GLOBAL ENERGY PERFORMANCE ASSESSMENT AND OPTIMIZATION OF DATA CENTER COOLING SYSTEMS

YINGBO ZHANG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Building Environment and Energy Engineering

Global Energy Performance Assessment and Optimization of Data Center Cooling Systems

YINGBO ZHANG

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Yingbo Zhang (Name of student)

ABSTRACT

Abstract of thesis entitled: Global Energy Performance Assessment and Optimization of Data Center Cooling Systems

Submitted by: Yingbo Zhang
For the degree of: Doctor of Philosophy
at The Hong Kong Polytechnic University in December 2024

As the digital backbone of our increasingly interconnected world, energy-intensive data centers pose a significant challenge to global decarbonization. Global data center electricity use in 2021 was 220-320 TWh, around 0.9%-1.3% of global electricity demand. The energy use per square foot in data centers can be 100 times that of typical office buildings. Notably, the cooling energy required to keep the servers in data centers from overheating is on par with that of the servers themselves, representing 30%–40% of the total energy consumption of data centers. Developing highly efficient cooling systems in data centers is a key challenge for the decarbonization of the data center industry.

Therefore, this study aims to conduct a comprehensive assessment of the energy performance of data centers and develop cutting-edge methods or technologies to fundamentally improve the energy efficiency of data centers. Firstly, a thorough review of next-generation high-temperature data centers and the categorization of existing methods/strategies for enhancing the energy efficiency of data center cooling systems is conducted. Secondly, the global energy impact of high-temperature data centers is comprehensively assessed. Thirdly, the energy performance of data center cooling systems under various conditions is analyzed systematically. A comprehensive methodology for the optimal design of data center cooling systems is developed considering progressive loading and life-cycle energy performance. Fourthly, the optimal dispatch strategy and design scenario for energy storage systems in data centers are investigated to unlock their great flexibility for smart grid services. Lastly, the energy, economic and carbon impacts of the national initiative ‘Eastern Data, Western Computing’ are assessed comprehensively by analyzing three major migration routes.

To tackle the problem of high and increasing data center energy consumption, 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. This study conducts a comprehensive review of

high-temperature data centers, especially their key advantages and the primary challenges associated with their implementation, as well as the existing efforts and latest technologies to tackle the bottlenecks. Future perspectives for the development and applications of the high-temperature data center are also discussed. Furthermore, the global energy impacts of high-temperature data centers are quantified and analyzed. The trade-off between cooling energy savings and server power rise is critically analyzed and discussed. Moreover, quantitative guidance and targets for developing 'ideal' and 'recommendable' servers for high-temperature data centers are established for IT and server professionals to further develop IT equipment and servers that take the data center cooling energy into account. 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 the current typical space temperature of 22°C.

To develop a life-cycle optimal design for data center cooling systems, the energy performance of data center cooling systems is systematically analyzed under full-range cooling loads and climate conditions. The energy performance of typical cooling systems is quantified under a typical progressive loading throughout the data center's lifecycle. An optimal design method is developed for centralized cooling systems with multiple chillers under progressive loading. The optimal designs in different climate zones are determined according to the energy performance under full-range loads and ambient temperatures. Free cooling hours, cooling energy, and life-cycle costs of the optimized designs and conventional designs are analyzed and compared comprehensively. The results show that the optimized cooling systems could operate more energy-efficiently, despite decreased free cooling hours (13-860). Significant cooling energy savings over the lifespan could be achieved, i.e., 4-22%, corresponding to the PUE reductions of 0.02-0.11, depending on climate conditions and control strategies.

This study also pioneers the idea of utilizing surplus energy storage capacity in data centers to offer grid flexibility services, considering progressive loading throughout their lifecycle. Two optimization problems are formulated for optimal energy storage dispatch in operation and for storage system design optimization respectively. The objective for optimal dispatch is to minimize the electricity cost, by efficiently allocating battery and cold storage capacities. The objective for design optimization is to minimize the life-cycle costs including investments and operation cost savings under typical loading conditions and electricity markets. Two typical electricity markets (the Guangdong electricity market and the CAISO electricity market) and

four investment scenarios for energy storage systems are considered. The impacts of discount rates and battery prices on the life-cycle economic benefits of energy storage systems are also analyzed comprehensively. The participation of data centers in grid flexibility services demonstrates significant economic benefits. Over its lifetime, the battery storage can achieve economic benefits of $1.6 million, which is 1.29 times its total investment. The cold storage can achieve economic benefits of $0.35 million, which is 2.39 times its total investment.

To facilitate the decarbonization of data centers, the Chinese government launched an ambitious initiative, called ‘Eastern Data, Western Computing’. The national initiative aims to migrate computing workloads from electricity-deficient Eastern regions to renewable-rich Western regions. A comprehensive assessment is conducted concerning its energy, economic and carbon impacts by analyzing three major migration routes. Future perspectives and challenges on carbon emission reduction of the initiative are analyzed. Potential policy suggestions and actionable insights are proposed to address these challenges. We found that ‘moving bits’ is much more energy efficient than ‘moving watts’, but not necessarily beneficial for decarbonization. The national initiative shows significant energy-saving potential, 332-942 GWh (4.8-12.5%) annually, attributed to reduced cooling energy and eliminated power-transmission loss. However, no economic benefit is observed if considering the high capital costs for constructing duplicated data centers in Western regions. The carbon emission benefits in different routes are significantly different. Shanghai-Sichuan route could reduce carbon emissions by up to 2803 $KtCO_2e$ (79.6%) annually, whereas Beijing-Inner Mongolia route exhibits a notable increase (1164 $KtCO_2e$ (24.9%)) in carbon emissions.

PUBLICATIONS ARISING FROM THE THESIS

Journal Papers

[1] **Yingbo Zhang**, Kuishan, Xiuming Li, Hangxin Li and Shengwei Wang*. Research and Technologies for next-generation high-temperature data centers–State-of-the-arts and future perspectives. *Renewable and Sustainable Energy Reviews*. 2023;171:112991.

[2] **Yingbo Zhang**, Hangxin Li and Shengwei Wang*. The global energy impact of raising the space temperature for high-temperature data centers. *Cell Reports Physical Science*. 2023;4(10):101624.

[3] **Yingbo Zhang**, Hangxin Li and Shengwei Wang*. Life-Cycle Optimal Design and Energy Benefits of Centralized Cooling Systems for Data Centers Concerning Progressive Loading. *Renewable Energy*. 2024;230:120847.

[4] **Yingbo Zhang**, Hangxin Li and Shengwei Wang*. Energy performance analysis of multi-chiller cooling systems for data centers concerning progressive loading throughout the lifecycle under typical climates. *Building Simulation*. Beijing: Tsinghua University Press, 2024, 17(10): 1693-1708.

[5] **Yingbo Zhang**, Hong Tang, Hangxin Li and Shengwei Wang*. Optimal Design and Dispatch of Hybrid Storage Systems in Data Centers, Unlocking the Flexibilities of Data Centers in Electricity Markets. *Energy*. Under review.

[6] **Yingbo Zhang**, Hangxin Li and Shengwei Wang*. Eastern Data, Western Computing": The energy, economic and carbon benefits for China's data centers. *Engineering*. Under review.

[7] Xiuming Li, Mengyi Li, **Yingbo Zhang**, Zongwei Han, Shengwei Wang*. Rack-level Cooling Technologies for Data Centers – A Comprehensive Review. *Journal of Building Engineering*. 2024;90: 109535.

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest and sincerest gratitude to my supervisor, Professor Shengwei Wang. Throughout my research journey, he has been incredibly supportive, patient, and encouraging. He has not only guided me on how to carry out research studies but has also taught me how to be a good presenter. His mentorship has taught me the importance of patience and perseverance in conducting high-impact research work. I am truly grateful for his guidance and for shaping me into a better researcher. I am also grateful to Professor Fu Xiao, whose invaluable advice and care have been instrumental during my three years of PhD studies. Their wisdom and guidance have greatly contributed to my research work.

Furthermore, I extend my appreciation to all the team members of the BEAR research group, especially Dr. Kui Shan and Dr. Hangxin Li. Their continuous support, follow-up on my research work, and insightful feedback have been invaluable. I am also thankful to Dr. Hong Tang, Dr. Wenxuan Zhao, and Dr. Ao Li for generously sharing their research experience and wisdom with me. Special thanks to Miss Jing Zhang for her academic assistance, modeling guidance, and engaging discussions on various topics. Her support and encouragement have meant a lot to me.

I would like to acknowledge The Hong Kong Polytechnic University and my supervisor again for providing me with the opportunity to undertake a three-month exchange study at DTU in Denmark. The exchange program broadened my horizons, enabling me to gain new perspectives, insights, and interdisciplinary knowledge. This experience has contributed significantly to my personal and academic growth.

To my badminton buddies at Hong Kong Polytechnic University, who played alongside me and engaged in friendly matches, I am grateful for the laughter and camaraderie we shared on and off the badminton court. Their companionship has made my PhD journey more enriching and exciting.

Lastly, I am deeply grateful to my parents, Yan Zhang and Lihong Yang, for their unconditional love and sacrifices. They have consistently shown me trust, support, and understanding. I would also like to express my heartfelt appreciation to all my friends for their encouragement during challenging times.

TABLE OF CONTENTS

GLOBAL ENERGY PERFORMANCE ASSESSMENT AND OPTIMIZATION OF DATA CENTER COOLING SYSTEMS I

CERTIFICATE OF ORIGINALITY I

ABSTRACT II

PUBLICATIONS ARISING FROM THE THESIS V

ACKNOWLEDGEMENTS VI

TABLE OF CONTENTS VII

LIST OF FIGURES X

LIST OF TABLES XIV

NOMENCLATURE XVI

CHAPTER 1 INTRODUCTION 1

1.1 Background and motivation 1

1.2 Aim and objectives 2

1.3 Organization of thesis 3

CHAPTER 2 LITERATURE REVIEW 6

2.1 Overview of high-temperature data centers 6

2.2 Energy-Efficient Technologies for Data Center Cooling 29

2.3 Opportunities for data centers to interact with the power grid 31

2.4 Computing workload migration in geo-distributed data centers 33

2.5 Summary of identified research gaps 33

CHAPTER 3 THE GLOBAL ENERGY IMPACT OF RAISING THE SPACE TEMPERATURE FOR HIGH-TEMPERATURE DATA CENTERS 35

3.1 Development of the cooling system model and typical operation modes 36

3.2 Performance of data center cooling systems under different ambient air temperatures 41

3.3 Impact of higher space temperatures on free-cooling hours worldwide 43

3.4 Impact of raising space temperature on cooling-energy savings worldwide 45

3.5 Impact of raising space temperature on data center energy performance and 'ideal servers' ... 46

3.6 Summary ... 49

CHAPTER 4 ENERGY PERFORMANCE ANALYSIS OF MULTI-CHILLER COOLING SYSTEMS FOR DATA CENTERS CONCERNING PROGRESSIVE LOADING THROUGHOUT THE LIFECYCLE UNDER TYPICAL CLIMATES ... 51

4.1 Description of multi-chiller cooling systems in the referenced data centers ... 51

4.2 Development of the cooling system model ... 55

4.3 Outline of energy performance assessment and specific cooling system designs ... 59

4.4 Energy performance under full-range loads and climate conditions ... 62

4.5 Summary ... 71

CHAPTER 5 LIFE-CYCLE OPTIMAL DESIGN AND ENERGY BENEFITS OF CENTRALIZED COOLING SYSTEMS FOR DATA CENTERS CONCERNING PROGRESSIVE LOADING ... 73

5.1 Formulation for optimizing data center cooling system concerning progressive loading ... 73

5.2 Description of baseline scenario and economic analysis ... 78

5.3 Results for optimized designs of cooling systems in different climate conditions ... 79

5.4 Energy and cost benefits of optimized designs in different climate conditions ... 83

5.5 Summary ... 87

CHAPTER 6 UNLOCKING THE FLEXIBILITIES OF DATA CENTERS IN SMART GRID MARKETS: OPTIMAL DISPATCH AND DESIGN OF ENERGY STORAGE SYSTEMS CONSIDERING PROGRESSIVE LOADING ... 89

6.1 Formulation of optimization problems considering effective use of surplus capacity of storage systems in flexibility markets ... 89

6.2 Description of energy storage systems in the referenced data center ... 96

6.3 Optimal dispatch results utilizing surplus capacity in data centers ... 97

6.4 Life-cycle economic benefits under different electricity markets ... 99

6.5 Impacts of discount rate and battery price on life-cycle economic benefits ... 102

6.6 Summary ... 105

CHAPTER 7 "EASTERN DATA, WESTERN COMPUTING": THE ENERGY, ECONOMIC AND CARBON BENEFITS FOR CHINA'S DATA CENTERS ... 107

7.1 Motivations for “Eastern Data, Western Computing” 107
7.2 Development of data transmission model 109
7.3 Analysis of energy saving, economic benefits and carbon emission reductions 111
7.4 Energy performance of cooling systems and data-transmission energy 117
7.5 Impacts on energy, economy and carbon emissions 119
7.6 Future perspectives on carbon emission reduction and abatement cost 124
7.7 Discussion and policy implications 126
7.8 Summary 127
CHAPTER 8 CONCLUSIONS AND FUTURE WORKS 129
8.1 Main contributions of this study 129
8.2 Conclusions 130
8.3 Recommendation for future work 134
REFERENCE 136

LIST OF FIGURES

Fig. 1.1 Organization of main chapters....5

Fig. 2.1 Ambient temperature classification of IT equipment by ASHRAE [14] 7

Fig. 2.2 Main benefits and concerns of high-temperature data centers 8

Fig. 2.3 Air-side free cooling map of Europe [42] 9

Fig. 2.4 Key challenges to the implementation of high-temperature data centers....10

Fig. 2.5 Breakdown of hardware component errors in a large data center - Microsoft ...13

Fig. 2.6 Breakdown of hardware component errors in a data center - Baidu 13

Fig. 2.7 Distribution of average temperatures and failure rates of HDDs [60] 14

Fig. 2.8 The monthly probability of HDD failure as a function of temperature [32] 14

Fig. 2.9 Thermal margin of key components in servers [52]....17

Fig. 2.10 Server power rise vs Ambient temperature range [29]....18

Fig. 2.11 Server flow rate increase versus ambient temperature increase 19

Fig. 2.12 Server fan behavior versus inlet temperatures [14]....19

Fig. 2.13 Key Component Power versus temperature 20

Fig. 2.14 Total power estimation in data centers 21

Fig. 2.15 Main feasible solutions at different levels for implementing high-temperature data centers....23

Fig. 2.16 Hot-aisle/cold-aisle configuration [82]....24

Fig. 3.1 Schematic of a cooling system unit. 36

Fig. 3.2 Chiller COP at the temperature of leaving condenser water of 28°C....37

Fig. 3.3 The pressure drop of the chilled water loop 40

Fig. 3.4 Schematics of typical cooling operation modes 41

Fig. 3.5 Cooling system COP versus wet-bulb temperature....42

Fig. 3.6 Annual free cooling ratio and cooling energy savings in 19 climate zones 44

Fig. 3.7 Global maps of annual free-cooling ratio at different space temperatures....45

Fig. 3.8 Global maps of cooling-energy savings with reference to baseline space temperature 22°C 46

Fig. 3.9 Impact of raising space temperature on the annual free-cooling ratio and normalized server power and total power in data centers in different cities 47

Fig. 4.1 Schematic of the multi-chiller cooling system 52

Fig. 4.2 Procedure to select the optimal operation mode 53

Fig. 4.3 Procedure and steps of energy performance assessment 59

Fig. 4.4 Schematic of specific cooling system designs 60

Fig. 4.5 Electricity flow in a data center 61

Fig. 4.6 Energy performance of cooling system components with the specific design (Fig. 4.4(A)), chiller (A), chilled water pump (B), open cooling tower (C), and cooling water pump (D) 62

Fig. 4.7 Energy performance of cooling system components with the specific design (Fig. 4.4(B)), chiller (A), chilled water pump (B), closed-circuit cooling tower (C) and cooling water pump (D) 64

Fig. 4.8 Operating hours of three cooling modes in different cities 66

Fig. 4.9 Annual average data center PUE and cooling system COP at part-load ratios in different cities 67

Fig. 4.10 Comparison of cooling system COP at PLRs of 0.9 and 1.0 over the typical year in Kunming 68

Fig. 4.11 Annual average cooling system COP (A) and data center PUE (B) under full-range cooling loads in different cities 69

Fig. 4.12 A typical IT load growth (A) and cooling system COP (B) over a lifetime of 10 years 71

Fig. 4.13 A typical IT load growth (A) and cooling system COP (B) over a lifetime of 20 years 71

Fig. 5.1 Typical progressive IT loading of data centers 74

Fig. 5.2 Procedure and steps of energy performance assessment 75

Fig. 5.3 Global map of 19 climate zones 75

Fig. 5.4 Schematic of a data center cooling system in the baseline scenario. 78

Fig. 5.5 The capital cost of chillers versus capacity 79

Fig. 5.6 Cooling energy savings (A) and total cost savings (B) under different numbers of cooling units when adopting the CPS control strategy in Climate Zone 4A 80

Fig. 5.7 Cooling energy savings (A) and total cost savings (B) under different numbers of cooling units when adopting the OPR control strategy in Climate Zone 4A.....81

Fig. 5.8 Operation modes under full-range cooling loads and ambient air temperatures 83

Fig. 5.9 Energy performance of the cooling unit under different ambient temperatures and part load ratios under CPS control strategy (A) and OPR control strategy (B).....84

Fig. 5.10 Impact on free cooling hours worldwide under CPS (A) and OPR (B) control strategies 84

Fig. 5.11 Worldwide system COP increase under CPS and OPR control strategies.....85

Fig. 5.12 Worldwide cooling energy savings under CPS and OPR control strategies 86

Fig. 5.13 PUE reduction worldwide under CPS (A) and OPR (B) control strategies 87

Fig. 5.14 Total cost savings worldwide under CPS (A) and OPR (B) control strategies 87

Fig. 6.1 Typical progressive IT loading of data centers 90

Fig. 6.2 Flowchart of dispatch optimization of surplus energy storage in data centers...92

Fig. 6.3 Surplus capacities of different scenarios throughout the lifecycle 94

Fig. 6.4 The framework of life-cycle analysis 96

Fig. 6.5 (A) Time of Use (ToU) energy, regulation and operating reserve prices in Guangdong, China; Optimal dispatch results of (B) EES (C) TES. 98

Fig. 6.6 (A) Hourly energy, frequency regulation and operating reserve prices in 01/25, 04/25, and 09/27 from CAISO in the US; Optimal dispatch results in three typical days of (B) EES (C) TES. 99

Fig. 6.7 Life-cycle economic benefits of different scenarios of (A) EES, (B) TES under the Guangdong electricity market, China 100

Fig. 6.8 Life-cycle economic benefits of different scenarios of (A) EES, (B) TES under CAISO electricity market 101

Fig. 6.9 Life-cycle economic benefits of EES versus discount rates and battery price decline rates under the Guangdong electricity market 103

Fig. 6.10 Life-cycle economic benefits of EES versus discount rates and battery price decline rates under the CAISO electricity market 103

Fig. 6.11 Life-cycle economic benefits of TES under the Guangdong electricity market at discount ranging from 2% to 10% 104

Fig. 6.12 Life-cycle economic benefits of TES under the CAISO electricity market at discount ranging from 2% to 10% 105

Fig. 7.1 Data center energy consumption, major migration routes, performance of cooling systems, and data-transmission energy intensity 118

Fig. 7.2 Energy and economic benefits when data transmission on the backbone networks 120

Fig. 7.3 Carbon emission reductions when data transmission on the backbone networks 122

Fig. 7.4 Energy, economic and carbon benefits when data transmission on dedicated lines 124

Fig. 7.5 Quantitative carbon emission reductions considering potential increases or decreases in the difference in CO2e factor between the two locations involved in a route. (A) Beijing-Inner Mongolia route (B) Shanghai-Sichuan route (C) Guangzhou-Guizhou route 125

Fig. 7.6 Net carbon abatement cost considering future potential changes in CO2e factor difference of power consumption 126

LIST OF TABLES

Table 2.1 The relative failure rate of volume servers as a function of server inlet temperature [14] 11

Table 2.2 Temperature limits of components in typical servers [56] 12

Table 2.3 Evaluated Components AFR (Average Failure Rate) [37] 15

Table 2.4 Server performance at different inlet temperatures [34] 16

Table 3.1 The specification of the cooling system [129]. 36

Table 3.2 The empirical coefficients and parameters of the chiller model 37

Table 4.1 Specification of the data center cooling system in the case study 52

Table 4.2 Number of chillers, water-side economizers, and cooling towers in operation 53

Table 4.3 Number of operating cooling water pumps in operation 54

Table 4.4 The PLR data of chiller model Type 142 55

Table 4.5 The performance data of chiller model Type 142 56

Table 4.6 Selected climate zones and cities 60

Table 5.1 Climate zones and 19 representative cities [175] 75

Table 5.2 Optimal capacity combinations under different numbers of cooling units and optimal PLRs under different IT loadings (CPS) 79

Table 5.3 Optimal capacity combinations under different numbers of cooling units and optimal PLRs under different IT loadings (OPR strategy) 81

Table 5.4 Optimized designs in different climate conditions 82

Table 5.5 Optimal combinations and corresponding PLRs under progressive loading 82

Table 6.1 Specifications of energy storage systems in the referenced data center 96

Table 7.1 Historical development of energy intensity of data transmission from 2004 to 2021 109

Table 7.2 The analytical parameters of the data-transmission model [241, 242] 110

Table 7.3 Transmission loss for cross-region power transmission in typical cases 112

Table 7.4 Characteristics of the 10 MW* typical data center used for analysis [1] 112

Table 7.5 Energy consumption breakdown in a typical data center [3, 249].....112

Table 7.6 The amount of data in data centers [250, 251] 113

Table 7.7 CO_2 emissions factors of major cities involved in this study [227].....114

Table 7.8 CO_2 Emissions factors associated with raw material and production of data transmission and power transmission 114

Table 7.9 The amortized cost for data centers and dedicated fiber-optic lines [226]....115

Table 7.10 Electricity prices in major cities involved in this study.....115

Table 7.11 Nomenclature for Equations 116

NOMENCLATURE

Abbreviations

PUE	*Power usage effectiveness*
COP	*Coefficient of performance*
HVAC	*Heating, ventilation, and air conditioning*
DIMM	*Dual Inline Memory Module*
PSU	*Power Supply Unit*
HDD	*Hard Disk Drive*
SSD	*Solid State Drive*
SAS	*Serial Attached Small Computer System Interface*
BMC	*Burst Mode Controller*
EEPROM	*Electrically Erasable Programmable read only memory*
CPU	*Central Processing Unit*
PCH	*Paging Indicator Channel*
MEM	*Memory Device*
PCB	*Printed Circuit Board*
NIC	*Network Interface Card*
PLR	*Part load ratio*
CPS	*Constant pressure setting*
OPR	*Near-optimal pressure resetting*
EE	*Energy efficiency*
TCOSP	*Total coefficient of the system performance*
COSP	*Coefficient of the system performance*
LTC	*Life-cycle total cost*
N	*Number of cooling units*
L	*IT loading at different stages*
CC	*Capital cost ($)*
OC	*Operating cost ($)*
CP	*Capacity (kW)*
P	*Power consumption (kWh)*
Q	*Cooling load (kW)*
T	*Temperature (℃)*

W	*Energy consumption (kWh)*
V	*Air flow rate (m^3/h)*
CRAH	*Computer room air handler*
EES	*Electrical energy storage*
TES	*Thermal energy storage*
Cap	*Capacity*
FR	*Frequency regulation*
SR	*Spinning reserve*
ToU	*Time of Use*
MILP	*Mixed integer linear programming*
MIQP	*Mixed integer quadratic programming*
CAISO	*California Independent System Operator*
SOC	*Battery state of charge*
AGC	*Automatic Generation Control*
REV	*Revenue*
LCC	*Life-cycle cost*
INV	*Investment*
C	*Cost*
OM	*Operation and maintenance costs*
COSP	*Coefficient of performance of the cooing system*
TMY	*Typical meteorological year*
r	*Discount rate*
d	*Annual decline rate of battery price*
c	*Capacity of each cooling unit (kW)*
p	*Part load ratio of each cooling unit*
a	*Empirical coefficients for the chiller model*
m	*Flow rate (m^3/s)*
f	*The speed of the pump*

Symbols

h_D	*Mass transfer coefficient*
A_v	*Surface area of water droplets per tower cell exchange volume*
V_{cell}	*Total tower cell exchange volume*

C_{pw}	*Constant pressure specific heat of water*
C_s	*The saturation-specific heat*
h_{sat}	*Heat transfer coefficient at saturated air condition*
h_T	*Heat transfer coefficient between the fluid and the air at T*
c_{pump}	*The parameter of the pump model*

Greek letters

π	*Price*
τ	*Time weighting factor*
ε	*Effectiveness*
γ	*The ratio of flow rate to design flow rate*
η	*Efficiency*

Subscripts

i	*Cooling unit number*
s	*Stage*
j	*Part load ratio number*
ch	*Chiller*
ct	*Cooling tower*
hx	*Heat exchanger*
des	*Design*
in	*Inlet*
out	*Outlet*
cd	*Condenser*
wet	*Wet bulb*
cwp	*Chilled water pumps*
cooling	*Cooling system*
IT	*IT equipment*
w	*Water*
req	*required*
op	*operation*
wet	*wet-bulb*
SA	*supplied air*

CHAPTER 1 INTRODUCTION

1.1 Background and motivation

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% of global electricity demand [4]. It is estimated that global data center electricity use will increase to 848 TWh by 2030 [5]. Notably, the cooling energy required to keep the servers in data centers from overheating is on par with that of the servers themselves, representing 30%–40% of the total energy consumption of data centers [6, 7].

