Center, White Paper), pp. 1–14.
21. Habibi Khalaj, A., Scherer, T., Siriwardana, J., and Halgamuge, S.K. (2015). Multi-objective efficiency enhancement using workload spreading in an operational data center. Appl. Energy *138*, 432–444. https://doi.org/10.1016/j.apenergy.2014.10.083.
22. Asgari, S., MirhoseiniNejad, S., Moazamigoodarzi, H., Gupta, R., Zheng, R., and Puri, I.K. (2021). A gray-box model for real-time transient temperature predictions in data centers. Appl. Therm. Eng. *185*, 116319. https://doi.org/10.1016/j.applthermaleng.2020.116319.
23. Gözcü, O., Özada, B., Carfi, M.U., and Erden, H.S. (2017). Worldwide Energy Analysis of Major Free Cooling Methods for Data Centers, pp. 968–976.
24. Ham, S.-W., Kim, M.-H., Choi, B.-N., and Jeong, J.-W. (2015). Simplified server model to simulate data center cooling energy consumption. Energy Build. *86*, 328–339. https://doi.org/10.1016/j.enbuild.2014.10.058.
25. Lei, N., and Masanet, E. (2020). Statistical analysis for predicting location-specific data center PUE and its improvement potential. Energy *201*, 117556. https://doi.org/10.1016/j.energy.2020.117556.
26. ASHRAE (2021). Standard 169-2021, Climatic Data for Building Design Standards.
27. Taylor, S.T. (2012). Optimizing design & control of chilled water plants: Part 4: Chiller & cooling tower selection. ASHRAE J. *54*, 60–70.
28. ASHRAE (2020). ASHRAE Handbook HVAC Systems and Equipment.
29. https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/.
30. https://energyplus.net/.
31. Huang, S., Fu, S., Zhang, Q., and Shi, W. (2015). Characterizing Disk Failures with Quantified Disk Degradation Signatures: An Early Experience, pp. 150–159.
32. Sankar, S., Shaw, M., and Vaid, K. (2011). Impact of Temperature on Hard Disk Drive Reliability in Large Datacenters, pp. 530–537.
33. Reducing Data Center Cost with an Air Economizer. https://www.intel.com/content/dam/doc/technology-brief/data-center-efficiency-xeon-reducing-data-center-cost-with-air-economizer-brief.pdf.
34. Alihosseini, Y., Zabetian Targhi, M., Heyhat, M.M., and Ghorbani, N. (2020). Effect of a micro heat sink geometric design on thermo-hydraulic performance: A review. Appl. Therm. Eng. *170*, 114974. https://doi.org/10.1016/j.applthermaleng.2020.114974.
35. Kandasamy, R., Ho, J.Y., Liu, P., Wong, T.N., Toh, K.C., and Chua, S.J. (2022). Two-phase spray cooling for high ambient temperature data centers: Evaluation of system performance. Appl. Energy *305*, 117816. https://doi.org/10.1016/j.apenergy.2021.117816.
36. Pan, C., Zhai, J., and Wang, Z.L. (2019). Piezotronics and Piezo-phototronics of Third Generation Semiconductor Nanowires. Chem. Rev. *119*, 9303–9359. https://doi.org/10.1021/acs.chemrev.8b00599.
37. Niemann, J., Bean, J., and Avelar, V. (2011). Economizer Modes of Data Center Cooling Systems, *160* (Schneider Electric Data Center Science Center Whitepaper).
38. Potts, Z. (2011). Free Cooling Technologies in Data Centre Applications (SUDLOWS White Paper).
39. Lui, Y.Y. (2010). Waterside and Airside Economizers Design Considerations for Data Center Facilities. Build. Eng. *116*, 98–108.
40. Braun, J.E. (1988). Methodologies for the Design and Control of Central Cooling Plants (The University of Wisconsin-Madison).
41. Taylor, S.T. (2014). How to design & control waterside economizers. ASHRAE J. *56*, 30–36.
42. TRNSYS 18, A transient systems simulation program, http://sel.me.wisc.edu/trnsys.
43. Ma, Z., Wang, S., Xu, X., and Xiao, F. (2008). A supervisory control strategy for building cooling water systems for practical and real time applications. Energy Convers. Manag. *49*, 2324–2336. https://doi.org/10.1016/j.enconman.2008.01.019.
44. Lu, Y., Wang, S., Zhao, Y., and Yan, C. (2015). Renewable energy system optimization of low/zero energy buildings using single-objective and multi-objective optimization methods. Energy Build. *89*, 61–75. https://doi.org/10.1016/j.enbuild.2014.12.032.
45. Cheung, H., and Wang, S. (2019). Optimal design of data center cooling systems concerning multi-chiller system configuration and component selection for energy-efficient operation and maximized free-cooling. Renew. Energy *143*, 1717–1731. https://doi.org/10.1016/j.renene.2019.05.127.
46. Pontes, R.F., Yamauchi, W.M., and Silva, E.K. (2019). Analysis of the effect of seasonal climate changes on cooling tower efficiency, and strategies for reducing cooling tower power consumption. Appl. Therm. Eng. *161*, 114148. https://doi.org/10.1016/j.applthermaleng.2019.114148.
47. Evans, T. (2012). The Different Technologies for Cooling Data Centers, *59* (APC white paper).