Data center cooling systems can be broadly classified into two main categories: *i),* air-cooled data centers [8] including centralized (chilled-water) cooling systems (air/water-cooled chillers) [9], direct expansion cooling systems [10], direct/indirect air (evaporative) cooling systems [11]; *ii),* liquid-cooled data centers [12], including direct/indirect liquid cooling [13]. Liquid cooling is considered a more efficient method, but often hampered by the conservative minds of designers and operators, and technological and economic challenges, such as maintenance and investment [14]. Centralized cooling systems with water-side economizers are the most widely used in large data centers due to their high reliability [15] and physical practicality [16].

Data center cooling technology has progressed significantly over time [17]. Improvements in air-handling design, particularly for data centers, were the first major steps for increasing cooling efficiency [18]. 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 [19]. A similarly improved coordination control strategy between the server and cooling sides can achieve a better match of cooling supply and demand [20]. Furthermore, recently developed AI-based real-time transient temperature predictions of server CPUs and the cold chamber help to better manage airflow in data centers [21]. Although these technologies can improve cooling efficiency, the resulting energy savings are limited as they only improve the efficiency of the ‘cold distribution’. However, there are still many significant challenges faced by data center cooling systems.

1) Existing research has proposed different approaches to enhance the energy efficiency of data center cooling systems. However, few studies systematically investigate and quantify the global energy-saving potential of high-temperature data centers. Understanding and quantifying this potential is crucial as it serves as a fundamental solution for improving the efficiency of data center cooling systems by changing the cooling mechanism.
2) Current studies on energy performance for data center cooling systems primarily focus on design conditions and overlook the system's life-cycle operation. However, it is essential to quantify the energy performance of these systems under partial loads throughout the data center lifetime and across various climate conditions. A thorough understanding of the system's energy performance over its life cycle is crucial for informing improved design and control strategies.
3) Existing studies overlook the progressive loading throughout data centers' lifecycle, which results in low energy-efficiency operation of data center cooling systems throughout the lifecycle. Therefore, there is a significant research gap in developing optimal designs and control strategies for data center cooling systems under progressive loading conditions.
4) Data centers can act as pivotal players in grid stability by providing flexible services, which can also generate revenue to reduce their energy costs, creating a win-win situation. However, the optimal dispatch and system designs of the energy storage systems in data centers to provide grid flexibility services have not been investigated.
5) The migration of computing workloads in geographically distributed data centers has emerged as a significant trend in the era of Artificial Intelligence (AI). The Chinese government launched an ambitious initiative, called 'Eastern Data, Western Computing' to facilitate the decarbonization of data centers. The national initiative aims to migrate computing workloads from electricity-deficient Eastern regions to renewable-rich Western regions. Currently, there is a lack of comprehensive assessment regarding its energy, economic, and carbon impacts on the data center industry.

1.2 Aim and objectives

This PhD study aims to conduct a comprehensive assessment of the energy performance of data centers and develop cutting-edge methods or technologies to fundamentally improve the energy efficiency of data centers. The following objectives are addressed to accomplish the research aim:

1) Conduct a thorough and systematic literature review of state-of-the-art research on high-temperature data centers. Categorize and analyze systematically the main benefits and the

major bottlenecks for implementing high-temperature data centers as well as the existing efforts and latest technologies to tackle the bottlenecks.

2) Quantify the global energy impacts of high-temperature data centers, and critically analyze the trade-off between cooling energy savings and server power rise. Establish quantitative guidance and targets for developing 'ideal' and 'recommendable' servers for high-temperature data centers.
3) Conduct a comprehensive assessment and analysis of the energy performance of centralized cooling systems for data centers under full-range loads and climate conditions. Quantify the life-cycle energy performance of centralized cooling systems in data centers under progressive loading.
4) Develop optimal designs for centralized cooling systems in data centers concerning life-cycle energy performance and progressive loading. Identify optimal designs of the cooling system in different climate conditions, and analyze the energy benefits of these designs under varying climate conditions.
5) Develop optimal dispatch strategies to effectively utilize the surplus capacity of energy storage systems in data centers to provide grid flexibility services considering progressive loading. Identify the optimal design scenario of energy storage systems by analyzing life-cycle economic benefits. Discuss the impacts of discount rates and battery prices on the life-cycle economic benefits of energy storage systems in data centers.
6) Conduct a comprehensive assessment of the energy, economic and carbon impacts of the national initiative 'Eastern Data, Western Computing'. Discuss the energy and carbon emissions trade-offs associated with some routes in the initiative. Analyze the future perspectives and challenges on carbon emission reduction posted by the initiative. Propose policy suggestions and actionable insights to address the potential challenges.

1.3 Organization of thesis

This chapter presents the background and motivations for the investigation of high-temperature data centers, the performance assessment and optimization of data center cooling systems, the optimal dispatch and design of energy storage systems in data centers, and the comprehensive assessment of the national initiative 'Eastern Data, Western Computing'. Other chapters of this thesis are organized as follows.

Chapter 2 presents an overview of research and technologies for next-generation data centers. The main benefits and the major bottlenecks for implementing high-temperature data centers, as well as the existing efforts and latest technologies to tackle these bottlenecks, are categorized

and analyzed systematically. The technologies and methods for energy efficiency improvements in data center cooling systems are also summarized and presented. In addition, the opportunities for data centers to interact with the power grid and the trends and benefits of computing workload migration in geo-distributed data centers are overviewed and discussed.

Chapter 3 presents the global energy impact of raising the space temperature for high-temperature data centers. The trade-off between cooling-energy savings and server power rise is critically analyzed. Quantitative guidance and targets are established for developing 'ideal' and 'recommendable' servers, considering the server performance associated with the thermal environment.

Chapter 4 presents a thorough assessment of the energy performance of multi-chiller cooling systems throughout the entire lifecycle. The energy performance of the cooling system is systematically analyzed under full-range cooling loads and climate conditions. The energy performance of typical cooling systems is quantified under typical progressive loading experienced throughout the data center's lifecycle.

Chapter 5 presents an optimal design method for centralized cooling systems with multiple chillers under progressive loading. The optimal designs in different climate zones are determined according to the energy performance under full-range loads and ambient temperatures. Free cooling hours, cooling energy, and life-cycle costs of the optimized designs and conventional designs are analyzed and compared comprehensively.

Chapter 6 presents a pioneering approach that leverages the surplus capacity of energy storage systems for emergencies in data centers to participate in flexible grid services. Optimal dispatch strategies of energy storage systems are developed to minimize electricity costs. The life-cycle economic benefits of proposed design scenarios are quantified and analyzed under two electricity markets. The impacts of discount rates and battery prices on the life-cycle economic benefits of energy storage systems are discussed comprehensively.

Chapter 7 presents a comprehensive and quantitative assessment of the energy-saving potentials, economic benefits, and carbon emission reductions of the national initiative 'Eastern Data, Western Computing'. The energy and emissions trade-offs associated with each route in the initiative are discussed. Future perspectives and challenges on carbon emission reduction of the initiative are analyzed. Potential policy suggestions and actionable insights are proposed to address these challenges.

Chapter 8 summarizes the main contributions and conclusions of the work conducted in this PhD project and gives recommendations for future research on the subjects concerned.

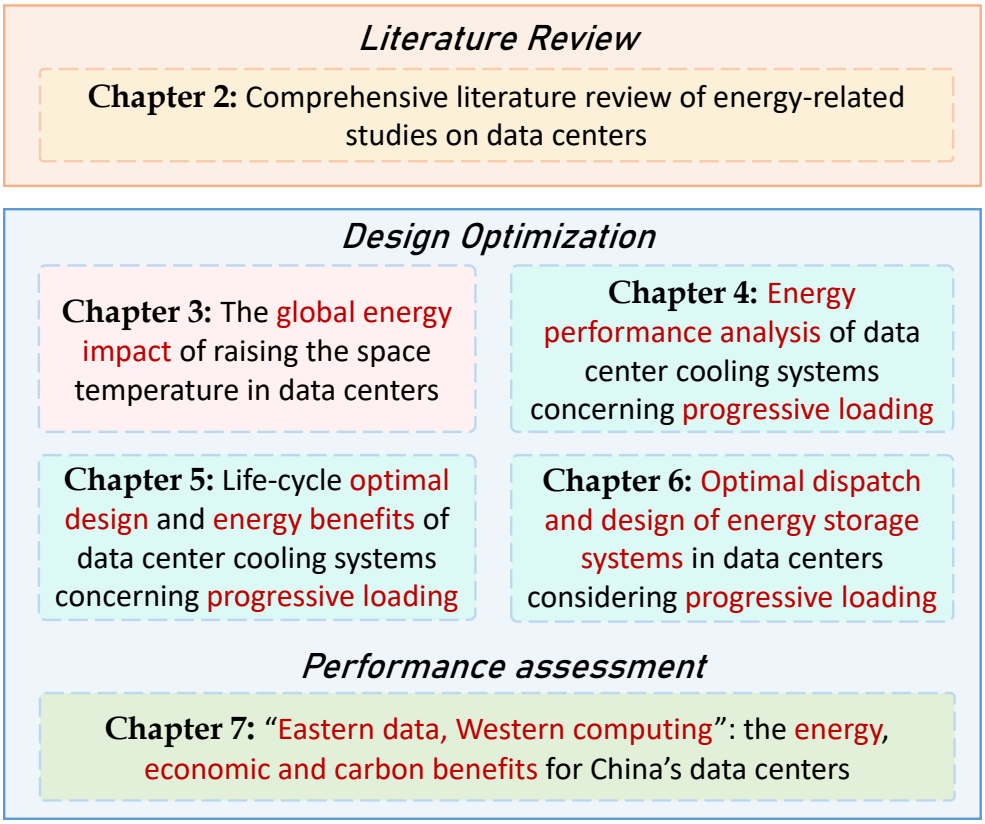


Fig. 1.1 Organization of main chapters

CHAPTER 2 LITERATURE REVIEW

Ever-increasing energy consumption in data centers has emerged as a growing global concern. Developing highly efficient cooling systems in data centers is of vital importance to the sustainability of the data center industry. High-temperature data centers are a fundamental solution to save a large amount of cooling energy by changing cooling mechanisms. 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. Furthermore, adopting life-cycle optimal designs for data center cooling systems considering progressive loading is also an effective means to largely reduce cooling energy. In addition, data centers can act as pivotal players in grid stability by providing flexible services. By utilizing the schedulable capacity of energy storage systems for emergencies in data centers to participate in flexible grid services, data centers can potentially reduce their energy costs and even generate revenue, creating a win-win situation.

This chapter provides a comprehensive overview of high-temperature data centers, emphasizing their key advantages and the primary challenges associated with their implementation, as well as the existing efforts and latest technologies to tackle the bottlenecks. The chapter also presents an overview of the technologies and strategies adopted to enhance energy efficiency in data center cooling systems. Additionally, the opportunities and main benefits for data centers to interact with the power grid and the trends and benefits of computing workload migration in geo-distributed data centers are also overviewed and presented.

2.1 Overview of high-temperature data centers

Cooling typically constitutes 30%~40% of the total power consumption in data centers [6, 7]. Scientists and engineers are attempting various means to enhance the efficiency of data center cooling systems. Currently, approaches associated with efficiency enhancement have been widely investigated, including airflow distribution optimization [22, 23], containment of aisles [24, 25], supply and demand match of cooling energy [26, 27], and performance improvement in servers [28]. Although efficiency enhancement is an effective means to reduce energy consumption, it can only save energy to a certain amount. As a fundamental solution, increasing the space temperature in data centers (the ambient temperature for servers) can dramatically reduce cooling energy demand by adopting free cooling.

In the earlier stages, data centers were typically operated in a temperature range of 20-21°C with a common notion of "cold is better"[29], and even some were as cold as 13°C [30]. Due

to the conservative suggestions by manufacturers and conventional wisdom, engineers and operators tend to follow conventional practices. Usually, mechanical cooling is needed to maintain a low space temperature in data centers, but it consumes a huge amount of energy. Thus, raising the space temperature in data centers is expected to reduce the mechanical cooling energy by shortening chiller operating hours. It has been reported that increasing the space temperature in data centers by 1K results in the saving of the total power consumption by 4~5% [31, 32].

The high-temperature data center is regarded increasingly as a trend in the data center profession and industry. In 2011, ASHRAE (American Society of Heating, Refrigeration, and Air Conditioning Engineers) further expanded the range of allowable environmental conditions of IT equipment and adopted two new classes in their updated guidelines (class A3 and class A4) based on the updated critical information (IT design and failure data) provided by IT manufacturers [14]. The new Class A3 environment has the upper temperature bound of 40°C and the Class A4 environment even has the upper temperature bound of 45°C, as shown in Fig. 2.1.

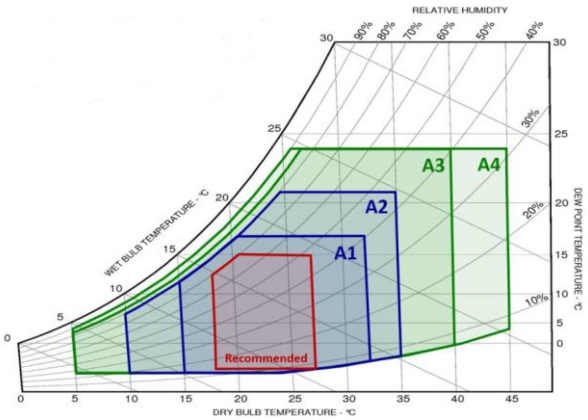


Fig. 2.1 Ambient temperature classification of IT equipment by ASHRAE [14]

In fact, if IT equipment could withstand long-term operation at the temperature of 45°C, the worldwide deployment of highly economized and even chiller-free data centers could be realized [33]. Some data center operators are considering operating their data centers at higher temperatures [34]. Beaty et al. [35, 36] reported comprehensive analysis and discussions on the workflow for raising the space temperature from the cooling perspective. Baidu reported that it is possible to design a kind of server that can support an ambient temperature of 50°C, according to the temperature specification of key components in servers [37]. Intel and Microsoft reported that most servers worked well under higher ambient temperatures and outside air [38]. A chiller-less economized data center test conducted by ASHRAE shows that

short-term high temperatures have little impact on data centers [29]. The data center utilized outdoor air to cool their computer rooms throughout the year and worked well. In addition, some researchers believe that temperature variations of outdoor air could be well regulated via proper control [39] and have little impact on server operations [40].

2.1.1 Main benefits and bottlenecks for high-temperature data centers

Fig. 2.2 summarizes the main benefits and concerns of high-temperature data centers today. It outlines four primary advantages and four main concerns associated with raising the space temperature in data centers. These main advantages and concerns are discussed and elaborated as follows.

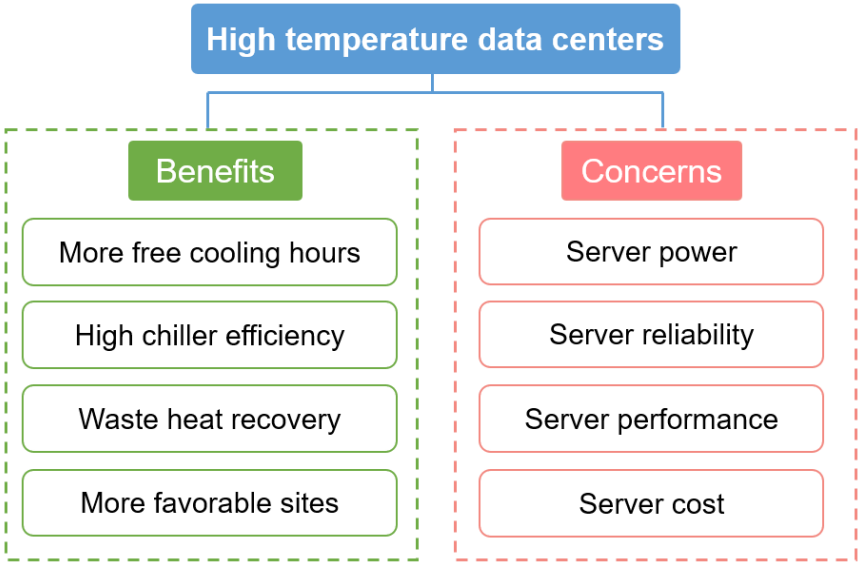
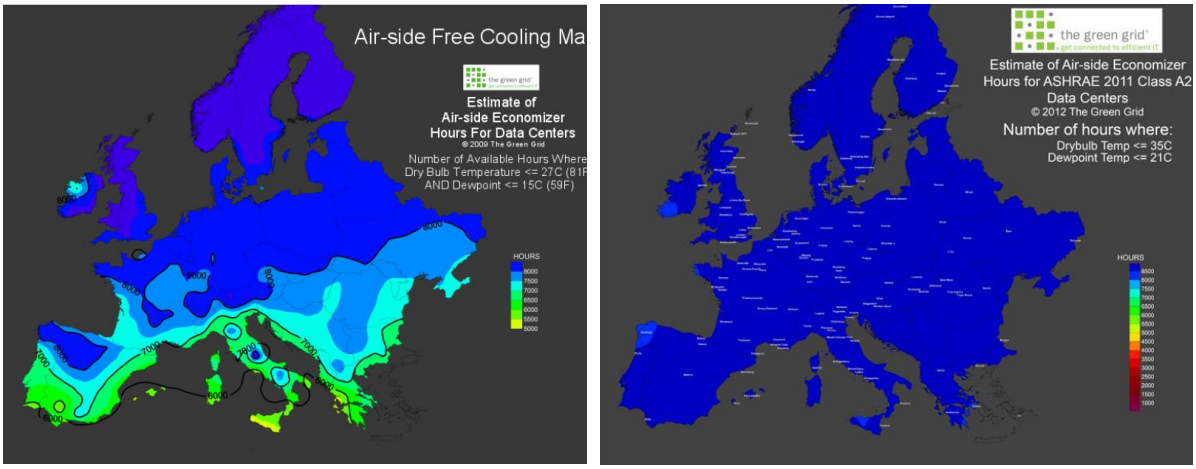


Fig. 2.2 Main benefits and concerns of high-temperature data centers

i. More free cooling hours and great energy-saving potential: Free cooling is widely employed in large data centers in mild and cold regions by utilizing outdoor cold air (when the outside temperature is lower than indoor space temperature in data centers) [41]. "Free cooling is the single biggest opportunity for efficiency, greater than all other options combined", according to the data center director of Google [33]. A study suggested that when the ambient temperature of servers in data centers rose from 27°C to 35°C, the airside free cooling hours in a year increased from 80% to 99 % in Europe, as shown in Fig. 2.3 [42]. Large Internet companies, like Google, Microsoft, and Intel, have been aggressive in operating their data centers at high temperatures (i.e., above 27°C). Microsoft reported that raising the space temperature in data centers by 2-4°C in one of their Silicon Valley data centers could save $250,000 in annual energy costs [30]. It was also estimated that the space temperature of 35°C, combined with containments, could lead to 22% energy savings [43].



Space temperature in data centers of 27°C *Space temperature in data centers of 35°C*

Fig. 2.3 Air-side free cooling map of Europe [42]

ii. Higher chiller efficiency: Higher space temperature in data centers allows for higher chilled water temperature. Every degree (K) increase in the chilled water temperature can bring about a 2% improvement in the chiller efficiency [44, 45].

iii. More favorable site options: Because of more free cooling hours, the high-latitude area in the Pan-Arctic region has become a hotspot for data center site selection [46, 47]. Google has deployed a number of its data centers in high latitudes near the North Pole to minimize cooling costs [48]. It was reported that 91~99% of data centers in North America, Europe and Japan could use free cooling throughout the year if Class A3 equipment was used (dry-bulb temperature range of Class A3: 5~40°C) [49]. Thus, raising the ambient temperature of servers could enable the deployment of highly economical and even chiller-less data centers worldwide.

iv. Improved waste heat recovery: The main barrier to recovering waste heat from data centers is that the heat collected, although plentiful, is of too low quality [50]. Increasing the space temperature in data centers makes it possible to recover more heat of higher quality or temperature. The heat can be used by the neighboring buildings or district heating systems.

Despite these advantages, it is observed that very few data centers have attempted high-temperature operations, as reported [51]. This is due to concerns, mainly system reliability, computing performance, power consumption, and cost issues related to high-temperature operation [52]. For example, it is usually considered that high temperatures will affect the reliability of servers, subsequently impacting overall system reliability. If these concerns are well addressed, the high-temperature environment will become a revolutionary turning point in the data center industry.

2.1.2 Temperature-sensitive IT components, implementation limitations and existing efforts

Servers are the core elements of a data center. There are four main limitations concerning servers when implementing high-temperature data centers, as depicted in Fig. 2.4, including *i.* the impact on server reliability; *ii.* the impact on server performance; *iii.* the impact on system power consumption; *iv*. the tradeoff between server cost premium and operating cost savings.

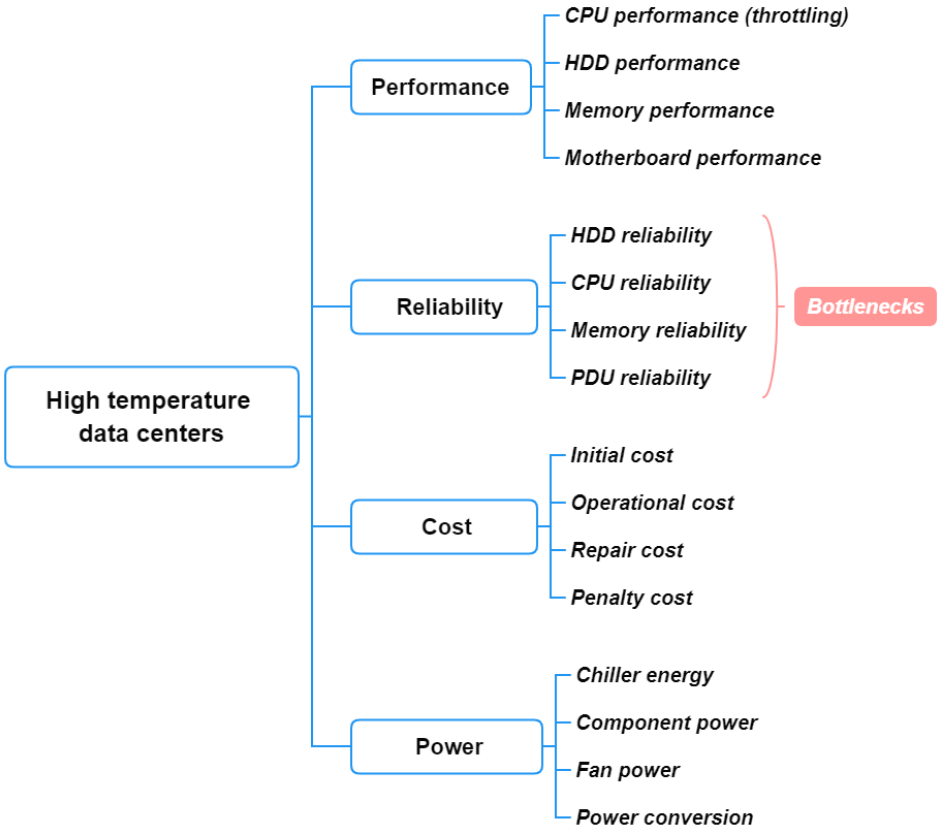


Fig. 2.4 Key challenges to the implementation of high-temperature data centers

Server reliability vs high-temperature

The reliability of servers is the main bottleneck for high-temperature data centers. It is commonly understood that a high-temperature environment would reduce the reliability and availability of servers. In this regard, the latest Thermal Guidelines for Data Processing Environments [14], published by ASHRAE, reported that the relative failure rate of volume servers is a function of server inlet temperature (the ambient temperature of servers) under continuous operating conditions (7×24×365), as shown in Table 2.1. Here, the 'X-factor' denotes the relative failure rate compared to the baseline of a server inlet temperature of 20°C (i.e. an X-Factor is 1.00 at a server inlet temperature of 20°C). It is noticeable that the average failure rate increases as the server inlet temperature rises.

Table 2.1 The relative failure rate of volume servers as a function of server inlet temperature [14]

Dry Bulb Temperature (°C)	X-Factor of Average Failure Rate	Lower Bound of X-Factor	Upper Bound of X-Factor
15	0.72	0.72	0.72
17.5	0.80	0.87	0.95
20	0.88	1.00	1.14
22.5	0.96	1.13	1.31
25	1.04	1.24	1.43
27.5	1.12	1.34	1.54
30	1.19	1.42	1.63
32.5	1.27	1.48	1.69
35	1.35	1.55	1.74
37.5	1.43	1.61	1.78
40	1.51	1.66	1.81
42.5	1.59	1.71	1.83
45	1.67	1.76	1.84

In reality, there is a specific high-temperature period (excursion time) for each location, mainly during hot summers. Thus, the impact on server reliability is much more benign than under continuous high-temperature operating conditions (7×24×365), even in the harshest scenario (chiller-free operation to achieve free cooling throughout the year).

Additionally, the guidelines pointed out that the reliability data was intended to capture most of the volume server market (belonging to Class A2, rated to 35°C [33], and very few products currently exist for the Class A3 environment [29]). However, some specific server information was not disclosed. For example, there is no definite data on the failure rate of servers that could function under high temperatures with improved heat sinks and advanced materials.

The impacts of high temperature on system reliability are further analyzed below from two aspects: the impacts of high-temperature excursion on data centers, and the temperature-sensitive components in servers.

High-temperature excursion

High-temperature “excursion time” is a period that requires special attention for the operation of high-temperature data centers. This has been studied by some IT companies. Google tested the temperature limits of its hardware by running servers at higher temperatures. The results showed that the servers operated just fine without any increase in failures [53]. A Google data center in Dublin was optimized to use fresh air to cool tens of thousands of servers throughout the year, without the need for air-conditioners or chillers. Despite annual high temperature and temperature fluctuations, the servers functioned well. It indicates that servers are much sturdier than previously imagined and can withstand high-temperature excursions for a certain time

[54]. Intel conducted a comparative study over 10 months using about 900 production blade servers [55]. The test room was divided equally into two side-by-side compartments: one used conventional air-conditioners, while the other used air economizers to cool down servers with 100% outside air at temperatures of up to 33°C. Servers in the economizer compartment were subjected to considerable variations in temperature and humidity, as well as poor air quality. The failure rates were 3.83% and 4.46% for the conventional air-conditioned compartment and the air economizer compartment, respectively, showing insignificant differences [55].

Temperature-sensitive components

To further understand the reliability of servers, it is important to investigate the reliability of individual server components, especially those that are prone to failures. Table 2.2 shows the detailed temperature limitations of server components, referring to Intel's Romley Sandy Bridge-EN platform design guide and component specification [56]. It is evident that the component with the lowest temperature limit is hard disk drives (HDDs).

Table 2.2 Temperature limits of components in typical servers [56]

Key Components	Temperature in SPEC (°C)
HDD (Hard Disk Drive)	60
SAS (Serial Attached Small Computer System Interface)	70
Crystal Oscillator	70
Clock Generator	70
BMC (Burst Mode Controller)	70
EEPROM (Electrically Erasable Programmable read-only memory)	85
CPU@95W TDTS	89
PCH (Paging Indicator Channel)	92.7
MEM@DDR3 (Memory Device)	95
PCB (Printed Circuit Board)	105
NIC (Network Interface Card)	123
Regulator	125
Ferrite Bead	125
Diode/Triode/Transistor	125

A few investigations also indicate that hard disk drives (HDDs) and memory units are the most frequently replaced components in modern data centers, with typically 1%-5% of drives needing replacement annually [57, 58]. Microsoft Corporation observed and recorded actual data on different kinds of failure types over two years from their typical large-scale data centers, which house more than 100,000 servers, as shown in Fig. 2.5 [59]. HDDs accounted for 71% of the known failures, making them the most dominant failing part.

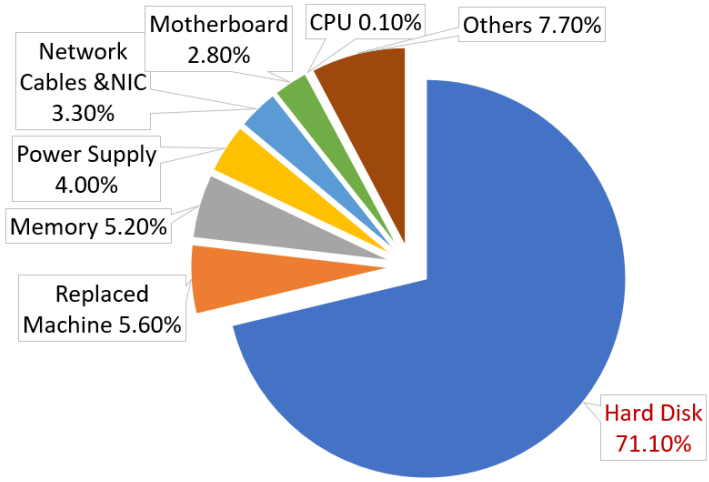


Fig. 2.5 Breakdown of hardware component errors in a large data center - Microsoft

Similarly, Baidu Corporation recorded statistical data on storage servers operating in their data centers between the period from April to June 2012, as illustrated in Fig. 2.6. The failure rate of HDDs was the highest, accounting for approximately 98% of all system failures, followed by memory units and motherboards. The failure rates of other system components are close to zero.

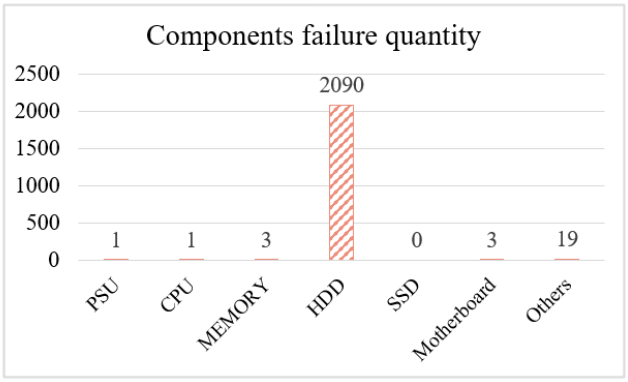


Fig. 2.6 Breakdown of hardware component errors in a data center - Baidu

Hard disk drives (HDDs)

Given that HDDs are the most critical components, studies have been conducted to investigate the relationship between temperature and HDD failure rates. A specific test by Microsoft Corporation revealed that an increase in server inlet temperature had a certain impact on HDD failure rates, but redesigning the layout of server components could maintain a very low level of HDD failure rate, even when the inlet temperature reached up to 40°C [40]. A recent study by Google presented a very interesting result showing that very low temperatures were actually more detrimental to disk reliability than higher temperatures [32]. Pinheiro et al. [60] observed

a rapid drop in disk failure rates with increasing temperatures, followed by a slight increase when the temperature exceeded 45°C, as shown in Fig. 2.7.

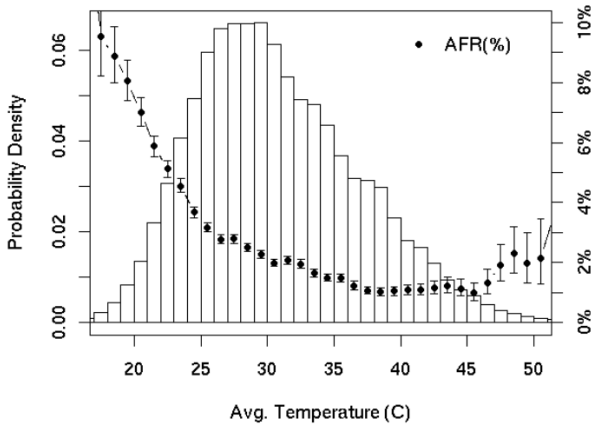


Fig. 2.7 Distribution of average temperatures and failure rates of HDDs [60]

Nosayba et al. [32] collected data on HDD replacements from January 2007 to May 2009 across 19 different Google data centers, covering 5 different HDD models. The failure rates of two models increased significantly when the temperature rose from 40°C to 55°C, while the failure rates of the other three models increased slightly, as shown in Fig. 2.8. This suggests that different models or types of HDDs respond differently to high-temperature operation. A study by Alibaba group reported that maintaining HDD inlet temperature between 50 and 55°C was a reasonable range to achieve acceptable reliability levels in high-temperature data centers [52].

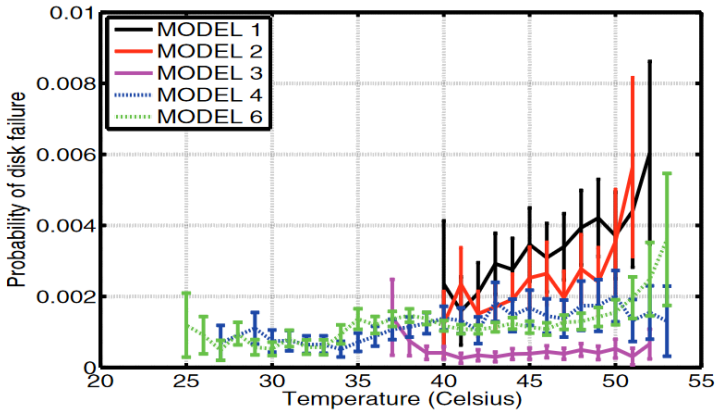


Fig. 2.8 The monthly probability of HDD failure as a function of temperature [32]

The failures of HDDs are likely caused by the vibration due to the rotation of server fans. It has been reported that HDD performance degradation is proportional to the speed of the server fans, and an unoptimized fan control algorithm could reduce HDD performance by up to 60% [61]. A study on solid state drives (SSDs) shows SSDs have better reliability compared to HDDs [62], but the price of SSDs is three to five times that of HDDs. In this context, Chan

et.al proposed a hybrid SSD-HDD system and a novel fan control method to minimize the impact of vibration [63]. Zhang et al. [52] also suggested that replacing HDDs with SSDs in high-temperature data centers could reduce the failure rate of servers.

Memories

High temperatures are considered to negatively impact the reliability of hardware components in memory due to significant physical changes in materials. Some studies show that high-temperature operation is expected to increase leakage current in memory chips and lead to a higher likelihood of flipped bits in the memory array [64, 65]. However, Schroeder et al. [66] reported that temperature had a surprisingly low effect on memory errors.

Table 2.3 shows the results of a comprehensive study conducted by Baidu Corporation concerning the effect of high temperatures on all key components of servers [37]. It can be observed that when raising the ambient temperature to 50°C, the failure rates of both motherboards and memory units doubled. However, an ambient temperature range of 35°C to 40°C is considered acceptable for high-temperature data centers. Overall, while high temperatures do affect the reliability of some servers, some of the latest IT products demonstrate the feasibility of operating high-temperature data centers around 35°C or even 40°C [33].

Table 2.3 Evaluated Components AFR (Average Failure Rate) [37]

Ambient temperature	35°C or 40°C	50°C
Components AFR	AFR (%)	AFR (%)
Motherboard	2.62%	5.06%
Memory	0.32%	0.70%
HDD	0.73%	0.92%
FAN	1.79%	2.50%
Power Supply	2.92%	2.92%

Server performance vs high-temperature

Server performance mainly refers to the performance of the central processing unit (CPU), memory and hard disk drive. It is generally believed that component performance degradation, such as frequency reductions for CPUs and memory, is associated with higher ambient temperatures. However, some studies indicate that this is not always the case.

IBM conducted a particularly well-defined and controlled study focused on server performance within the Class A3 environment. They selected nearly 70 different CPUs as test samples, with a server inlet temperature baseline of 25°C for tests of each piece of equipment, and then repeated the tests at 35°C (the upper limit for Class A2) and 40°C (the upper limit of class A3).

As summarized in Table 2.4, the performance at 35°C and 40°C was almost the same as at 25°C, with a maximum reduction of 2% in some cases. The results indicate that there was no significant performance degradation at higher temperatures up to a certain extent [34]. Similarly, Wang also studied the impact of CPU temperature on server performance and concluded that CPU temperature has no obvious impact on server performance [67].

Table 2.4 Server performance at different inlet temperatures [34]

Test conditions			**Server inlet temperature**		
System	**Model**	**Exerciser**	**25°C**	**35°C**	**40°C**
X3550M4	130W	Linpack Turbo-on	1.00	1.00	1.00
	115W	Linpack Turbo-on	1.00	1.00	1.00
X3650M4	115W Best	Linpack Turbo-on	1.00	1.00	1.00
	115W Best	SPECjbb2005	1.00	1.00	0.99
	115W Best	SPECint_rate2006	1.00	1.00	1.00
	115W Best	SPECfp_rate2006	1.00	1.00	1.00
X3650M4	130W	Linpack Turbo-On	1.00	1.00	1.00
X240	130W Best	Linpack Turbo-On	1.00	0.99	0.99
	130W Best	SPECjbb2005	1.00	0.99	0.99
	130W Best	SPECint_rate2006	1.00	1.00	1.00
	130W Best	SPECfp_rate2006	1.00	1.00	1.00
	130W Worst	Linpack Turbo-On	1.00	0.99	0.98
	130W Worst	Linpack Turbo-Off	1.00	1.00	1.00
	130W Worst	SPECjbb2005	1.00	0.99	0.99
	130W Worst	SPECint_rate2006	1.00	1.00	0.99
	130W Worst	SPECfp_rate2006	1.00	1.00	0.99
dx360M4	115W Worst	Linpack Turbo-On	1.00	1.00	1.00
	115W Worst	SPECjbb2005	1.00	1.00	1.01
	115W Worst	SPECint_rate2006	1.00	1.00	1.00
	115W Worst	SPECfp_rate2006	1.00	1.00	1.00
AVERAGE			**1.00**	**1.00**	**1.00**

Researchers at the University of Toronto tested the performance of HDDs, CPUs, and memory across a wider range of workloads and temperatures [32]. The test temperatures were much higher than those in the IBM tests, making it easier to identify performance degradation if it fell outside the range of normal statistical error. The tests were conducted inside a thermal chamber where temperature could be controlled in 0.1K increments from -10°C to 60°C – a range much wider than the typical temperature range in data centers today. The results showed

that there was no throttling down in CPU and memory performance on any of the benchmarks up to 55°C. As for the hard drives, they observed a decrease in “throughput” at an ambient temperature of 40°C for the Seagate 73GB and Hitachi SAS drives, at 45°C for the Fujitsu and Seagate 500GB SAS drives, and at 55°C for the Hitachi Deskstar. This indicates that different types of products have different responses to high-temperature operation.

Alibaba Group also studied the performance of key components operating at different ambient temperatures, as shown in Fig. 2.9. Here, the term ‘thermal margin’ refers to the additional operating margin available after meeting safety limit requirements. As observed, all key components, except for HDDs during long-term operation, maintained sufficient thermal margin even at a server inlet temperature of 40°C [52]. These results are consistent with those reported by the University of Toronto.

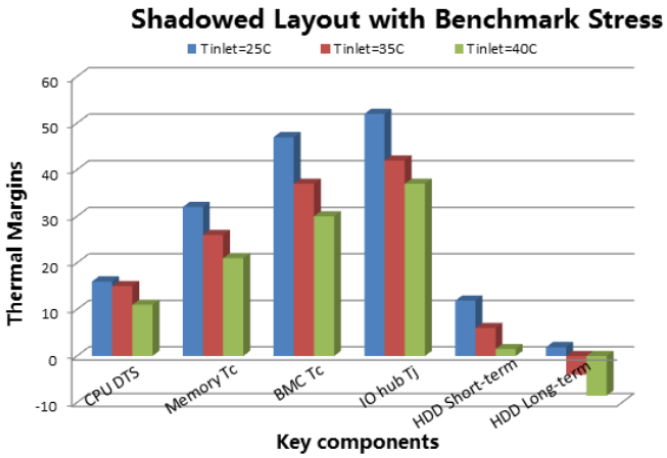


Fig. 2.9 Thermal margin of key components in servers [52]

Baidu Corporation conducted one-month experiments to explore the relationship between the performance of key server components and ambient temperatures. They discovered that when the ambient temperature increased from 25°C to 50°C, the performance of the CPU, memory, and HDD remained stable. However, when the HDD operated at temperatures ranging from 54 to 57°C, its performance declined by approximately 30%. When the temperature reached 57°C, the performance dropped dramatically, by an average of about 40% to 80% [37].

Based on these test results, it can be concluded that the HDD is a key limitation for high-temperature operations. It seems that different types of HDDs respond differently to high-temperatures. Therefore, HDDs capable of operating over a wider or higher temperature range are necessary for high-temperature data centers.

Power consumption vs high-temperature

Server power

As the ambient temperature of servers increases, the energy consumption of server fans also increases to maintain all components below certain critical thermal thresholds. It has been reported that server fans contribute up to 14% of the total power consumption in data centers [68]. Additionally, the power consumption of servers slightly increases as the inlet temperature rises, due to increased leakage current in some silicon devices [67].

ASHRAE studied the relationship between server power and server inlet temperature, as shown in Fig. 2.10 [29]. For Class A2 servers, the power consumption could increase by 4% to 8% when the inlet temperature rises from 15°C to 30°C. The increase could be between 7% to 20% when the inlet temperature rises from 15°C to 35°C. The power increase in Class A3 servers could be similar to that of Class A2 servers, although Class A3 servers would likely require improved heat sinks and/or fans to adequately cool the components.

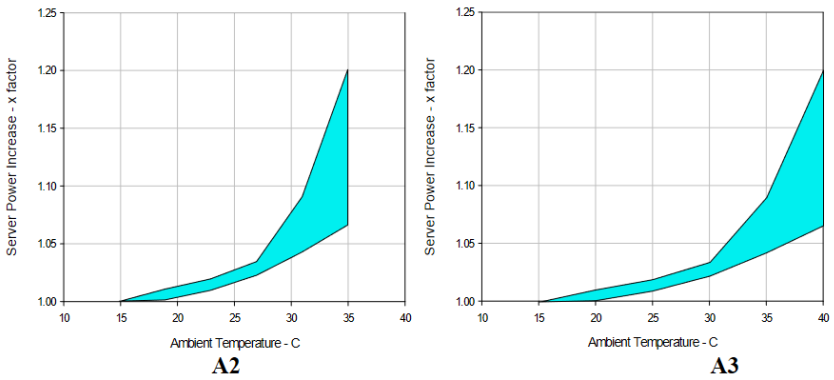


Fig. 2.10 Server power rise vs Ambient temperature range [29]

Fan power rise

In the latest thermal guidelines for data processing environments from ASHRAE [14], the relationship between server airflow rate and ambient temperature was tested, as shown in Fig. 2.11. The server airflow rate increases as the ambient temperature rises. The blue area primarily results from different types of servers configured with various fan control algorithms. The guidelines also list two typical fan control algorithms: fixed fan speed control and variable fan speed control, as illustrated in Fig. 2.12. In most cases, servers equipped with variable fan speed control show better energy-saving potential.

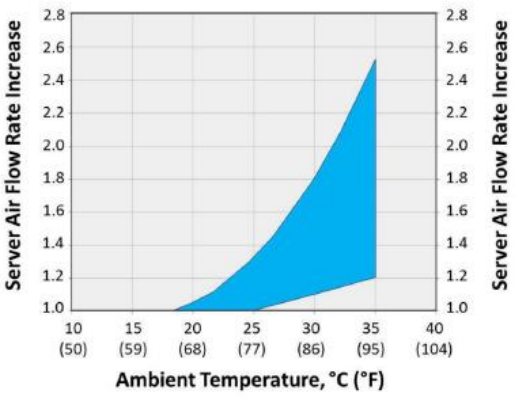
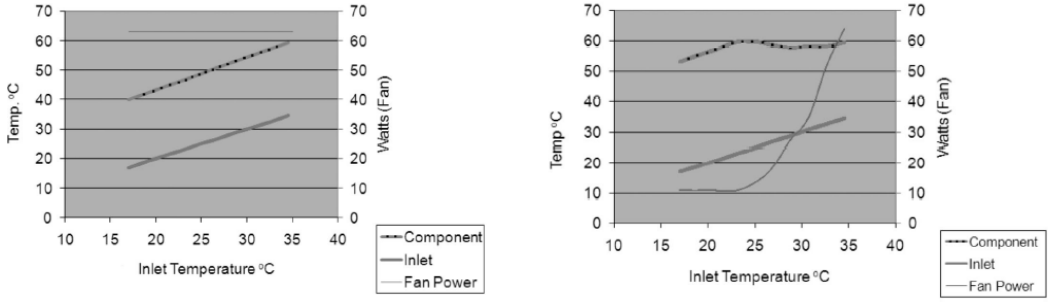


Fig. 2.11 Server flow rate increase versus ambient temperature increase



Fixed fan speed Variable fan speed

Fig. 2.12 Server fan behavior versus inlet temperatures [14]

Schneider [69] divided IT server fan behavior into three categories based on server configurations, including: i) servers configured with a high-stress load, ii) servers configured with a moderate load, and iii) servers configured with a light load. They concluded that servers configured with a light load were usually equipped with the constant fan speed control algorithm, while the other two types were often equipped with the variable fan speed control algorithm. According to the Fan Affinity Laws, shown in Eq. (2.1), fan power increases cubically with the rotational speed [70]. Thus, fan power will increase significantly when the server detects an increase in ambient temperature.

$$\text{fan power} \propto (\text{fan speed})^3 \tag{2.1}$$

Leakage power

Patterson et al. [71] reported that the temperature has little effect on the dynamic power of CPUs but significantly influences leakage. Leakage power is primarily induced by leakage current in some silicon devices. In the past, it did not draw much attention since the leakage current was an order of magnitude less than the chip's normal operating current [72]. In later generations of CPUs (e.g., 0.35-0.18 micron technologies), leakage remained a small fraction of the total power. However, as semiconductor technology has recently advanced into the

nanometer era, the continued shrinking geometries in silicon have caused the leakage to increase by as much as 50%, due to shorter leakage paths [73-75]. Fallah et al. [76] reported that every 1K increase across multiple process generations causes a 2% increase in leakage. As technology progresses to 45 nm and beyond, these temperature-dependent leakage rates are expected to rise further. Thus, leakage power needs to be considered when raising the ambient temperature.

Fig. 2.13 shows the effect of ambient temperature on the power of key server components. As the ambient temperature increases, fan power substantially increases, while CPU and memory powers increase slightly, and the HDD power decreases slightly [56]. It can be concluded that fan power is the main contributor to the rise in server power when the ambient temperature increases.

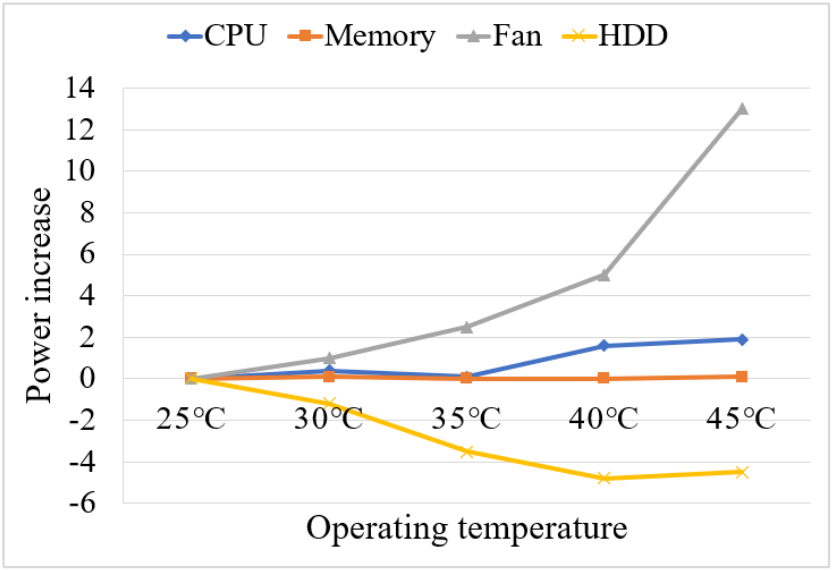


Fig. 2.13 Key Component Power versus temperature

Total power consumption

Generally, the power consumption of chiller plants will decrease when higher temperatures are set in computer rooms due to the higher chilled water temperature supplied. Reduced chiller energy consumption results from increased chiller efficiency due to a higher chilled water supply temperature, a reduced cooling load due to higher space temperatures, and less chiller operation time due to increased free cooling time.

However, server fan power and computer room air handler (CRAH) fan power may increase. The CRAH must supply more cooling air to servers to match the increased airflow from server fans. Thus, server power and CRAH fan power may offset the power saving in chiller plants.

Moss et al. [69] found that when the server inlet temperature was between 25-27 ℃, the total energy consumption was the lowest for a chilled water system. However, their study did not consider the free cooling and advanced servers, which can withstand higher temperatures with low fan power. Seaton [77] studied server power and cooling energy consumption at different server inlet temperatures without considering free cooling. The results also show that server inlet temperatures within 25-27℃ had the lowest total energy consumption, in line with David's results. Similarly, Muroya et al. [78] found that the total power consumption of data centers reached a minimum when the server inlet temperature was around 25℃, as shown in Fig. 2.14.

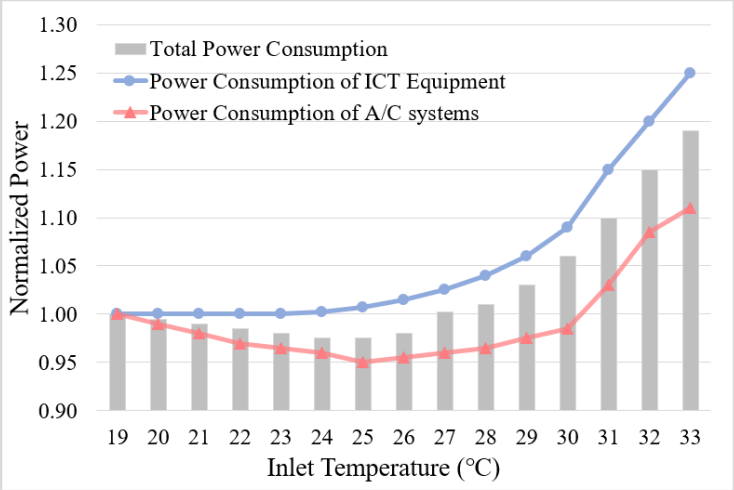


Fig. 2.14 Total power estimation in data centers

The above studies considered the facility's energy consumption when adopting mechanical cooling only. In practical applications, free cooling is often adopted to save more cooling energy. Seaton [77] further estimated the cooling costs of increased inlet temperatures while considering air-side free cooling and found that total cooling energy consumption would decrease further when the server inlet temperature was raised to 40°C.

System cost vs server inlet temperature

Server cost

Currently, servers are categorized into four classes: A1, A2, A3, and A4. The newer server classes (A3 and A4) that support wider environmental envelopes are generally more expensive and less commonly used. Class A1 servers, found in low-price markets and older data centers built long ago, have an upper-temperature limit of 30°C [79]. Class A2 servers, which are the standard base products, have an upper-temperature limit of 35°C. Class A3 servers, which are becoming the new standard, cost a premium over class A2 servers and have an upper-

temperature limit of 40°C. Class A4 servers, which are even more expensive compared to class A3 servers, have an upper-temperature limit of 45°C. High-temperature data centers require class A3 and class A4 servers due to their wider temperature limits.

For high-temperature data centers, the cost premiums for higher-class servers, which are usually equipped with optimized heat removal mechanisms and more robust components cannot be overlooked. Features such as improved heat sink designs and advanced materials are available but come at a high cost [29]. It is estimated that adopting cooling improvements to maintain the server performance will incur a 1%−2% premium for class A3 servers over class A2 servers, and a 5%−10% premium for class A4 servers over class A2 servers. Additionally, if the cooling system cannot be enhanced due to server volume constraints, many server designs may require non-cooling component improvements, such as the adoption of advanced materials, to achieve class A3 or class A4 operation. In these cases, there would be a cost premium of 10% to 15% over a class A2 server [29].

Total life-cycle cost

Although the capital investment in high-temperature data centers will increase due to the premiums of higher-class servers, the operating cost will decrease due to more free cooling hours. The increase in capital cost is primarily due to server premiums, while the reduction in capital cost mainly involves investments in chiller plants, particularly for chiller-less plants. The reduction in operating cost is primarily associated with increased chiller efficiency and reduced chiller operating hours.

Baidu Corporation conducted a total cost of ownership analysis, considering cooling operating costs and maintenance costs [56]. The results show that the cooling system investment cost would be greatly reduced and the overall cost-benefit would increase by 7.48% if the chillers were not used and air-side free cooling mode was employed. Rubenstein et al. [80] investigated the operating cost of data centers when increasing the space temperature in computer rooms, and concluded that while increasing temperature indeed reduced chiller power consumption, the overall cost savings may not be realized due to increased server power and decreased reliability.

These studies analyzed the correlation between the operating cost and the space temperature in data centers without considering the premiums for higher-level servers, the capital cost of chiller-free cooling plants, and the operating cost savings due to more free cooling time. It can

be observed that current research lacks a detailed economic analysis of high-temperature data centers.

2.1.3 Current technologies, feasible solutions and future perspectives

This section summarizes current technologies and feasible solutions for implementing high-temperature data centers at the room, rack level, server and chip levels. It introduces and elaborates on the available optimization, improvement and innovation methods at each level, as shown in Fig. 2.15. Additionally, this section presents and discusses the current state and future perspectives of data centers.

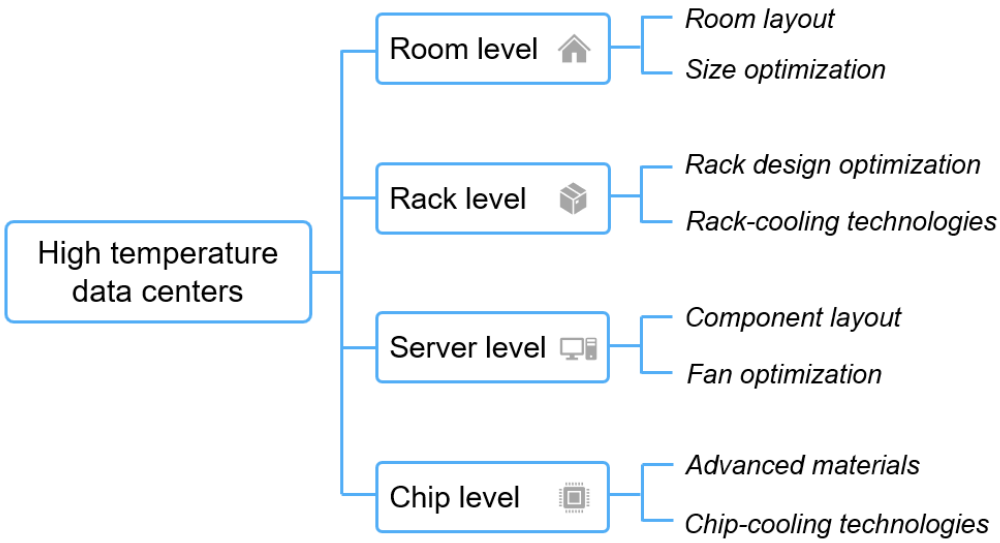


Fig. 2.15 Main feasible solutions at different levels for implementing high-temperature data centers

Room level

Optimizing airflow management is an effective way to enhance cooling efficiency in data centers, which can reduce hot air recirculation and cold air bypass. Most data centers have hot spot problems in that the temperature of local areas is significantly hotter than the average temperature in the computer rooms. The temperature of the hot spots could be 8-10K higher than the average temperature in the computer rooms [32]. Therefore, more cooling energy and lower supply air temperature are needed for the reliable operation of all servers, due to the "shortboard effect".

Good airflow management allows servers to operate at a higher ambient temperature. With good airflow management, 24°C cold air from computer room air handlers can result in a maximum server inlet temperature of 25-26°C somewhere in the data center (without hot spots). In contrast, with very poor airflow management, 13°C cold air could easily result in server inlet

temperatures ranging anywhere from 25°C up to over 32°C due to hot spots. Thus, the improvement methods at room level primarily aim to optimize airflow management and eliminate hot spots.

Containment strategy

The aisle containment strategy, including cold aisle containment and hot aisle containment, is an effective means to improve airflow distribution in data centers and has been widely used [81]. Containment can reduce the mixing effects of hot and cold airstreams that would cause unwanted temperature rises at rack inlets. Fig. 2.16 shows a typical hot-aisle/cold-aisle configuration. The cold air moves from raised floors to the room through perforated tiles, is inspired by racks on the front side, gains the heat created by racks, and is expelled at the rear of each rack. The hot air is then aspired by the computer room air conditioning units and is rejected to the exterior environment.

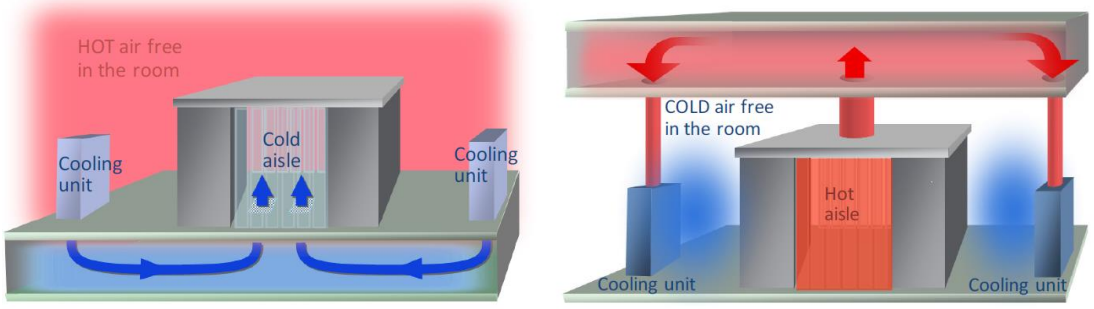


Fig. 2.16 Hot-aisle/cold-aisle configuration [82]

An experiment in a data center with a floor area of 52 m^2 shows that cold aisle containment could dramatically remove hot spots at the rack inlet compared with conventional open aisle configuration [83]. With cold aisle containment, the rack inlet temperature could be reduced by up to 40% without changing the room layout [84]. Other studies have found that the supplied air temperature could be raised by 3K [85], or increased from 18°C to 22°C by using aisle containment [86]. It is also found that hot-aisle containment could save 43% in annual cooling system energy costs and a 15% reduction in annualized PUE, compared with cold-aisle containment [82].

Room layout

Room layout mainly involves room shape optimizing, rack layout, and the location of computer room air handlers (CRAHs) [87]. A study found that locating CRAH units perpendicular to rack rows had better cooling efficiency than positioning them in line with rack rows [88]. Schmidt et al. [89] studied overhead/raised-floor air supply systems, and found that raised-

floor design could achieve a higher temperature of supply air. Another study also found that the raised-floor air supply with overhead return system had the best cooling performance among four different systems, since this system could use the lowest supply air flow rate to achieve the same average temperature at the same heat density and supply air temperature [90].

Structural size optimization

Structural size optimization mainly includes the optimizing plenum size, ceiling height and the shape of the perforated tile. An optimized structure allows computer room air handlers to supply higher-temperature cold air to computer rooms.

A study shows that optimizing the plenum depth, ceiling height, and cold aisle location can achieve a 12K decline in server inlet temperature [91]. Lu et al. [92] examined the effects of geometry configurations on the thermal performance of data centers and found that plenum with gradient cross-sections, created by inclined partitions, could reduce the inlet and outlet air temperatures by 1.7−2.9K. The depth of the raised floor and areas free of perforated tile also affect rack inlet temperatures. The results show that even a very slight variation of open tile areas (less than 8%) can lead to some rack inlet air temperatures seeing a reduction of as much as 10K [93]. An open area of perforated tiles affects flow uniformity alongside the tiles. A 25% open area of perforated tiles is recommended to maximize perforated tile airflow uniformity and temperature distribution uniformity [94]. Tile orientation also affects air flow distribution. The sensitivity in bulk flow rate and variations in flow distribution of different tile orientations have been studied [95]. The perforated tile geometry also has considerable impacts on the airflow characteristics in an aisle [96].

Rack level

At the rack level, the thermal performance is affected by the rack design [97] and the server placements within racks [98]. Rack-level modifications and optimization could improve heat dissipation by up to 50% and decrease temperature variability by 60% under the same cooling infrastructure [97, 98]. In addition, the innovative technologies applied in rack-level cooling are an effective way to enhance cooling efficiency.

Rack design and optimization

The server configuration has been found to have a strong correlation to failure rates when raising the inlet temperature. The correlation between failures and the location of servers within a rack should be considered when setting the inlet temperature of servers [40]. A study on server rack design proposed to decouple the fans and power supply units from the node and

place all fans on the rear side of the rack. A new rack design lowered the power consumption by 10% while allowing key components to keep enough margin to maintain reasonable reliability levels, even under an ambient temperature of 40°C [52]. Another similar design achieved a power reduction of 66.7% compared with general racks under the same test conditions [99].

It was also found that the heat generated by the IT equipment could be efficiently discharged with higher porosity. By increasing the porosity ratio of rack doors from 25% to 50%, the thermal environment inside the rack could be improved significantly [100].

The placement of servers affects rack thermal performance significantly. Wang et al. [101] proposed a drawer-type rack design with more hot aisle space and less cold aisle space. The design could reduce hot air recirculation and cold air bypass, and decrease the rack maximum inlet temperature by up to 13.3K. Chu et al. studied the effect of server layout and heat exchanger layout on airflow uniformity in a vertical direction [22]. Results show that the placement of servers and heat exchangers has a great impact on the airflow distribution at the server inlet. To obtain a reasonable server arrangement, a similar test was conducted to discover the relationship between servers' placements and airflow distribution. It was recommended to uniformly mount the servers from the middle of the rack to the upper for low load rate cases [102].

Innovative rack-cooling technologies

Yu et al. [103] applied the solid sorption heat pipe coupled with direct air cooling technology for rack-level cooling in data centers and found that this cooling technology could reduce the peak temperature of servers from 75.8 °C to 68.8 °C. Such a reduction of server peak temperature means that higher space temperature could be allowed in data centers. Li et al. [104] proposed a multi-split backplane cooling system at the rack level. A reliable operation control strategy could save a maximum of 12% of energy. Tian et al. [105] designed an internally cooled rack with a two-stage heat pipe. Their test results show that the new cooling solution essentially eliminates undesired air mixing and hot spots, and reduces annual cooling energy consumption by about 13%.

Server level

Today's data center cooling systems have evolved from traditional building HVAC systems. The primary difference lies in the serviced objects of the data center cooling systems primarily serving servers, whereas traditional cooling systems in offices and residential buildings

primarily serve people. The comfortable temperature range of the human body is narrow, but the server is man-made, and can be improved through structural and materials modifications. Redesigning servers is a feasible and effective means of implementing for high-temperature data centers.

A server chassis consists of sheet metal casings, CPUs, motherboards, power supplies, fans, memories, HDDs, and cables connecting these components [106]. The layout of servers and the positioning of each component within chassis are optimized for overall efficiency and cost-effectiveness.

Redesigning the server layout is an effective way to increase heat dissipation efficiency. A numerical study on a hybrid air/liquid-cooled server shows that basic server layout optimizations, such as changing the memory module angles and spacing, could enhance both the cooling effectiveness and the potential for waste heat recovery from the air stream. Such optimization could also decrease entropy generation by 15% [107].

Sarma et al. [108] conducted a CFD simulation on the thermal design of a 2-rack unit high computing server, summarized the existing problems, and provided reasonable thermal design recommendations. Hybrid cooling systems are proposed for servers. These systems combine air cooling for low-power components (e.g., power supplies, storage disk drives, and printed circuit boards) and water cooling for higher-power components (e.g., microprocessors and dual inline memory modules (DIMM)). This kind of server could accept water at a temperature of up to 45°C and air at a temperature of up to 50°C into the node, and saves 25% energy consumption in total [109].

To investigate the impact of the internal design of servers on their performance and cooling efficiency, servers were tested in a chimney exhaust rack. Internal recirculation due to unreasonable design could cause a hot air leakage from the server into the cold aisle, and subsequently affect the performance of adjacent servers. Due to the leakage, the power consumption of adjacent servers increased by around 20% [110].

Chip level

Chip-level cooling objects usually include processors, hard disks, and memories [111]. Generally, the schemes at the chip level could be divided into two categories: advanced chip materials and chip cooling.

Chip materials

The third-generation semiconductors with a wide bandgap, such as GaN/SiC equipment are considered to be of high performance, which further improves energy efficiency [46]. The wider bandgap allows the material to operate at a higher temperature, stronger voltage, and faster switching frequency [112]. When applied in data centers, such advanced silicon material will significantly improve energy efficiency in the data center industry.

Innovative chip-cooling technologies

Chip cooling enhancement technologies mainly include high-conductivity thermal interface materials [113, 114] and innovative chip cooling technologies [115]. Chip cooling enhancement technologies could effectively increase the heat dissipation of heat sources (mainly processors (60-75W each), and DIMMs (6W each) [50]).

Nanofluids, such as TiO_2 nanofluid and SiC nanofluid, are found to have a high-efficiency cooling potential [116, 117]. They could reduce CPU surface temperature by about 3.3K (8.2%) compared with deionized water using a cylindrical grooves heat sink [118]. Choi et al. [119] designed a new CPU cooler based on an active cooling heat sink combined with heat pipes. The total fan power consumption could be effectively reduced by 66.2% when using a heat pipe-embedded heat sink to replace the conventional heat sink [120].

Chip-scale refrigeration technologies are viable for thermal-limited applications, such as thermo-electric coolers [121] and small-scale refrigeration cycles [122]. Deng et al. [123] studied a two-stage multichannel liquid-metal cooling system for the thermal management of a high-heat-flux-density chip array. The experimental results show that the proposed liquid-metal cooling system can accommodate a heat flux of 50-200 W/cm^2 with a convective heat transfer coefficient exceeding 20,000 W/(m^2·K). Liang et al. [124] compared the heat pipe, thermoelectric system and vapor compression refrigeration for electronics cooling. They concluded: *i.* heat pipe is more attractive for cooling large devices at higher temperatures, *ii.* two-stage TE system could be used for cooling devices at lower temperatures, and *iii.* VCR system was capable of dissipating much higher heat flux (200 W/cm^2) at lower temperatures than all other technologies [124]. Chowdhury [125] integrated the thermo-electric coolers fabricated from nanostructured Bi_2Te_3-based thin-film superlattices into state-of-the-art electronic packages. Naphon et al. [126] studied HDD cooling with a vapor chamber, and found that the average HDD temperature with the vapor chamber cooling system was 15.21% lower than those without the vapor chamber cooling system.

Current state and future perspectives

At present, the space temperature in data centers is still very conservative, particularly for those used for rental purposes. It is observed that very few data centers tried the high-temperature operation due to the concern about reliability, although the allowing temperature limit has been increased by some professional associations, such as ASHRAE.

Several issues need to be addressed before adopting the high-temperature operation in practice. The most important issue is to enhance the high-temperature tolerance of server components, especially the HDDs and CPUs. Perhaps an increase in the space temperature in data centers by 5-10K could make a huge difference in cooling energy use. Some advanced materials, such as GaN, the third-generation semiconductor material, maybe the breakthrough for electronic components. In addition, optimizing data center cooling systems is another approach to making the high-temperature operation of data centers possible. The optimal design, control, and choice of IT equipment are the main contributors to cooling efficiency enhancement. In high-temperature data centers, free cooling can be adopted to realize chiller-less or even chiller-free cooling plants or operations.

2.2 Energy-Efficient Technologies for Data Center Cooling

2.2.1 Energy performance assessment

Existing studies on the energy efficiency of chilled-water cooling systems have largely concentrated on the influence of climatic conditions. Lei et al. [127] used Sobol's method to develop a model capable of predicting power usage effectiveness (PUE) across different cities. Their findings highlight that climate parameters are the most crucial factors in predicting PUE for water-side economizer systems. Díaz et al. [128] conducted a study to explore the potential of water-side economizers under various climate conditions. They found that the coefficient of performance (COP) of the system could be increased by over 10% during wintertime in coastal climates. Cheung et al. [129] investigated the energy efficiency of the multi-chiller cooling system across different climate zones. They found that cooling energy savings could be achieved by up to 15% in climate Zone 3B by adopting their optimal design.

2.2.2 Maximizing free cooling hours

Li et al. [41] investigated the energy performance of data center cooling systems with a water-side economizer by optimizing free cooling switchover temperature and cooling tower approach temperature. Their results show that significant energy savings, up to 10%, could be achieved by optimizing these parameters. Cheung et al. [129] proposed an optimal design of

data center cooling systems concerning multi-chiller system configuration and component selection for energy-efficient operation and maximized free-cooling. The results show that the optimized design could reduce the annual energy consumption by 3-15% depending on the climate conditions.

2.2.3 AI-based control strategies or optimization

Li et al. [130] optimized the control of data center cooling systems by adopting the emerging deep reinforcement learning method. Their results show that up to 15% of the cooling cost can be reduced by the predicted control setting on the real trace. Han et al. [131] proposed demand-based cooling control strategies to improve the energy efficiency of data centers, and found that the system COP (Coefficient of Performance) could be improved significantly. Ran et al. [132] proposed a joint optimization of job scheduling and cooling control for data center energy efficiency using deep reinforcement learning. They reported that their algorithm can save up to 15% energy consumption and show a better tradeoff between energy saving and service quality.

2.2.4 Incorporating heat recovery systems

Yu et al. [133] studied the energy benefits of heat recovery in a data center in Harbin, and found that heat recovery through heat pumps can meet the thermal demand of the subsidiary buildings. Ebrahimi et al. [50] reviewed and discussed the most promising methods and technologies for recovering data center low-grade waste heat in an effective and economically reasonable way. Chen et al. [134] conducted an experimental study of an integrated data center cooling and waste heat recovery system in different climate conditions. The carbon emissions can be reduced by 244.7 tons annually by adopting the proposed system for a 96kW data center.

2.2.5 Integration with renewable energy sources

G. Rostirolla et al. [135] reviewed and summarized the challenges and solutions for the integration of renewable energy in data centers. He et al. [136] analyzed data center power supply systems based on multiple renewable power configurations. The results show that a maximum renewable penetration of 28.31% can be achieved for a hybrid renewable power system. Han et al. [137] developed a shared energy storage business model for data center clusters considering renewable energy uncertainties. The daily cost of data center clusters can be reduced by 26.36% compared to the baseline scenario. Liang et al. [138] proposed a solar-driven combined cooling and power system for a data center. They found that carbon emissions can be reduced by 257,420 kg per day under design conditions by adopting this system.

2.3 Opportunities for data centers to interact with the power grid

Data centers, labeled 'electricity hogs', have placed considerable strain on the grid. However, this challenge also presents a unique opportunity for data centers to act as pivotal players in grid stability by providing flexible services. By leveraging this opportunity, data centers can potentially reduce their energy costs and even generate revenue, creating a win-win situation. Current research on the interaction of data centers and power grid falls into the following categories.

2.3.1 Renewable energy integration

Data centers can integrate renewable energy sources, such as solar or wind, into their operations. They can consume renewable energy directly or even generate excess energy to feed back into the grid. He et al. [136] analyzed data center power supply systems based on multiple renewable power configurations. The results show that a maximum renewable penetration of 28.31% can be achieved for a hybrid renewable power system. Han et al. [137] developed a shared energy storage business model for data center clusters considering renewable energy uncertainties. The daily cost of data center clusters can be reduced by 26.36% compared to the baseline scenario. Liang et al. [138] proposed a solar-driven combined cooling and power system for a data center. They found that carbon emissions can be reduced by 257,420 kg per day under design conditions when adopting this system.

2.3.2 Utilizing servers and auxiliary equipment for demand response

As Lawrence Berkeley National Laboratory (LBNL) reported “data centers are excellent candidates with great potential for smart grid demand response” [139]. There are some potential demand-side resources to participate in load shedding and load-shifting in data centers.

IT equipment

Servers can adjust their energy consumption through power management systems, such as Dynamic Voltage/Frequency Scaling (DVFS) [140] and power-capping [141]. Chen et al. [142] reported that participating in smart grid demand response programs, i.e., regulation services and frequency control, can reduce data center electricity costs by up to 68.3%, while meeting the service level agreements (SLAs) for quality of service (QoS). Zhang et al. [143] demonstrated that data centers can save their electricity costs by 10% while abiding by all the QoS constraints in a real-world scenario through workload scheduling and CPU power limiting. On the other hand, data centers have delay-tolerant workloads, which can be shifted in time in response to electricity prices or other grid requests. Ghamkhari et al. [144] studied the

coordination between data center and power grid and proposed a linear programming problem to reduce the maximum power flow in the power grid by scheduling the interactive type workload to geo-distributed data centers. Hu et al. [145] investigated cost-efficient workload scheduling, including both the non-elastic interactive workload and the elastic batch workload, for the coordination between the cloud service provider with geo-distributed data centers and smart grids. Their results show that the cost of smart grids can be significantly reduced, by up to 20%, and the load variations of smart grids can be well smoothed simultaneously, by scheduling computing workloads.

Cooling systems

Data centers can operate under a broad range of temperatures [14], which results in a large range of cooling energy. Ghatikar et al. [146] conducted a field test on the ability of data center cooling systems to participate in demand response. Their results show that the response time of activating/deactivating redundant chillers/Computer Room Air Conditioner units (CRACs) is 2 – 8 min. Fu et al. [147] proposed a synergistic control strategy by adjusting the frequency of the servers and the chilled water supply temperature setpoint simultaneously to track the regulation signal from the electrical market. They concluded that the proposed synergistic control strategy can provide an extra regulation capacity of 3% of the design power when chillers are activated, compared with a server-only control strategy. *Backup system.* Backup generators powered by diesel or natural gas are usually configured to start two to four seconds after a utility outage or voltage fluctuation [148]. However, utilizing backup generators for demand response programs is not environmentally friendly [147].

2.3.3 Utilizing energy storage systems for demand response

Data centers can integrate energy storage systems, such as batteries, to store excess energy when demand is low and use it during peak times. This can help to reduce strain on the grid during periods of high demand. Zhang et al. [149] proposed a strategy to leverage thermal storage to cut the electricity bill for cooling, and found that the electricity bill for cooling can be reduced by 15.8% to 20.8% under different CPU utilization levels. Yang et al. [150] found that data center micro-grid tie-line power fluctuations can be effectively regulated by UPS battery group dynamical management. Aksanli analyzed the economic feasibility of using energy storage devices in data centers to reduce their maximum power demand. Guo et al. [151] studied the co-planning problem of networked Internet data centers and battery energy storage systems in a smart grid system. Their results show that the system's quality-of-service,

economics, and reliability can be significantly enhanced, by coordinately planning the data centers' and battery energy storage systems' locations and sizes.

2.4 Computing workload migration in geo-distributed data centers

Fiber-optic communication technology plays a key role in data transmission [152], which enables effective 'communication' among geographically distributed data centers. Over the past four decades, significant progress has been made in this field [153, 154], including the development of wavelength-division multiplexing [155] and space-division multiplexing [156]. Despite these advancements, network delay remains a challenge in long-distance data transmission [157, 158]. In the EDWC initiative, workloads that can tolerate delays, such as storage and backup, will be migrated. Whereas workloads requiring timely responses, such as web search and videoconferencing, will continue to be processed at local data centers in the East. It is noteworthy that data centers typically have a significant proportion of delay-tolerant workloads, more than 50% of total workloads [159]. This provides substantial temporal and spatial flexibility for geographically distributed data centers [160]. Existing studies have shown an increasing interest in the flexible scheduling of workloads in geo-distributed data centers, with a focus on optimizing energy use and cost [161, 162] and maximizing renewable utilization [163]. A study shows that up to 40% of the operational cost can be reduced through load distribution and scheduling in geographically distributed data centers [164]. By migrating data center workloads from the fossil-fuel-heavy regions to the renewable-heavy regions, up to 239 $KtCO_2e$ can be reduced per year [1].

2.5 Summary of identified research gaps

A comprehensive review of the current studies on the energy aspects of data centers is presented in this chapter. Based on the above literature review, research gaps are identified and summarized as follows:

1) Few studies systematically investigate and quantify the global energy-saving potential of high-temperature data centers. Understanding and quantifying this potential is crucial as it serves as a fundamental solution for improving the efficiency of data center cooling systems by changing the cooling mechanism.
2) Most studies on the energy performance of data center cooling systems focus on design conditions. Few studies consider the characteristics of the system's life-cycle operation. However, it is crucial to thoroughly understand the energy performance of these systems under full-range loads and climate conditions throughout their entire life cycle.

3) A significant research gap in developing optimal designs and control strategies for data center cooling systems is the overlooking of progressive loading throughout the data centers' lifecycle. This results in low energy-efficiency operation throughout the lifecycle. An optimal design and control strategy under progressive loading throughout the data centers' lifecycle is an effective means to reduce a large amount of cooling energy.
4) There is a lack of studies on the optimal dispatch and system design of the effective use of surplus capacity in energy storage systems in data centers to provide flexible grid services. By leveraging this opportunity, data centers can potentially reduce their energy costs and even generate revenue, creating a win-win situation.
5) The migration of computing workloads in geographically distributed data centers has emerged as a significant trend in the era of Artificial Intelligence (AI). The Chinese government launched an ambitious initiative, called 'Eastern Data, Western Computing' to facilitate the decarbonization of data centers. The national initiative aims to migrate computing workloads from electricity-deficient Eastern regions to renewable-rich Western regions. However, there is a lack of comprehensive assessment regarding the energy, economic, and carbon impacts of this initiative.

CHAPTER 3 THE GLOBAL ENERGY IMPACT OF RAISING THE SPACE TEMPERATURE FOR HIGH-TEMPERATURE DATA CENTERS

Currently, approaches associated with efficiency enhancement have been widely investigated, including airflow distribution optimization [22, 23], containment of aisles [24, 25], supply and demand match of cooling energy [26, 27], and performance improvement in servers [28]. Although efficiency enhancement is an effective means to reduce energy consumption, it can only save energy to a certain extent. As a fundamental solution, high-temperature data centers can dramatically reduce cooling energy demand by changing the cooling mechanism. It adopts a different cooling mechanism and makes 'chiller-free' data centers possible, facilitating the transition from chiller-based cooling to completely chiller-free cooling in data centers. The high-temperature data center is increasingly regarded as a trend in the data center profession and industry. In 2011, ASHRAE (American Society of Heating, Refrigeration, and Air Conditioning Engineers) further expanded the range of allowable environmental conditions for IT equipment and adopted two new classes in their updated guidelines (class A3 and class A4) based on the updated critical information (IT design and failure data) provided by IT manufacturers [14]. The new Class A3 environment has the upper temperature bound of 40°C and the Class A4 environment even has the upper temperature bound of 45°C.

This chapter presents the global energy impact of raising the space temperature for high-temperature data centers. The trade-off between cooling-energy savings and server power rise is critically analyzed. Quantitative guidance and targets are established for developing 'ideal' and 'recommendable' servers, considering the server performance associated with the thermal environment. Chapter 3.1 introduces the development of the cooling system model. Chapter 3.2 presents the energy performance of data center cooling systems under different ambient and space temperatures. Chapters 3.3 and 3.4 analyze the impacts of raising the space temperature in data centers on free cooling hours and cooling energy savings, respectively. Chapter 3.5 analyzes the impact of raising space temperature on overall data center energy performance and identifies 'ideal servers'.

3.1 Development of the cooling system model and typical operation modes

A commonly used cooling system is selected for the assessment. Its cooling load is assumed to be 4000 kW. The specifications of the corresponding components are selected according to the cooling load, as shown in Table 3.1.

The cooling unit consists of a chiller, a heat exchanger, a cooling tower, a chilled water pump, and a cooling water pump, as shown in Fig. 3.1. The cooling unit can be divided into two subsystems: the cooling water loop and the chilled water loop. A parallel configuration is used in the cooling water loop, and a serial configuration is used in the chilled water loop according to the common engineering practice [15, 165] and comparative assessment [166].

Table 3.1 The specification of the cooling system [129].

Equipment	Design parameter
Water-cooled chiller	Design cooling capacity: 4200 kW
	Chilled water supply temperature: 13-20℃
Variable-speed chilled water pump	Design flow rate: 620000 kg/h
	Design pressure head: 500 kPa
	Design power consumption: 110kW
Constant-speed cooling water pump	Design flow rate: 1240000 kg/h
	Design pressure head: 350 kPa
	Design power consumption: 145 kW
Open cooling tower	Design power consumption: 74 kW
Dry coolers	Design power consumption: 220 kW

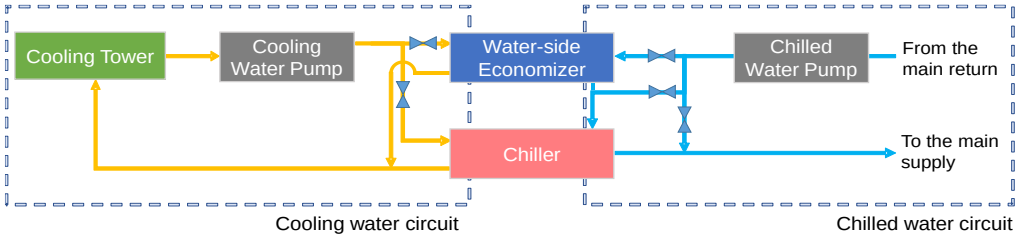


Fig. 3.1 Schematic of a cooling system unit.

3.1.1 Chiller model and the assumption for heat exchangers

Braun's method is used to model chillers [167], in which two variables are involved, i.e., the load and the temperature difference between the leaving condenser and the chilled water flows. Eq. (3.1) correlates the chiller power and variables. In addition, the testing data for the chiller used in this study are from a major manufacturer (Trane), which is designed for high-temperature applications, allowing the chilled water supply temperature to reach up to 20°C.

$$\frac{P_{ch}}{P_{des}} = a_0 + a_1 X + a_2 X^2 + a_3 Y + a_4 Y^2 + a_5 XY \tag{3.1}$$

$$\mathrm{X} = \frac{\dot{Q}_e}{\dot{Q}_{des}} \tag{3.2}$$

$$\mathrm{Y} = \frac{(T_{cwr} - T_{chws})}{\Delta T_{des}} \tag{3.3}$$

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 Eq. (3.1) (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 3.2. An example of chiller COP at the temperature of leaving condenser water of 28°C is shown in Fig. 3.2. In addition, a 1.5°C temperature approach for heat exchangers is assumed [16, 166].

Table 3.2 The empirical coefficients and parameters of the chiller model

Parameter	Value
a_0	0.4623
a_1	0.1659
a_2	0.1947
a_3	-0.3905
a_4	0.1469
a_5	0.4147
Q_{des}	4200 kW
ΔT_{des}	285 K
P_{des}	457.7 kW

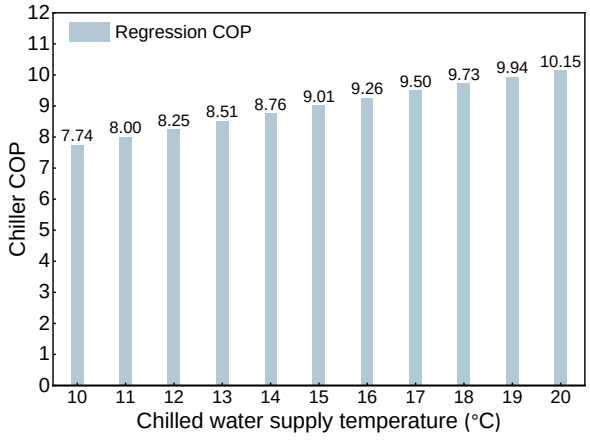


Fig. 3.2 Chiller COP at the temperature of leaving condenser water of 28°C

3.1.2 Cooling tower model

In this study, heat rejection equipment includes open cooling towers and dry coolers [168]. 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 [169]. Braun [167] 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, assuming that the Lewis number equals one. For a counterflow cooling tower, it is described in Eq. (3.4).

$$\varepsilon_a = \frac{1-\exp(-NTU(1-m^*))}{1-m^* exp(-NTU(1-m^*))} \tag{3.4}$$

where,

$$\text{NTU} = \frac{h_D A_v V_{cell}}{m_a} \tag{3.5}$$

$$m^* = \frac{m_a C_s}{m_{w,i} C_{pw}} \tag{3.6}$$

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 using the water inlet and outlet conditions and psychometric data as described in Eq. (3.7).

$$C_s = \frac{h_{s,w,i}-h_{s,w,o}}{T_{w,i}-T_{w,o}} \tag{3.7}$$

The wet-bulb temperature is used as a primary input for the open cooling tower model because the enthalpy can be approximated as a formula related to wet-bulb temperature [170], 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 for these two models can be found in TRNSYS 18 component mathematical manual [169].

According to the desired cooling capacity and outdoor conditions, the airflow rate can be determined. The fan power grows cubically with the rotational speed ideally, (k = 3) as shown in Eq. (3.8) [171]. 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 [171].

$$\frac{W_{ct}}{W_{ct,design}} = \left(\frac{Q_{ct}}{Q_{ct,design}}\right)^k \tag{3.8}$$

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 air flow rate, and $Q_{ct,design}$ is the air flow rate at the design condition.

In the free cooling mode, the speed of the cooling tower fans is controlled to set the outlet water temperature of the heat exchangers to 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 Eq. (3.9). 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 temperature setpoint. The cooling tower control setting given by Eq. (3.9) is used as a proximate or near-optimal setting [129]. 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 [172].

$$T_{ct,out} = \max(T_{wet} + 5℃, T_{\min,\mathrm{cd}}) \tag{3.9}$$

3.1.3 Pump model

The cooling water pumps are usually constant-speed pumps and they are assumed to operate 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 pump efficiency, as shown in Eq. (3.10).

$$W_{cwp} = \frac{\Delta p_{cwp} \times m_w}{\eta_{cwp}} \tag{3.10}$$

The pressure head of pumps (equal to the pressure drop of the chilled water loop) is set to be linear with respect to the water flow rate in operation [171] (Fig. 3.3). This control strategy is nearly optimal for the energy-efficient control of the chilled water loop.

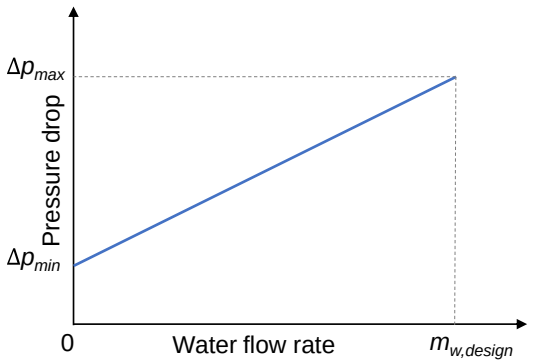


Fig. 3.3 The pressure drop of the chilled water loop

3.1.4 Computer room air handlers

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

3.1.5 Typical operation modes

Fig. 3.4 shows three typical operation modes for water-cooled multi-chiller cooling systems in data centers, mechanical cooling mode, partial free cooling mode and free cooling mode, respectively.

Mechanical cooling mode: There are two circuits in the cooling systems. One is the cooling water circuit, and the other is the chilled water circuit. In the cooling water circuit, the heated cooling water from a chiller flows to a cooling tower and then is cooled through the heat dissipation of the cooling tower. The cooled cooling water enters a chiller. In the chilled water circuit, the chilled water from a chiller enters the computer room air handler and then cools the hot air in the computer room. The heated chilled water goes back to the chiller. The mechanical cooling mode does not involve a plate heat exchanger.

Partial free cooling mode: The cooling water from a cooling tower enters a heat exchanger. Meanwhile, the heated chilled water from hot air in a computer room also enters a heat exchanger. The heat exchanger precools the heated chilled water. The pre-cooled chilled water enters a chiller and then is further cooled by the chiller to achieve the desired chilled water supply temperature. The chilled water further cooled by the chiller enters a computer room to exchange heat with the hot air. Meanwhile, the heated cooling water from the heat exchanger and the chiller both return to the cooling tower. This mode involves chillers and heat exchangers. Heat exchangers are used to precool and handle part of the cooling load.

Free cooling mode: In the cooling water circuit, the cooling water from a cooling tower enters the heat exchanger, and then exchanges heat with heated chilled water from a computer room. After the heat exchange, the heated cooling water returns to the cooling tower for the cooling water circuit. For the chilled water circuit, the chilled water from the heat exchanger enters the computer room air handler and then exchanges heat with hot air in the computer room. The free cooling mode does not involve a chiller, and heat exchangers handle all cooling loads.

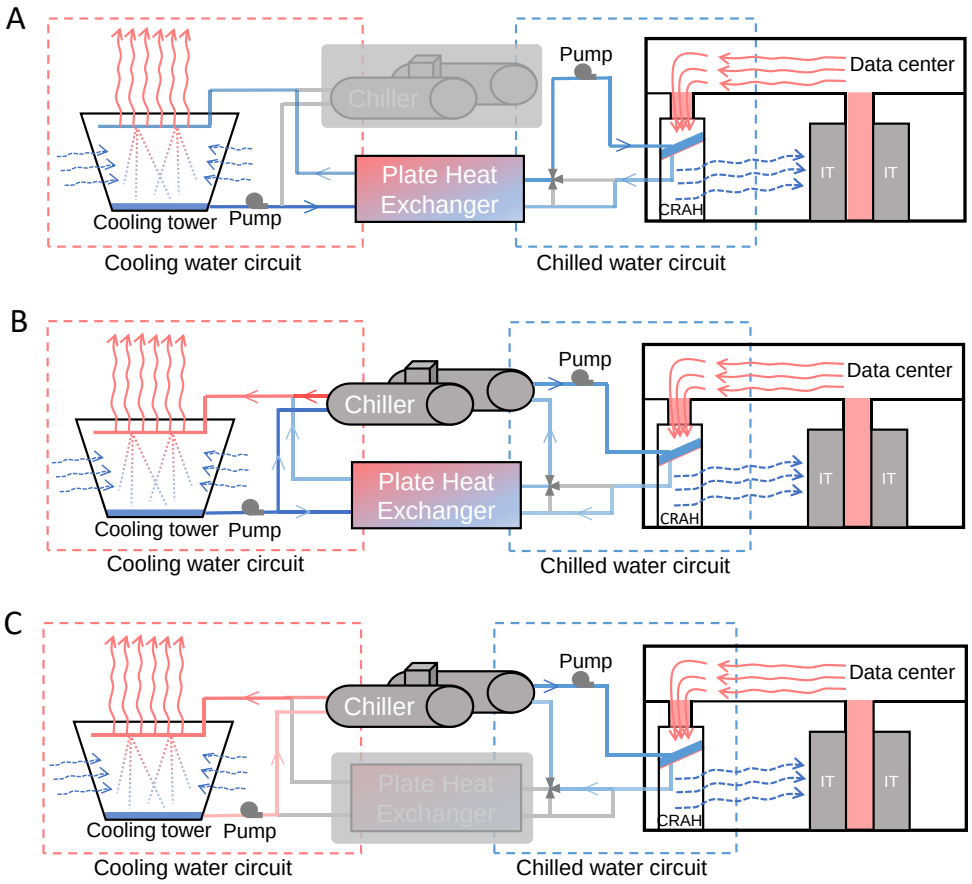


Fig. 3.4 Schematics of typical cooling operation modes

(A) Free cooling mode; (B) Partial free cooling mode; (C) Mechanical cooling mode

3.2 Performance of data center cooling systems under different ambient air temperatures

A basic cooling system model is developed based on fundamental principles and mathematical formulas combined with the test data of cooling equipment performance. The cooling system comprises cooling distribution equipment, cooling equipment and heat rejection equipment. 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.

Fig. 3.5 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 [173]. 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 [168], with this heat transfer mechanism providing an inferior cooling efficiency [173].

The cooling system COP is dramatically reduced once the chiller begins operating. In both partial-free-cooling and mechanical-cooling modes (which require chillers), the cooling system COP is around 3–5, whereas in 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 is only slightly better in partial-free-cooling than in mechanical-cooling mode because only a small percentage of the cooling load can be handled by free cooling in partial-free-cooling mode, with low-COP chiller cooling required for the remaining load. Furthermore, water pumps are required to supply cooling water to the chiller and heat exchangers, which consume more energy than other modes.

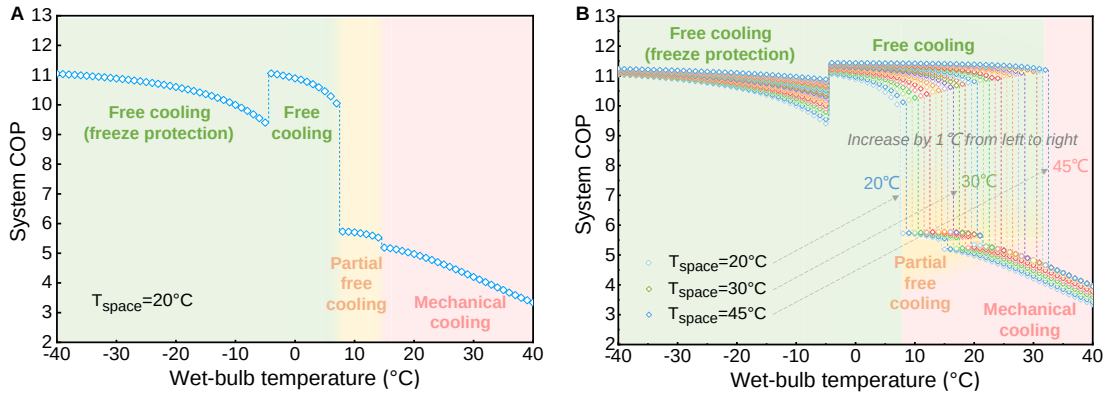


Fig. 3.5 Cooling system COP versus wet-bulb temperature

(A) at the space temperature of 20°C. (B) at space temperatures ranging from 20°C to 45°C.

Fig. 3.5(A) 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 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 require the chillers to operate at higher condensing temperatures, resulting in lower COP for the chillers.

As the space temperature increases, data centers can achieve higher cooling system COP, as shown in Fig. 3.5(B). This is primarily because data centers can operate in free-cooling mode at higher space temperatures, and thus high cooling system COP is achieved over an extended range of ambient air wet-bulb temperatures. Secondarily, higher space temperatures allow for higher chilled water supply temperatures, and thus higher chiller COP 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 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 to 45°C, the free-cooling range is extended by 10K (i.e., a wet-bulb temperature from 7°C to 17°C) and then 25K (i.e., from 7°C to 32°C), respectively. The increase in cooling system COP is in the range of 0.8%–1.7% for every 1K rise in data center space temperature.

3.3 Impact of higher space temperatures on free-cooling hours worldwide

As the space temperature increases, more free-cooling hours become possible. Using the weather data [174] of 57 representative cities from 19 climate zones [175] worldwide, we assessed the potential free-cooling hours at various space temperatures for these cities. Fig. 3.6(A) shows the mean annual free-cooling ratio (i.e., the percentage of free-cooling hours over a year) across 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 the 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 achievable in hot climate zones, such as 0% in the extremely hot–humid Zone 0A. However, if the space temperature is increased to 35°C, over 99% annual free-cooling ratio can be achieved in Zone 5A, while it is also achievable even in Zone 0A if the space temperature is further increased to 38°C. Importantly, all climate zones can achieve nearly 100% free cooling year-round if the space temperature is increased to 41°C.

Therefore, a new concept, the ‘global free-cooling temperature’, has been defined as the data center space temperature at or above which the central cooling system can work in free-cooling mode year-round in nearly all climate zones worldwide. Similarly, the ‘free-cooling temperature’ of a particular city has been defined as the data center space temperature at or

above which the central cooling system can work in free-cooling mode year-round in that city. According to the above analysis, the value of the global free-cooling temperature is 41°C.

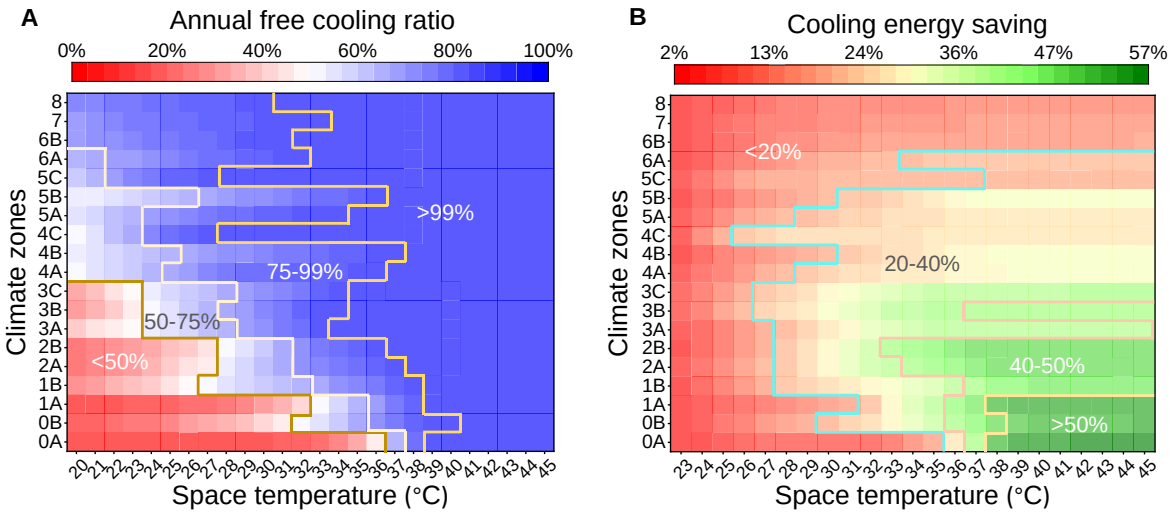


Fig. 3.6 Annual free cooling ratio and cooling energy savings in 19 climate zones

(A) Annual free-cooling ratio upon raising data center space temperature from 20°C to 45°C in 19 climate zones. (B) Annual cooling-energy savings compared with baseline temperature of 22°C in 19 climate zones.

Fig. 3.7 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 further analyze this, we consider the percentages of global land area (1.391 billion km^2) that would allow 50%, 75% or 100% of annual free-cooling at a given space temperature: for a typical space temperature of 22°C (see Fig. 3.7(A)), these percentages are 58.7%, 49.9% and 0%, respectively; for a 27°C space temperature (see Fig. 3.7(B)), they become 69.7%, 58.7% and 0%, respectively; for a 32°C space temperature (see Fig. 3.7(C)), they become 87.0%, 76.5% and 36.5%, respectively; for a 35°C space temperature (see Fig. 3.7(D)), they become 93.9%, 80.8% and 60.2%, respectively; for a 40°C space temperature (see Fig. 3.7(E)), they become 99%, 99% and 93.7%, respectively; and for a 45°C space temperature (see Fig. 3.7(F)), 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.

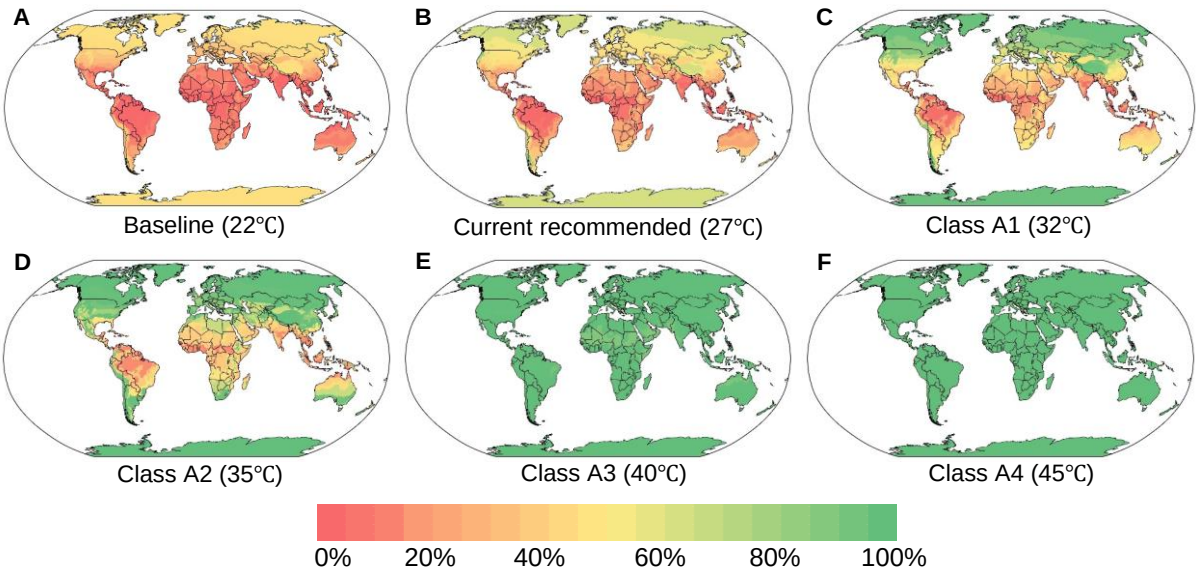


Fig. 3.7 Global maps of annual free-cooling ratio at different space temperatures

(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).

3.4 Impact of raising space temperature on 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 Fig. 3.5). Fig. 3.6(B) 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% of cooling energy could be saved if the space temperature were 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 were raised to 45°C. In addition, every degree (K) of space temperature increase could result in 2%–6% of cooling-energy savings, depending on the climate conditions.

Fig. 3.8 shows the global map of cooling-energy savings when raising the space temperature from the baseline 22°C to 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 further analyze this, we consider the percentages of global land area that could allow 20%, 40% and 50% of annual cooling-energy savings for a given space temperature: for a 27°C space temperature (see Fig. 3.8(A)), these percentages are 2.9%, 0% and 0%,

respectively; for a 32°C space temperature (see Fig. 3.8(B)), they become 44.0%, 0% and 0%, respectively; for a 35°C space temperature (see Fig. 3.8(C)), they become 47.4%, 11.1% and 0%, respectively; for a 40°C space temperature (see Fig. 3.8(D)), they become 53.7%, 36.9% and 19.1%, respectively; and for a 45°C space temperature (see Fig. 3.8(E)), 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.

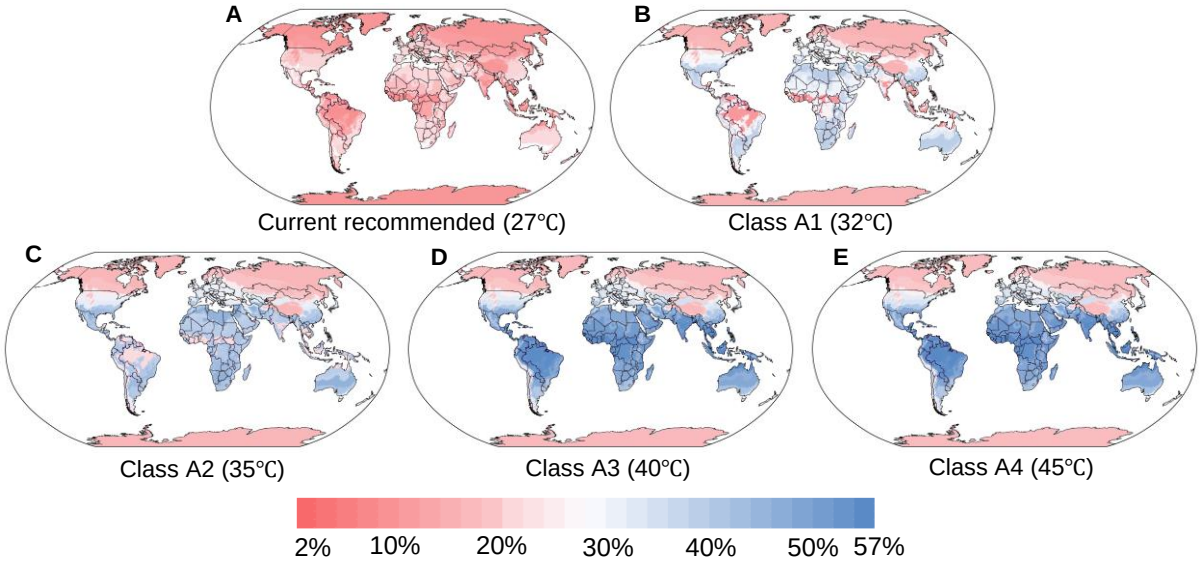


Fig. 3.8 Global maps of cooling-energy savings with reference to baseline space temperature 22°C

(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).

3.5 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. This is therefore an important issue, which needs to be investigated for the future development of server equipment.

Current servers can be categorized into four (thermal) environment classes: Class A1, A2, A3 and A4. Servers of the latest generations (i.e., Class 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 [29]. Notably, Class A3 and A4 servers can work at higher space temperatures and demonstrate less server power increase when working at the same high space temperatures (server inlet temperature) as Class A1 or A2 servers [14]. Generally, the main contribution to server power increase comes from the server fan, as the fan runs faster to prevent the server components from overheating; a relatively minor part of the power increase is due to server components operating to manage the same computing load at higher temperatures [176]. When only considering the server performance associated with the thermal environment and data center cooling energy, it is optimal for servers to operate at higher space temperatures. 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 increase of current and future servers.

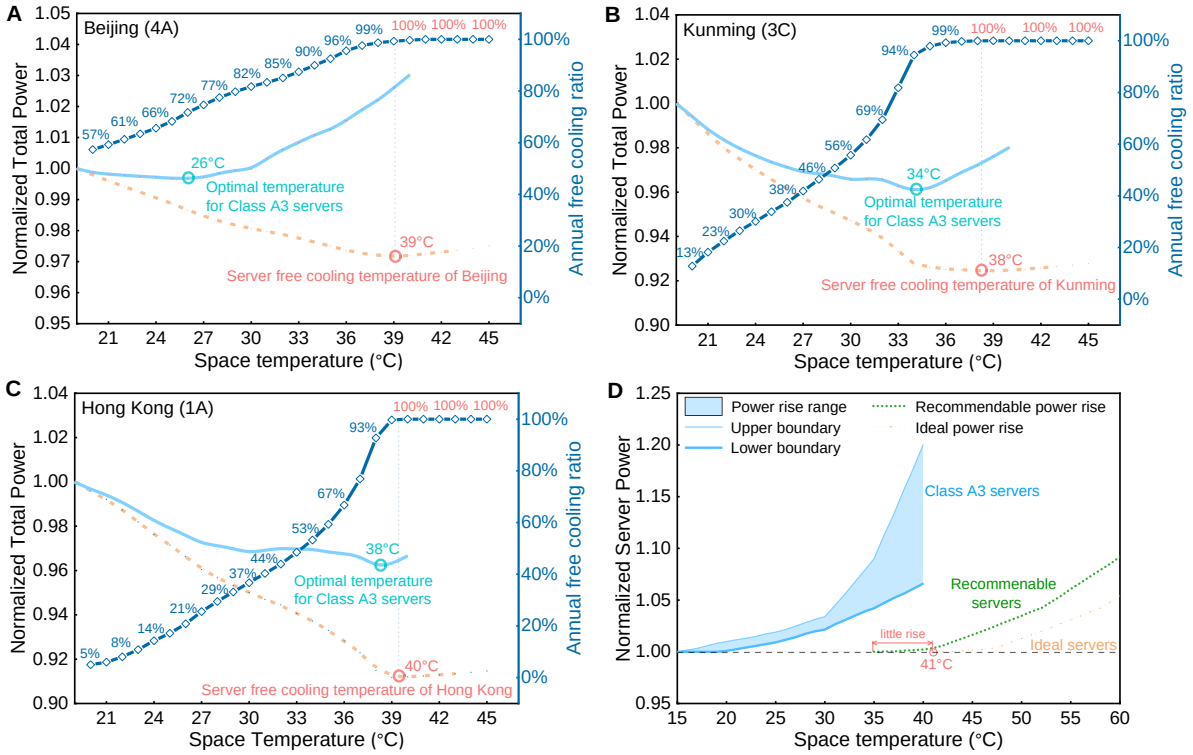


Fig. 3.9 Impact of raising space temperature on the annual free-cooling ratio and normalized server power and total power in data centers in different cities

Normalized total power is calculated against the baseline power at 19°C. Normalized server power is calculated against the baseline power at 15°C [14]. (A) Normalized total power in Beijing. (B)

Normalized total power in Kunming. (C) Normalized total power in Hong Kong. (D) Normalized power rise range of Class A3 servers and the estimated power rise of ideal servers.

Fig. 3.9(A)-(C) shows the normalized total power and the annual free-cooling ratio when raising the space temperature in three cities from three representative climates, namely, Beijing, Kunming and Hong Kong. Fig. 3.9(D) shows the power rise range of Class A3 servers [14] and the power rise trend of ideal servers (dotted lines; note: the power rise of current Class A3 servers at a given space temperature varies, depending on the server models and manufacturers [14]). The normalized total power of adopting Class A3 servers shown in Fig. 3.9(A)-(C) corresponds to the lower boundary of the Class A3 server power rise in Fig. 3.9(D).

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 Fig. 3.9(A)-(C) 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 Fig. 3.9(A)-(C), the free-cooling temperatures of Beijing, Kunming and Hong Kong are 39°C, 38°C and 40°C, respectively. As previously elaborated, to facilitate free cooling of data centers worldwide, 'ideal servers' should be able to work reliably without an increase in the server power when the temperature is no higher than the global free-cooling temperature (i.e., 41°C). This means that the power of ideal servers should not increase with temperature when the temperature is below 41°C (see Fig. 3.9(D)). 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 increases in server power when below the global free-cooling temperature would be highly recommendable – we, therefore, name these 'recommendable servers'.

3.6 Summary

This chapter presents the first comprehensive worldwide investigation of the cooling-energy savings of high-temperature data centers. The results show that every 1K increase in the space temperature could enhance the cooling system COP by 0.8% to 1.7% and reduce cooling energy consumption by 2% to 6%, depending on the climate conditions. 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. Meanwhile, 13% to 56% of 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, typically 20–25°C, while ASHRAE recommends a temperature range of 18–27°C. However, we found that the optimal space temperature in each city depends on the types 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 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 a 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.

Current data center cooling systems are developed from traditional building heating, ventilation and air-conditioning systems, although the primary objects cooled are servers instead of people. 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 [29] and third-generation semiconductors [177]. 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 an upper-temperature limit 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. 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 and chip-level cooling technologies should improve the server performance associated with thermal environments to eliminate barriers and 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.

CHAPTER 4 ENERGY PERFORMANCE ANALYSIS OF MULTI-CHILLER COOLING SYSTEMS FOR DATA CENTERS CONCERNING PROGRESSIVE LOADING THROUGHOUT THE LIFECYCLE UNDER TYPICAL CLIMATES

This chapter presents a thorough assessment of the energy performance of multi-chiller cooling systems throughout the entire lifecycle. Chapter 4.1 presents the specification of the cooling system in the referenced data centers. Chapter 4.2 introduces the cooling system model developed in TRNSYS 18. Chapter 4.3 outlines the energy performance assessment and introduces the specific cooling system designs in different climate conditions. Chapter 4.4 presents the energy performance under full-range loads and climate conditions, and discusses life-cycle energy performance concerning progressive loading.

4.1 Description of multi-chiller cooling systems in the referenced data centers

4.1.1 Specifications of the referenced data center

The cooling system design of the reference date center is shown in Fig. 4.1 [129]. The design considered distribution headers around all cooling towers and all cooling water pumps, and the combination of constant-speed cooling water pumps with different flow capacities. Distribution headers around all cooling towers allow the use of more cooling towers than chillers and water-side economizers, which increase the heat rejection area in the cooling process and subsequently enhance cooling efficiency. Distribution headers around all cooling water pumps can reduce the number of operating pumps under part-load conditions, and then increase the energy efficiency of all pumps. The combination of constant-speed cooling water pumps with different flow capacities could make cooling water pumps work efficiently under part-load conditions.

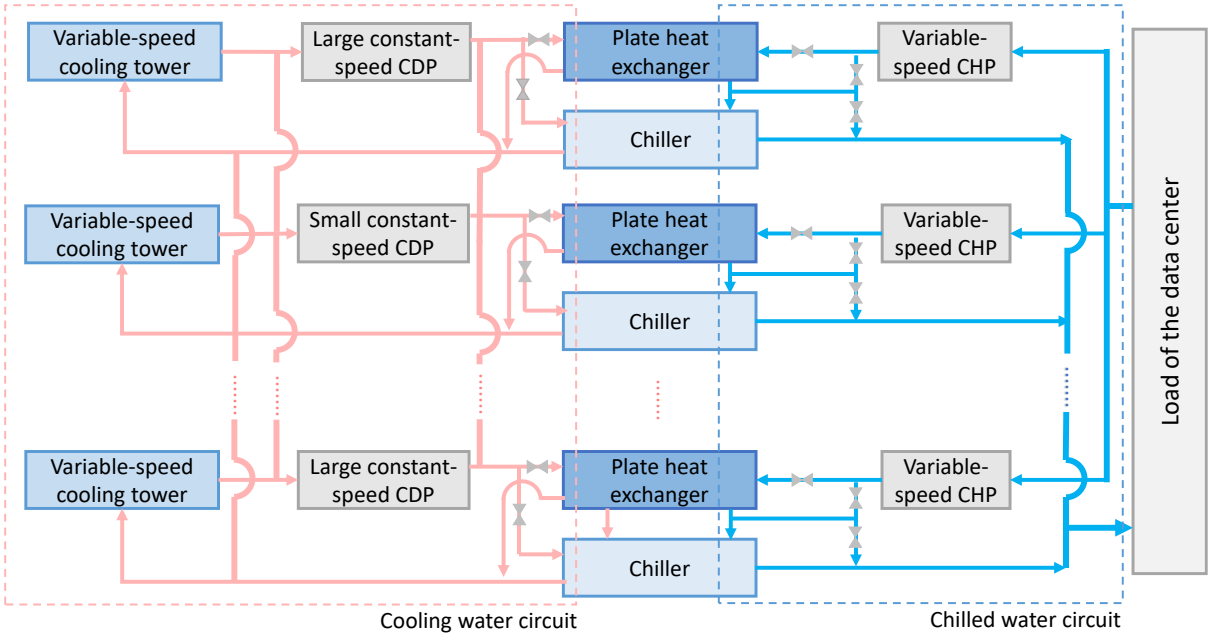


Fig. 4.1 Schematic of the multi-chiller cooling system

The total cooling load of the data center is 16,800 kW [129]. The data center cooling system is equipped with four water-cooled chillers, four water-side economizers, four variable-speed cooling towers, two large constant-speed cooling water pumps (CDP), two small constant-speed cooling water pumps, and four chilled water pumps (CHP). Detailed specifications for each piece of cooling equipment are provided in Table 4.1. For a fair comparison, both the selected open cooling tower and the closed-circuit cooling tower are equipped with axial fans and have largely the same design heat rejection capacity from the same manufacturer.

Table 4.1 Specification of the data center cooling system in the case study

Equipment	Design specification	Quantity
Water-cooled chiller	Design cooling capacity: 4200 kW	4
	Design chilled water outlet temperature: 13°C	
	Design chilled water flow rate: 544300 kg/h	
	Design cooling water flow rate: 620000 kg/h	
Water-side economizer	Design heat transfer rate: 4300 kW	4
	Design water flow rate: 620000 kg/h	
Variable-speed chilled water pump	Design flow rate: 620000 kg/h	4
	Design pressure head: 500 kPa	
	Design power consumption: 110kW	
Large constant-speed cooling water pump	Design flow rate: 1240000 kg/h	2
	Design pressure head: 350 kPa	
	Design power consumption: 145 kW	
Small constant-speed cooling water pump	Design flow rate: 620000 kg/h	2
	Design pressure head: 350 kPa	
	Design power consumption: 84 kW	
Variable-speed open cooling tower	Design power consumption: 74 kW	4

Variable-speed closed-circuit cooling tower	Design air flow rate: 235.6 m^3/s Design heat rejection rate: 16800 kW Design power for anti-freezing electric heater: 60 kW Design power consumption: 110 kW Design heat rejection rate: 16800 kW Design air flow rate: 268.88 m^3/s Design water flow rate: 1071429 kg/h	4

4.1.2 Typical control algorithms of the cooling system

This section elaborates on control algorithms used in the case study. The control algorithms involve the operation mode of the cooling system and the operation of the cooling equipment.

The operation mode of the cooling system: Fig. 4.2 shows the procedure for selecting the optimal operation mode. First, all three cooling modes will be simulated under each cooling load and ambient air temperature. Then, according to the simulation results, the cooling modes that can meet the cooling load will be identified. Lastly, the one exhibiting the lowest power consumption will be selected as the optimal operation mode.

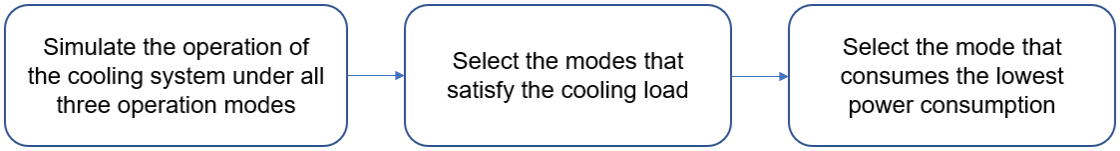


Fig. 4.2 Procedure to select the optimal operation mode

Number of chillers, heat exchangers and cooling towers in operation: The number of operating chillers and economizers varies with the operation mode. The principle of operating equipment quantity is to meet the cooling load. The number of cooling towers is larger than the number of operating chillers and waterside economizers to increase the heat rejection area. To avoid a lack of water in operating cooling towers, the number of cooling towers in operation is one more than the number of chillers and economizers in operation. Table 4.2 shows the number of chillers, heat exchangers, and cooling towers in operation in three cooling modes.

Table 4.2 Number of chillers, water-side economizers, and cooling towers in operation

Operation mode	Cooling load (kW)	Number of chillers in operation	Number of water-side economizers in operation	Number of cooling towers in operation
Mechanical cooling	< 4200 kW	1	0	2
	4200 kW -8400 kW	2	0	3
	8400 kW -12600 kW	3	0	4
	> 12600 kW	4	0	4
Partial free cooling	< 8500 kW	1	1	2
	> 8500 kW	2	2	3

Free cooling	< 4300 kW	0	1	2
	4300 kW -8600 kW	0	2	3
	8600 kW -12900 kW	0	3	4
	> 12900 kW	0	4	4

Chilled water supply temperature of chillers or heat exchangers: The chilled water supply temperature is set at 13°C. It is common to set the chilled water supply temperature at 13°C (then the supplying cold air is 20-22°C) or even higher, in line with the recommended space temperature (20-27°C) in data centers, according to ASHRAE standard [14]. In mechanical cooling and partial free cooling mode, the chilled water supply temperature is achieved by chillers. In free cooling mode, the chilled water supply temperature is achieved by controlling the fan speed of the cooling towers.

Number of pumps in operation: In the study, the number of chilled water pumps is the same as the larger number of operating chillers or the number of operating economizers. The number of cooling water pumps is shown in Table 4.3.

Table 4.3 Number of operating cooling water pumps in operation

Total required cooling water flow	Number of operating cooling water pumps
< 620000 kg/h	1 small pump
620000 kg/h -1240000 kg/h	1 large pump
1240000 kg/h - 1860000 kg/h	1 large pump and 1 small pump
> 1860000 kg/h	2 large pumps

Operating speed of variable speed water pumps: The number of operating chilled water pumps is the maximum of the number of operating chillers and the number of operating heat exchangers. The speed of chilled water pumps is controlled according to the required total chilled water flow rate (Eq. (4.1)). The speed limits of the chilled water pumps are 30 Hz to 50 Hz to protect their motors. Therefore, the frequency of chilled water pumps changes when the number of chillers or heat exchangers changes.

The speed of large and small constant-speed cooling water pumps always remains at 50 Hz.

$$m_{req,chw} = N_{op,ch}\dot{m}_{des,chw,ch} + N_{op,hx}\dot{m}_{des,chw,hx} \quad (4.1)$$

where, $m_{req,chw}$ is the required total chilled water flow rate, $N_{op,ch}$ is the number of operating chillers, $N_{op,hx}$ is the number of operating heat exchangers, $\dot{m}_{des,chw,ch}$ is the design chilled

water flow rate of the chiller (544300 kg/h), $\dot{m}_{des,chw,hx}$ is the design chilled water flow rate of the heat exchanger (620000 kg/h).

Operating speed of cooling tower fans: The speed of the variable-speed fans in cooling towers depends on the operation mode and the ambient wet-bulb temperature. In free cooling mode, the speed of cooling tower fans is controlled according to the principle that the water supply temperature of water-side economizers reaches 13°C. In mechanical cooling and partial free cooling mode, the supplied chilled water is controlled by chillers. The speed of the fans is controlled to maintain the cooling tower water outlet temperature at the set point, given by Eq. (4.2). The setpoint ensures optimal performance of the fan and considers the chiller minimum condenser water entering temperature [178]. In addition, the spray-water pump of the closed-circuit cooling tower is shut off for freeze protection if the outdoor temperature is below 0°C.

$$T_{ct,out,set} = Max(T_{amb,wet} + 5[℃], 18[℃]) \quad (4.2)$$

The temperature difference between air supply and return in computer rooms: the temperature difference between air supply and return in computer rooms is fixed at 10°C [179, 180]. In addition, since the environment of computer rooms is isolated, there is no humidification requirement in the cooling system [181].

4.2 Development of the cooling system model

The operation of the data center cooling system is simulated using TRNSYS 18, with the majority of cooling equipment having been well-validated by previous studies according to the TRNSYS 18 user manual [169]. In addition, typical meteorological year (TMY) weather files [182] were used in the simulations.

Chillers are modeled using Type 142, which relies on catalog data provided as external text files to determine chiller performance. The performance testing data of the chiller used in this study is from a major manufacturer (York). Table 4.4 and Table 4.5 summarize and present the PLR and performance input data of Type 142 used in this study.

Table 4.4 The PLR data of chiller model Type 142

PLR	Fraction of full power
0.15	0.165
0.2	0.195
0.3	0.262
0.4	0.328
0.5	0.398
0.6	0.480

0.7	0.580
0.8	0.698
0.9	0.843
1	1

Table 4.5 The performance data of chiller model Type 142

Chilled water supply temperature	Cooling water supply temperature	Chiller COP
7	18	7.021
7	22	6.310
7	24	6.207
7	28	5.979
7	32	5.478
10	18	8.020
10	22	7.182
10	24	6.407
10	28	6.154
10	32	5.807
12	18	9.729
12	22	8.554
12	24	7.964
12	28	7.071
12	32	6.208
15	18	11.438
15	22	9.908
15	24	9.298
15	28	8.073
15	32	7.034

Open cooling towers are modeled using Type 162. Closed-circuit cooling towers are modeled using Type 510. The ε-NTU method is used to model open cooling towers. Air effectiveness (ε_a) 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 one [167] For a counterflow cooling tower, it is described by Eq. (4.3).

$$\varepsilon_a = \frac{1-\exp(-NTU(1-m^*))}{1-m^* exp(-NTU(1-m^*))} \tag{4.3}$$

$$\text{NTU} = \frac{h_D A_v V_{cell}}{m_a} \tag{4.4}$$

$$m^* = \frac{m_a C_s}{m_{w,i} C_{pw}} \tag{4.5}$$

where, h_D is mass transfer coefficient, A_v is the surface area of water droplets per tower cell exchange volume, V_{cell} is the total tower cell exchange volume, m_a is the mass flow rate of air, C_{pw} is the constant pressure specific heat of water, and C_s is the saturation-specific heat.

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 [170], at a given atmospheric pressure. 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 as shown in Eq. (4.6) [171].

$$\frac{P_{ct}}{P_{ct,design}} = \left(\frac{Q_{ct}}{Q_{ct,design}}\right)^k \tag{4.6}$$

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

Closed cooling towers rely on the basic premise that the saturated air temperature is the temperature at the air-water interface and is also the temperature of the outlet fluid. The saturated air enthalpy can be calculated by Eq. (4.7).

$$h_{sat}\left(T_{fluid,out}\right) = h_{air}\left(T_{air,in}\right) + \frac{\dot{Q}_{fluid}}{\dot{m}_{air}\left(1-\exp\left[-\beta_{design}\left(\frac{\dot{m}_{air}}{\dot{m}_{air,design}}\right)^{y-1}\right]\right)} \tag{4.7}$$

$$\beta_{design} = Ln\left[\frac{h_{sat}\left(T_{fluid,out,design}\right)-h_{air}\left(T_{air,in,design}\right)}{h_{sat}\left(T_{fluid,out,design}\right)-h_{air}\left(T_{air,out,design}\right)}\right] \tag{4.8}$$

where, h_{sat} is heat transfer coefficient at saturated air condition, m is mass flow rate, Q is heat transfer rate, h_T is the heat transfer coefficient between the fluid and the air at the given temperature T, and y = 0.6 for most applications.

The outlet air enthalpy could be found from an energy balance on the cooling tower and can be expressed as:

$$\dot{Q}_{fluid,design} = \dot{m}_{fluid} Cp_{fluid}\left(T_{fluid,in,design} - T_{fluid,out,design}\right) \tag{4.9}$$

$$h_{air}\left(T_{air,out,design}\right) = h_{air}\left(T_{air,in,design}\right) + \frac{\dot{Q}_{fluid,design}}{\dot{m}_{air}} \tag{4.10}$$

The airflow rate is calculated by Eq. (4.11). The fan power is in cubic growth of the airflow rate.

$$\dot{m}_{air} = \dot{m}_{air,design}\gamma_{air} \tag{4.11}$$

$$\gamma_{air} = \frac{\dot{m}_{air}}{\dot{m}_{air,design}} \tag{4.12}$$

where, γ is the ratio of flow rate to design flow rate.

The pump models are modified based on Type 743. The pump flow rate is calculated according to the speed of the pump and the pressure difference of the pipelines. The pressure difference of the pipeline is estimated by Eq. (4.13), which is a variation from the widely used fan affinity law. Eq. (4.14) shows the relationship between pump speed, pump pressure difference and pump flow rate [129].

$$\Delta P_{pipe} = \Delta P_{des,pump}(\frac{\dot{m}_{pump}}{\dot{m}_{des,pump}})^2 \tag{4.13}$$

$$\Delta P_{pump} = c_{pump,0} f_{pump}^2 + c_{pump,1} f_{pump} \dot{m}_{pump}^2 \tag{4.14}$$

where, ΔP is the pressure difference, $\dot{m}_{pump}$ is the flow rate of operating pumps in its water circuit. $c_{pump,0}$ and $c_{pump,1}$ are the parameters of the pump model, f_{pump} is the speed of the pump. According to the manufacturers' specifications of pumps, $c_{pump,0}$ is 0.023, and $c_{pump,1}$ is -3.9E-13 for chilled water pumps; $c_{pump,0}$ is 0.0178, and $c_{pump,1}$ is -1.89E-13 for small cooling water pumps; $c_{pump,0}$ is 0.0177, and $c_{pump,1}$ is -1.32E-13 for large cooling water pumps.

Heat exchangers are modeled using Type 657. In the heat exchanger model, the energy transfer across the heat exchanger is given by Eq. (4.15)

$$Q_{HX} = \varepsilon C_{min}(T_{hot,in} - T_{cold,in}) \tag{4.15}$$

where, Q_{HX} is the energy transfer across the heat exchanger, ε is the heat exchanger's effectiveness, 0.857 in this study according to manufacturing testing data, C_{min} is the minimum of the hot and cold-side fluid thermal capacitances, $T_{hot,in}$ is the temperature of the fluid entering the hot side of the heat exchanger.

CRAH is modeled according to the Eq. (4.16). A variable-speed fan that changes airflow based on the cooling load was applied to the CRAH in the reference data center [183]. The difference between supply and return air temperatures of the computer rooms is assumed to be a constant of 10K [2].

$$\frac{P_{CRAH}}{P_{CRAH,des}} = \frac{\dot{V}_{SA}^3}{\dot{V}_{SA,des}^3} \tag{4.16}$$

where, P_{CRAH} is the power consumption of CRAH, $\dot{V}_{SA}$ is the air flow rate.

4.3 Outline of energy performance assessment and specific cooling system designs

4.3.1 Outline of energy performance assessment

The outline of the assessment procedure is elaborated as shown in Fig. 4.3.

i. The cooling system models are developed using the test data of cooling equipment and typical control algorithms using the software TRNSYS 18. The models consider the variations in system designs related to freeze protection for different climates.

ii. Python programming is used to change inputs, such as weather data and cooling loads, within the cooling system model created in TRNSYS 18. Additionally, Python programming is used to determine the optimal operation mode for each specific weather condition and cooling load.

iii. The energy performance of the cooling system components, as well as relevant metrics such as free cooling hours, cooling system COP, and data center PUE, are identified under full-range cooling loads and climate conditions.

iv. The energy performance of the cooling system is quantified throughout the lifecycle with a typical progressive loading scenario.

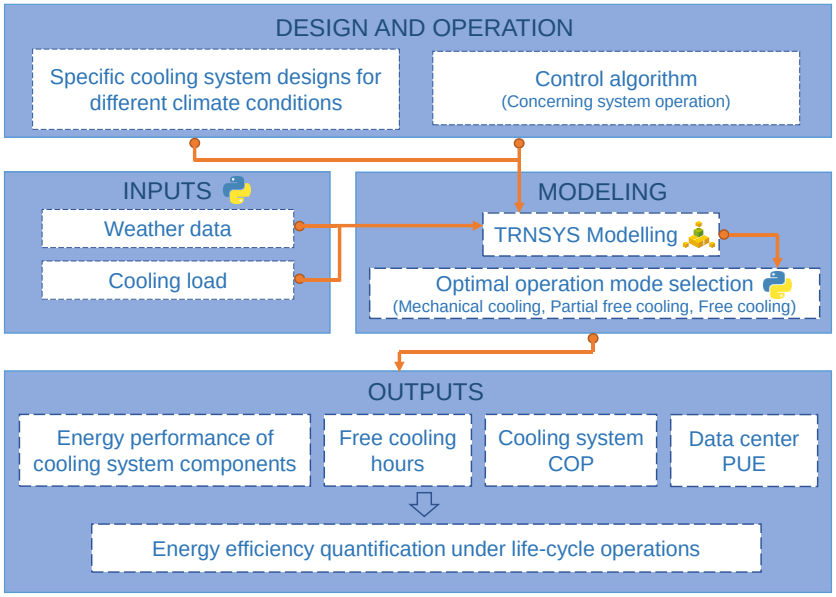


Fig. 4.3 Procedure and steps of energy performance assessment

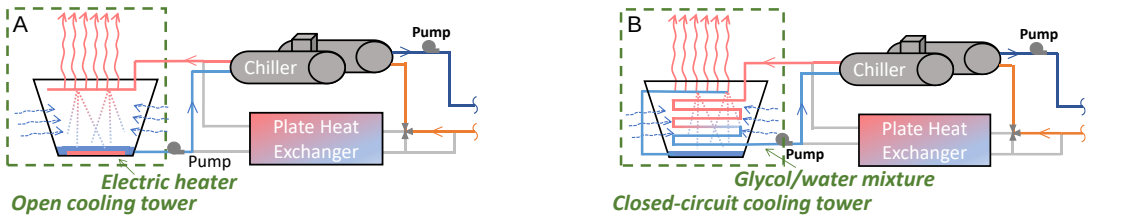
4.3.2 Cooling system designs for different climate conditions

In practice, the implementation of identical cooling systems in data centers situated in different climatic conditions often necessitates modification in design and operation to address the specific needs and constraints associated with each climate. Table 4.6 shows six cities located in different climate conditions and their specific designs.

Table 4.6 Selected climate zones and cities

Zone	City	Longitude and latitude	Specific design
Hot summer and warm winter	Hong Kong	22.3°N and 114.2°E	Fig. 4.4(A)
Temperate	Kunming	25.0°N and 102.7°E	Fig. 4.4(A)
Hot summer and cold winter	Shanghai	31.2°N and 121.4°E	Fig. 4.4(A)
Cold	Beijing	39.9°N and 116.3°E	Fig. 4.4(A)
Severe cold	Ulanqab	40.1°N and 110.3°E	Fig. 4.4(A)
Extremely severe cold	Harbin	45.8°N and 126.8°E	Fig. 4.4(B)

Fig. 4.4 shows specific designs for freeze protection in water-cooled cooling systems. Generally, open cooling towers are widely used in the cooling system due to their high heat rejection efficiency and low capital cost [173, 184], shown in Fig. 4.4(A). To prevent freezing, electric heaters are installed inside open cooling towers for freeze protection and are activated under necessary or extreme conditions. However, in regions experiencing exceptionally harsh winter, where outdoor air temperature can reach -35°C, such as Harbin. Electric heaters may prove inadequate for preventing freeze-related issues according to engineering cases. For such extreme cold environments, closed-circuit cooling towers and antifreeze solutions are adopted for freeze protection [168], as shown in Fig. 4.4(B). The antifreeze solution used in the closed-circuit cooling tower typically consists of a mixture of water and ethylene glycol.



Design for most regions Specific design for extremely cold regions

Fig. 4.4 Schematic of specific cooling system designs

4.3.3 Data center energy flow and performance metrics

Electricity energy flow in data centers

Fig. 4.5 shows the electrical energy flow in a data center [185, 186]. In a typical data center, electrical energy is supplied from the utility grid to power uninterruptible power supply (UPS) systems, cooling systems as well as other miscellaneous equipment (e.g., lighting and offices). UPS provides power to the IT equipment. Generally, there will be electricity losses at UPS. In this study, UPS electricity loss and other miscellaneous electricity consumption are considered to account for 12% of the IT electricity consumption [127, 179].

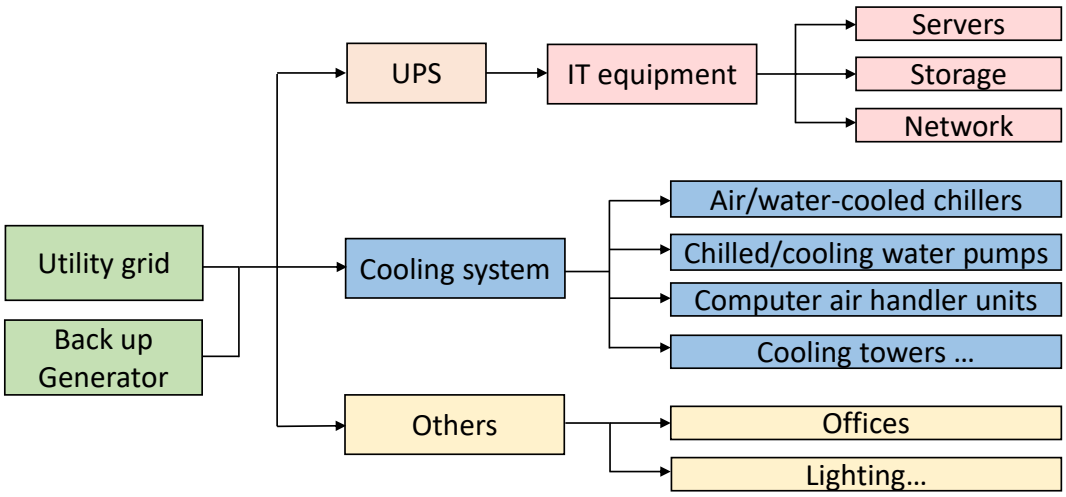


Fig. 4.5 Electricity flow in a data center

Power usage effectiveness (PUE)

In the data center industry, the most used metric that evaluates the energy effectiveness of the data center is Power Usage Effectiveness (PUE). PUE was proposed in 2006 [187] and promoted by the Green Grid (a non-profit organization of IT professionals) in 2007 [188]. PUE is an indicator of the energy efficiency of data centers. A lower PUE represents a more efficient data center. It means that more electrical energy is used by IT equipment instead of other equipment. A PUE of 1 is an ideal value [189].

$$PUE = \frac{Total\ Facility\ Power}{IT\ Equipment\ Power} = \frac{P_{IT}+P_{cooling}+P_{others}}{P_{IT}} \tag{4.17}$$

Energy efficiency of cooling systems

To further analyze the energy performance and efficiency of data center cooling systems, the coefficient of cooling system performance (COP) is proposed, given by Eq. (4.18) [129].

$$\text{Cooling system } COP = \frac{Q}{W_{system}} \tag{4.18}$$

where Q is the cooling load and W_{system} is the energy consumption of the data center cooling system, including the energy consumption of the cooling plant and the computer room air handler (CRAH).

In addition, the part load ratio (PLR) is defined by Eq. (4.19). The cooling load in data centers is dominated by the servers themselves. Therefore, the impact of weather-related cooling load variations in different cities is often considered negligible in studies related to data center cooling systems [128, 190, 191].

$$PLR = \frac{Actual\ cooling\ load}{Design\ cooling\ load} \tag{4.19}$$

4.4 Energy performance under full-range loads and climate conditions

4.4.1 Energy performance of cooling system components

The cooling system is designed with a total of four cooling equipment units. According to the control algorithms of the cooling system, the part-load ratios (PLRs) of 0.25, 0.5, and 0.75 are transition points where the number of chillers would increase or decrease.

Fig. 4.6 (corresponding to the specific design shown in Fig. 4.4(A)) and Fig. 4.7 (corresponding to the specific design shown in Fig. 4.4(B)) illustrate the energy performance of cooling system components under full-range loads and ambient wet-bulb temperatures. Notably, there are distinct dividing lines that fluctuate within the range of 3°C -9°C for four components. These wet-bulb temperatures serve as crucial switching points for different operation modes of the cooling system, such as the transition from mechanical cooling to partial free cooling or from partial free cooling to free cooling.

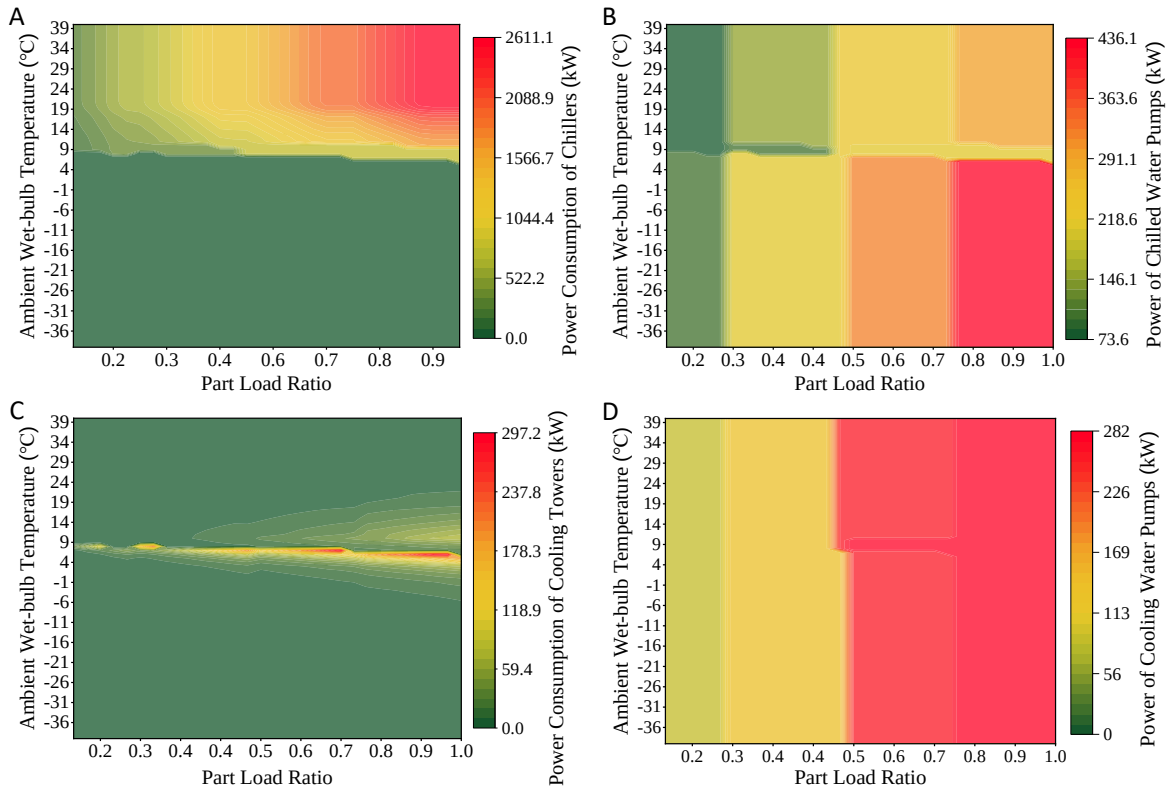


Fig. 4.6 Energy performance of cooling system components with the specific design (Fig. 4.4(A)), chiller (A), chilled water pump (B), open cooling tower (C), and cooling water pump (D)

In Fig. 4.6(A) and Fig. 4.7(A), the power consumption of chillers increases as wet-bulb temperature rises (with chillers operating when wet bulb temperature exceeds 9°C). This increase is due to a lower chiller COP when the cooling water temperature is higher. As the

ambient air temperature increases, the cooling water outlet temperature from the cooling tower also increases, according to the control algorithm of cooling towers. Additionally, it is observed that the power consumption of chillers increases as cooling system PLR increases, and increases slowly when PLR is near 0.2 (one chiller in operation), 0.4 (two chillers in operation), 0.6 (three chillers in operation) and 0.8 (four chillers in operation). This can be attributed to the fact that each chiller operates at a chiller PLR of 0.8 when the cooling system PLR is near 0.2, 0.4, 0.6 and 0.8. The chiller COP increases as the chiller PLR rises, reaches its peak at a chiller PLR of approximately 0.8, and then experiences a slight decrease as the PLR continues to increase to 1.

In Fig. 4.6(B, D) and Fig. 4.7(B, D), there are sudden increases in the power consumption of chilled water pumps and cooling water pumps when the PLR approaches 0.25, 0.5 and 0.75. This change is directly related to the variation in the cooling load, which impacts the number of chilled water pumps in operation. Notably, the power consumption of chilled water pumps varies across different operation modes in Fig. 4.6(B) and Fig. 4.7(B). This variation is attributed to the specific control algorithm of chilled water pumps. The change in power consumption of cooling water pumps (in Fig. 4.6(D) and Fig. 4.7(D)) results from adjustments in the number of cooling water pumps in operation.

The primary difference between the two specific designs lies in the energy performance of cooling towers. In general, closed cooling towers, as shown in Fig. 4.7(C), consume more power compared to open cooling towers shown in Fig. 4.6(C). Notably, there is a distinct dividing line at wet-bulb temperatures ranging from 5°C to 9°C for open cooling towers (Fig. 4.6(C)), and from 3°C to 7°C for closed-circuit cooling towers (Fig. 4.7(C)). The wet-bulb temperature in the dividing line is the switching temperature for different operation modes at different PLRs.

The switch temperatures for operation modes that adopt closed-circuit cooling towers are lower than those that adopt open cooling towers. This difference is attributed to the lower cooling efficiency of closed-circuit cooling towers, as they lack direct contact between the cooling water and the outdoor cold air [173]. In addition, the use of antifreeze solutions in closed-circuit cooling towers results in lower specific heat capacity and consequently a lower convective heat transfer coefficient in heat exchangers, leading to a larger approach temperature for heat exchangers.

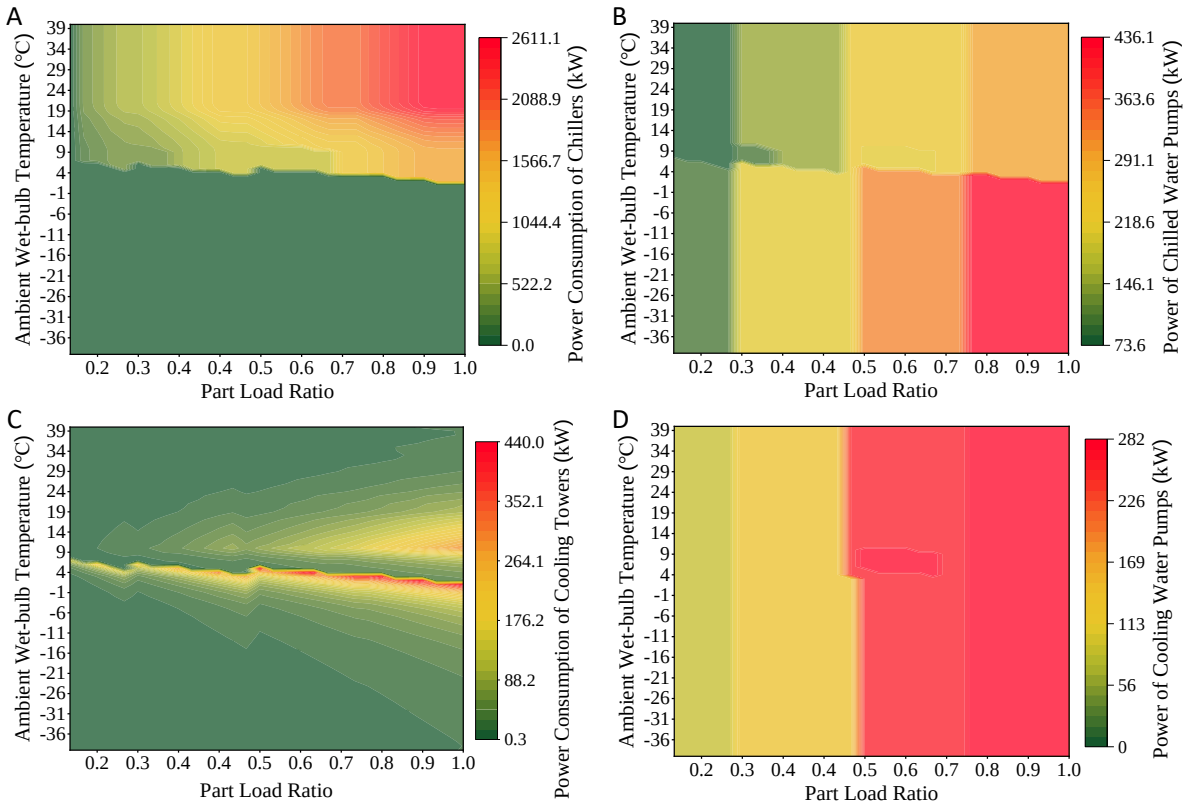


Fig. 4.7 Energy performance of cooling system components with the specific design (Fig. 4.4(B)), chiller (A), chilled water pump (B), closed-circuit cooling tower (C) and cooling water pump (D)

4.4.2 Overall energy performance of data centers

Free cooling hour under full-range cooling load

Fig. 4.8 illustrates the operating hours of three cooling modes (free cooling, partial free cooling, and mechanical cooling) at different part-load ratios (PLR) throughout the year in six representative cities. Among these cities, Ulanqab shows the maximum free cooling hours, reaching 5921 hours at a PLR of 1. Following Ulanqab, the cities of Harbin, Kunming, Shanghai, Beijing, and Hong Kong show decreasing free cooling hours. Hong Kong shows the minimum free cooling hours, only 158 hours. Furthermore, it can be observed that the number of free cooling hours decreases as the PLR increases in all cities. For example, in Kunming, the free cooling hours decrease by up to 1442 hours throughout the year when the PLR increases from 0.1 to 1. The variation in free cooling hours in different cities is mainly due to the change in the switching point from partial free cooling mode to free cooling mode. For example, when the partial load ratio (PLR) is 0.3, the switching wet-bulb temperature from partial free cooling mode to free cooling mode is 11℃. When the PLR is 1, the switching wet-bulb temperature from partial free cooling mode to free cooling mode is 8℃. This means that

the free cooling hours will decrease by the total hours at the wet-bulb temperatures of 9-11℃ in different cities. The total hours at wet-bulb temperatures of 11℃, 10℃, and 9℃ vary in different cities, resulting in varying decreases in free cooling hours when the PLR increases. The change in switching wet-bulb temperature can be attributed to the fact that the cooling tower capacity is “oversized” when the cooling load is lower than the designed cooling load. Essentially, the “oversized” cooling towers have the capability to handle the cooling load even at higher wet-bulb temperatures than the design condition.

It is worth noting that there are rebounds in free cooling hours when the PLRs are 0.25 and 0.5. This is due to the change in the number of cooling towers in operation. According to control algorithms, the number of cooling towers in operation increases from 2 to 3 when the PLR is 0.25, and from 3 to 4 when the PLR is 0.5. When more cooling towers are activated, the total heat rejection area of the cooling towers also increases. This means that there is an "oversized" cooling tower capacity at PLRs of 0.25 and 0.5. The cooling towers have the capability to handle the cooling load even at higher wet-bulb temperatures. Consequently, there are more available free cooling hours that can meet the cooling load once the number of cooling towers increases at these critical switching points. Meanwhile, there is no rebound in free cooling hours at a PLR of 0.75. This is because the design quantity of cooling towers is 4 in this study. There is no increase in the number of cooling towers at a PLR of 0.75.

Fig. 4.8 Operating hours of three cooling modes in different cities

Energy performance of data centers under full-range cooling loads

Fig. 4.9 shows the annual average data center PUE and cooling system COP at different part-load ratios (PLRs) in six representative cities. The difference in cooling system COP at different PLRs can be as high as 6. Generally, cooling system COP increases as the part-load ratio increases except for three significant drops in cooling system COP at the PLRs of 0.25, 0.5, and 0.75. As explained in Section 4.1, these PLRs act as transition points where the number of chillers in operation may increase or decrease. At these critical transition points, activating an additional chiller will result in multiple chillers operating at their off-design conditions and thus a decrease in cooling efficiency. For instance, when the part-load ratio is close to and below 0.25, only one chiller is in operation, which operates at its design conditions and thus

has a high chiller COP. However, when the PLR slightly exceeds 0.25, a second chiller is activated, causing both chillers to operate at their off-design conditions and resulting in a low chiller COP. As the PLR increases towards 0.5, both chillers can operate closer to their design conditions and thus a high chiller COP once again.

As the PLR increases, the data center PUE shows the opposite trend compared to the cooling system COP. This is due to the fact that a higher cooling system COP means lower cooling energy consumption, resulting in a lower PUE. The difference in data center PUE at different PLRs can be up to 0.14.

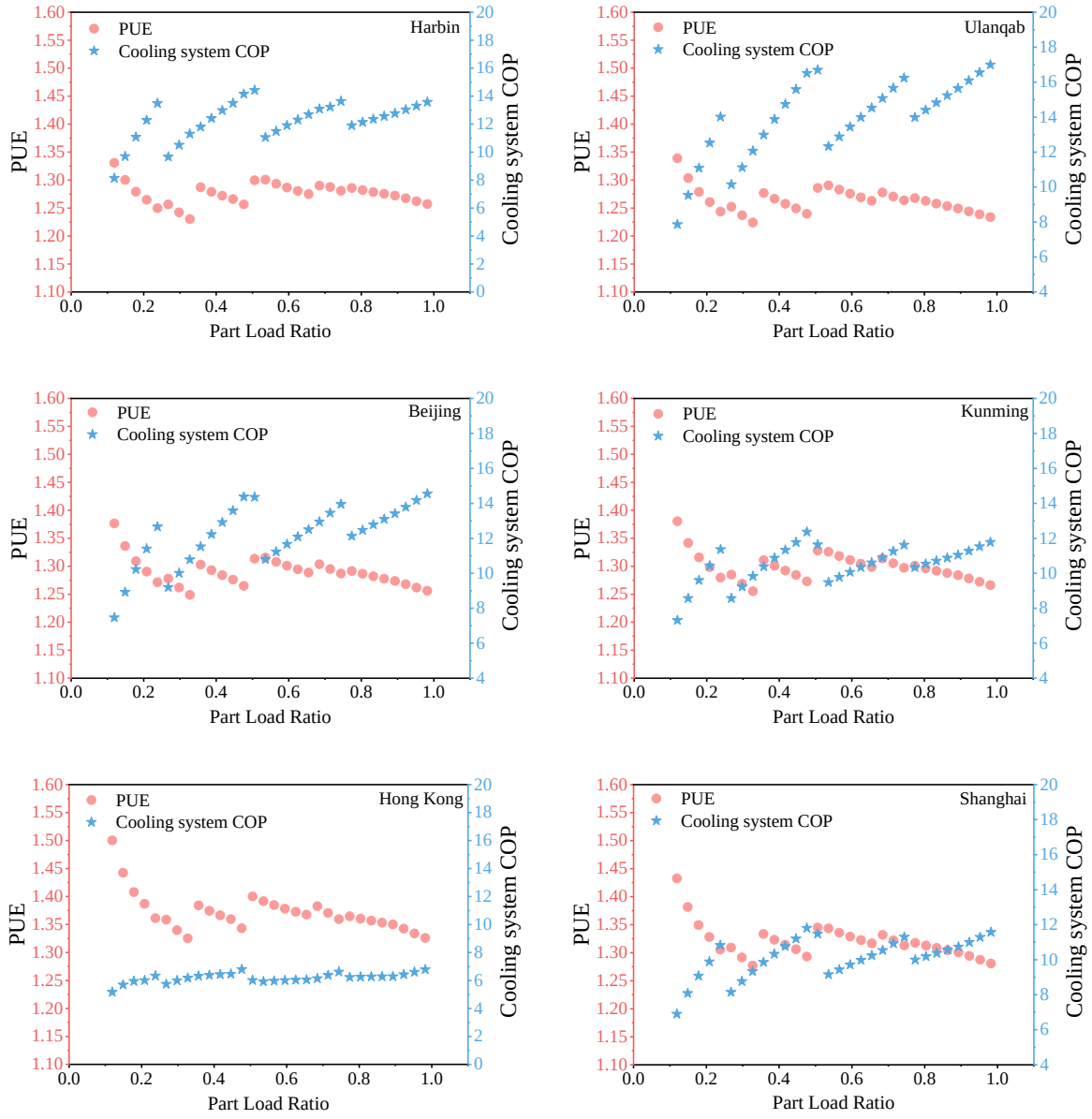


Fig. 4.9 Annual average data center PUE and cooling system COP at part-load ratios in different cities

Fig. 4.10 compares cooling system COP at PLRs of 0.9 and 1.0 over the typical year, taking Kunming as an example. There is no change in the number of cooling equipment at the PLRs of 0.9 and 1.0. It is observed that the cooling system COP at a PLR of 1.0 is consistently higher than that at a PLR of 0.9 even though the latter has higher free cooling hours. Several factors contribute to this observation. In free-cooling mode, the increased cooling loads necessitate a corresponding increase in energy consumption by the cooling towers. However, the increased energy consumption of the cooling towers is lower than the increased cooling loads. In mechanical-cooling mode and partial-free-cooling mode, the chillers play a key role in determining the total energy consumption. Similarly, the increased energy consumption of chillers is also lower than the increased cooling loads. Therefore, the cooling system operates more efficiently at a PLR of 1.0 compared to a PLR of 0.9.

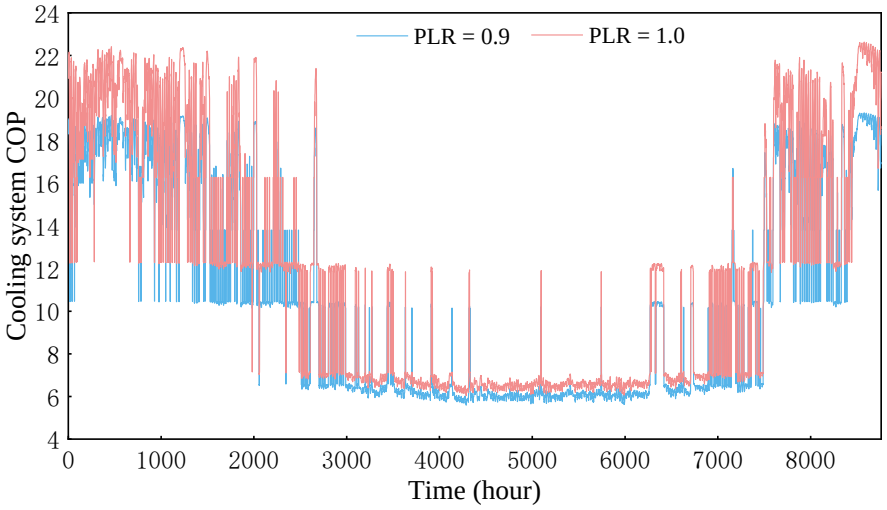


Fig. 4.10 Comparison of cooling system COP at PLRs of 0.9 and 1.0 over the typical year in Kunming

Impact of climate condition on energy performance

Fig. 4.11 (A) compares cooling system COP in six cities at different PLRs. The rankings of cooling system COP in these cities, from highest to lowest, are Ulanqab, Harbin, Beijing, Kunming, Shanghai, and Hong Kong. The primary reason for the difference in cooling system COP among these cities is the distribution of annual wet-bulb temperatures. The determinant is the number of hours below the critical wet-bulb temperature, which triggers the data center cooling system to switch from free-cooling mode to other modes. This indicates that Ulanqab experiences more favorable conditions for free cooling, with the highest number of hours below the critical wet-bulb temperature.

Harbin is characterized as an extremely cold region with harsh winter weather conditions. Consequently, closed-circuit cooling towers are adopted in Harbin, along with the use of an antifreeze solution to prevent freezing. Ulanqab benefits from a more favorable climate with consistently cool temperatures throughout the year. In Ulanqab, only electric heaters are required to address freezing concerns. It is worth noting that the cooling efficiency of closed-circuit cooling towers is lower than that of open cooling towers because the cooling water is not in direct contact with the outdoor cold air [173]. Furthermore, antifreeze solutions have a lower thermal conductivity than water, leading to a certain decrease in the cooling system's efficiency in Harbin.

Fig. 4.11 (B) compares data center PUE in six cities at different PLRs. Among these cities, Ulanqab exhibits the lowest PUE due to its excellent climate conditions and higher cooling efficiency. In addition, it can be observed that there are some slight fluctuations in PUE, except for the three major fluctuations in PUE due to the switch of chiller numbers. These slight fluctuations are attributed to changes in the number of computer room air handlers in operation at different PLRs.

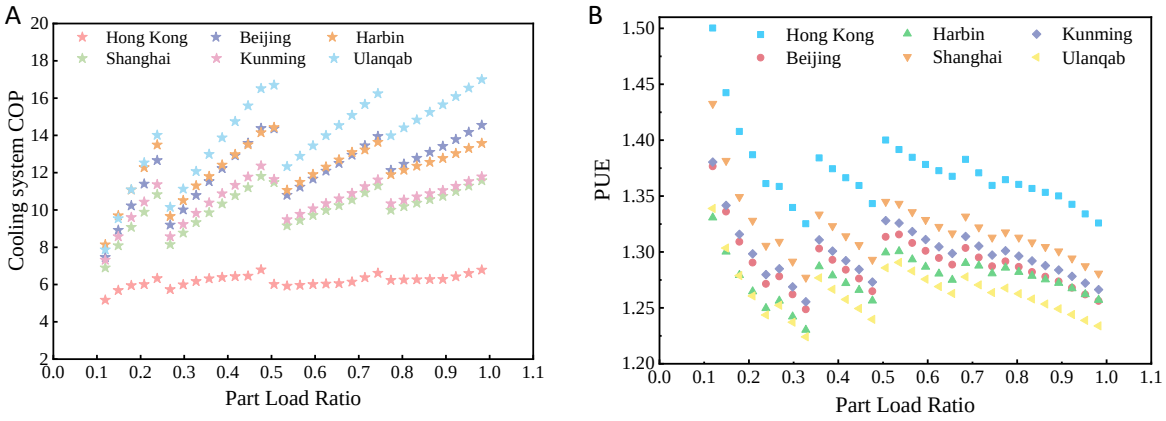


Fig. 4.11 Annual average cooling system COP (A) and data center PUE (B) under full-range cooling loads in different cities

4.4.3 Energy performance of the data center cooling system under progressive loading

The IT load within data centers is not static but experiences a progressive increase throughout their lifecycle, as shown in Fig. 4.12(A) [192]. The expected load of the data center starts at 30% and gradually ramps up to a final expected load value of 90%. However, it is worth noting that the actual start-up load is around 20% of the design load, gradually increasing to an ultimate actual load of approximately 60% of the design capacity. This indicates that most data centers operate at part load for the majority of their lifetime [193]. The final achieved IT load

is significantly lower than the initial design IT load. This progressive loading characteristic poses challenges to the high-efficiency operation of the cooling system throughout the data center's lifespan. As the IT load increases gradually, the cooling system may not operate at its optimal efficiency, leading to higher energy consumption and potential inefficiencies in cooling capacity utilization.

Fig. 4.12(B) illustrates the difference between the design cooling system COP and the actual cooling system COP in Beijing. It is observed that there is a significant gap between the design COP and the actual COP. The average cooling system COP throughout the lifecycle with the progressive increase in IT load is only 11.7, 2.9 lower than the design system COP. This indicates that the multi-chiller cooling system operates inefficiently for a major portion of its operational lifetime, resulting in substantial energy waste. In addition, it is noticeable that there is also a relatively low cooling system COP throughout the lifespan of the data center under the expected IT load increase. Fig. 4.13 shows the energy performance of the cooling system under an assumption of a 20-year life cycle (also adopting the progressive loading profile from reference [192]. It is observed that the cooling system operates at a system COP of 11.7 for the majority of its lifetime. However, it is important to note that further analysis of the energy performance of the cooling system is needed if the progressive loading profile differs.

The primary reason behind the inefficiency is that during the design phase, designers typically focus on the design IT load and incorporate extra capacity for emergencies. However, they often overlook the progressive increase in IT load and the ultimate IT load that the data center will experience. As a result, the cooling system is often oversized, and the match between the capacity and the quantity of cooling system components is not optimal, which would adversely affect the overall efficiency of the cooling system over the data center's lifetime.

To address this issue, it is crucial to consider the progressive loading nature of data centers when designing and operating the cooling system. This includes optimizing the cooling system design to match the evolving IT load profile, implementing adaptive control strategies to adjust cooling capacity in response to changing demands, and utilizing load modulation techniques to optimize energy efficiency during part-load conditions. By considering these factors, data center operators and designers can mitigate the energy inefficiencies associated with progressive loading and achieve greater energy efficiency throughout the data center's lifetime.

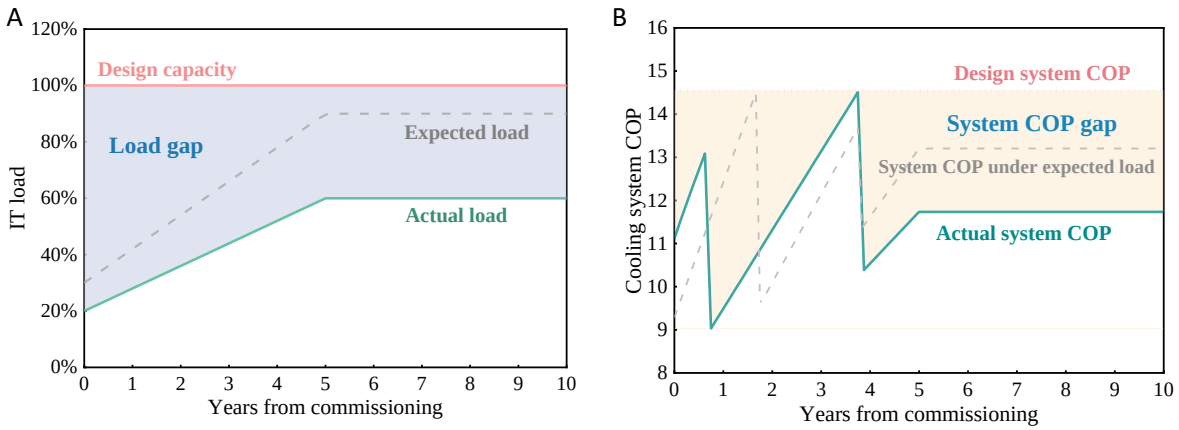


Fig. 4.12 A typical IT load growth (A) and cooling system COP (B) over a lifetime of 10 years

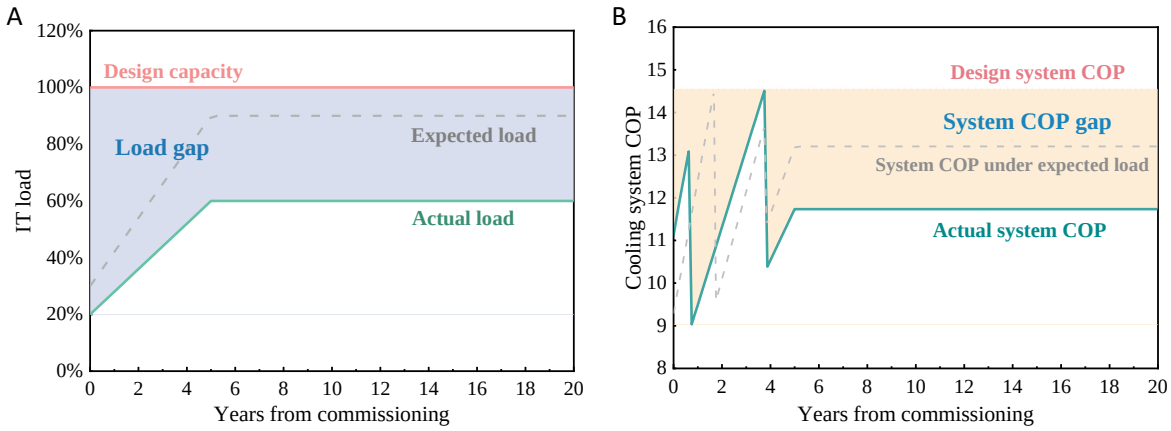


Fig. 4.13 A typical IT load growth (A) and cooling system COP (B) over a lifetime of 20 years

4.5 Summary

This chapter presents a pioneering assessment and quantification of the energy performance of multi-chiller cooling systems in data centers concerning progressive loading throughout the lifecycle. Through an extensive analysis of the energy performance of cooling system components under full-range cooling loads and wet-bulb temperatures, the significant impacts of progressive loading on cooling system COP, data center PUE, and free cooling hours are identified. The quantitative results offer valuable insights for designing optimal cooling systems and achieving high-efficiency HVAC systems and cooling plants for data centers, particularly at partial loads throughout the lifecycle.

Our findings reveal that there is a notable variance in cooling system COP at different PLRs. The difference in cooling system COP at different PLRs can be as high as 6, corresponding to

a difference in PUE of up to 0.14. Additionally, free cooling time could differ up to 1442 hours at different PLRs in the same location. Furthermore, on average, the cooling system COP throughout the lifecycle with a progressively increasing IT load is 2.9 lower than the COP under design conditions.

To address this inefficiency, it is imperative for the future design and operation of data centers to take into account the progressive nature of IT load increases. This involves the optimization of the cooling system design to match the evolving IT load profile and the implementation of adaptive control strategies to adjust cooling capacity in response to changing demands. By considering these factors, data center operators and designers can mitigate the energy inefficiencies associated with progressive loading and achieve greater energy efficiency throughout the data center's lifetime.

This study offers critical insights into the energy performance of multi-chiller cooling systems in air-cooled data centers and presents quantified results that can inform the development of next-generation, high-efficiency cooling solutions. By addressing the challenges identified and adopting the recommended strategies, the data center industry can move towards more sustainable and energy-efficient operations.

CHAPTER 5 LIFE-CYCLE OPTIMAL DESIGN AND ENERGY BENEFITS OF CENTRALIZED COOLING SYSTEMS FOR DATA CENTERS CONCERNING PROGRESSIVE LOADING

This chapter presents an optimal design method for centralized cooling systems with multiple chillers under progressive loading throughout the lifecycle. Chapter 5.1 presents the formulation of the optimization problem of centralized cooling systems in data centers. Chapter 5.2 identifies the optimal designs in different climate zones based on the energy performance under full-range loads and ambient temperatures. Chapter 5.2 also analyzes and compares free cooling hours, cooling energy, and life-cycle costs of the optimized designs and conventional designs.

5.1 Formulation for optimizing data center cooling system concerning progressive loading

5.1.1 Typical progressive IT loading of data centers

The IT load within data centers is not static but experiences a progressive increase over their lifetime, as illustrated in Fig. 5.1 [192]. It is assumed that the initial start-up load is around 20% of the design load, step wisely increasing to an ultimate load of approximately 60% of the design capacity. The loading progresses through stages, with the load being 20% in stage 1, 30% in stage 2, 40% in stage 3, 50% in stage 4, and 60% in stage 5. It is worth noting that most designs will incorporate the additional capacity for redundancy and future requirements [194]. Therefore, it is assumed that the design capacity is 1.1 times the 100% IT load.

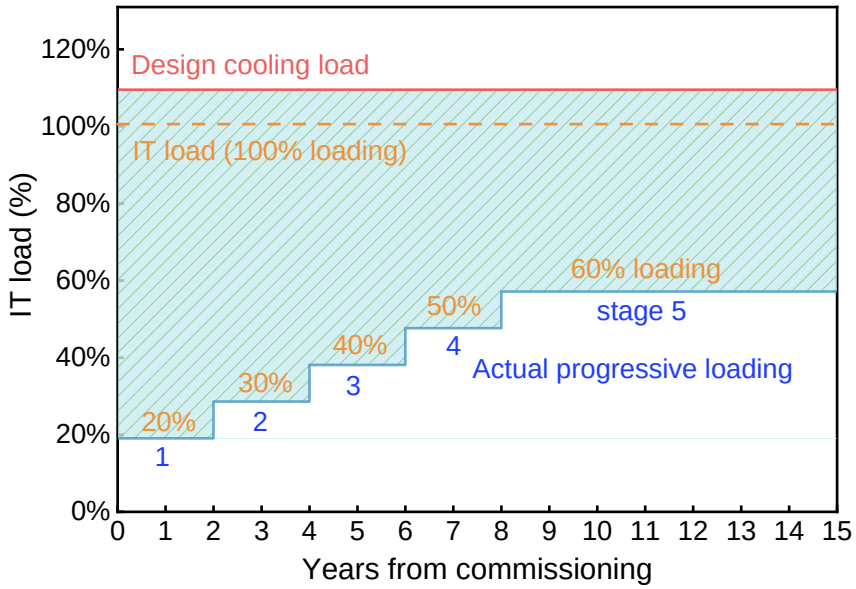


Fig. 5.1 Typical progressive IT loading of data centers

5.1.2 Outline of the optimal design of centralized cooling systems for data centers

The outline of the optimal design method is elaborated as shown in Fig. 5.2.

i. A basic cooling system model is developed using the test data of cooling equipment and typical control algorithms. The cooling system is simulated in three operation modes: mechanical cooling mode, partial free cooling mode, and free cooling mode, at wet-bulb temperatures ranging from −40°C to 40°C and under full-range cooling loads. The mode that satisfies the cooling load and consumes the lowest cooling energy (the highest COP) is selected for each outdoor condition.

ii. Two-step optimization is conducted. The first step is to maximize the life-cycle energy efficiency of the cooling system under different numbers of cooling units. Using the progressive loading data and energy performance data, the optimization problem of maximizing the life-cycle energy efficiency of the data center cooling system is formulated. The SLSQP optimization algorithm in Python is used to obtain the optimal capacity combination of cooling units and optimal PLRs of cooling units under each stage.

iii. The second step is to find the optimal number of cooling units. The total costs, including operating cost and capital cost under N = 3, 4, 5, 6 and 7, can be obtained according to the first-step results and the cost models of cooling equipment. Then, the optimal number of cooling units can be determined.

iv. A worldwide analysis of the optimal design is conducted using typical meteorological year (TMY) weather data [182] by selecting 19 representative cities (Table 5.1) in different

climate zones identified by ASHRAE (Fig. 5.3 [175]). The worldwide free cooling hours, cooling system efficiency, cooling energy and total cost of the optimized designs can be obtained.

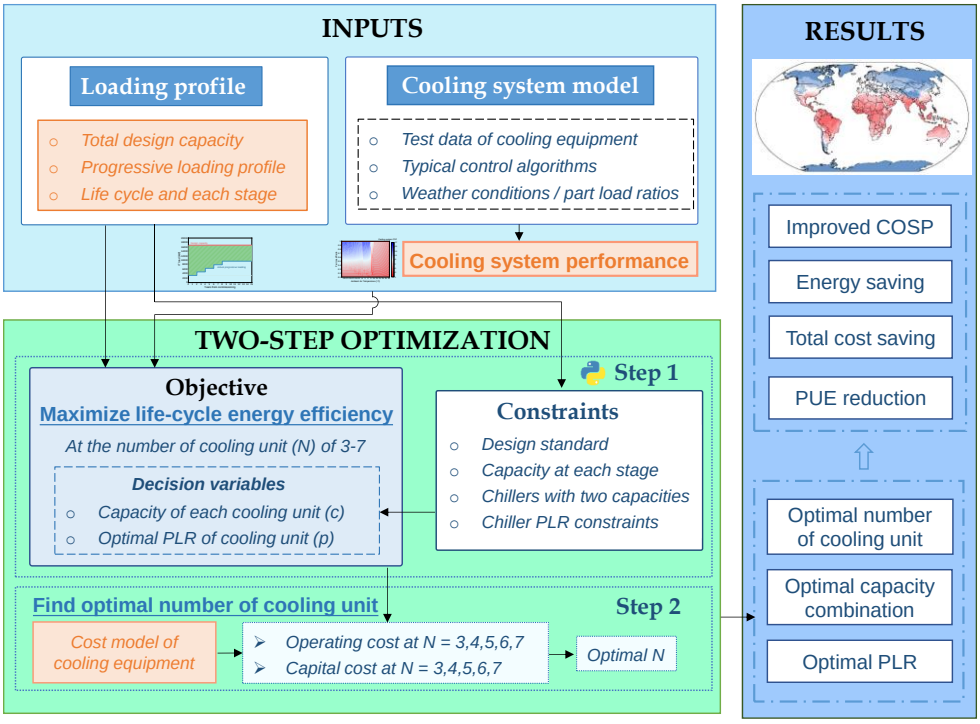


Fig. 5.2 Procedure and steps of energy performance assessment

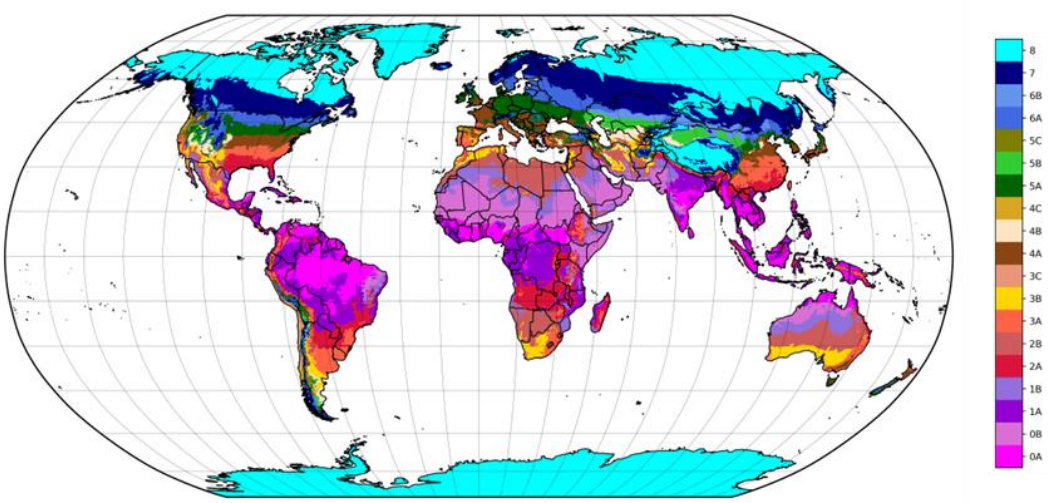


Fig. 5.3 Global map of 19 climate zones

Table 5.1 Climate zones and 19 representative cities [175]

Zones	Features	City
0A	Extremely Hot Humid	Bangkok (Thailand)
0B	Extremely Hot Dry	Abu Dhabi (United Arab Emirates)
1A	Very Hot Humid	Hong Kong (China)

1B	Very Hot Dry	New Delhi (India)
2A	Hot Humid	Fuzhou (China)
2B	Hot Dry	Cairo (Egypt)
3A	Warm Humid	Guiyang (China)
3B	Warm Dry	Adelaide (Australia)
3C	Warm Marine	Kunming (China)
4A	Mixed Humid	Beijing (China)
4B	Mixed Dry	Madrid (Spain)
4C	Mixed Marine	Seattle (USA)
5A	Cool Humid	Yingkou (China)
5B	Cool Dry	Taiyuan (China)
5C	Cool Marine	Comox (Canada)
6A	Cold Humid	Changchun (China)
6B	Cold Dry	Chifeng (China)
7	Very Cold	Harbin (China)
8	Subarctic/Arctic	Mohe (China)

5.1.3 Objectives and constraints of the optimization problem

The first step is to maximize the life-cycle energy efficiency of the cooling system under different numbers of cooling units by adopting the SLSQP (Sequential Least Squares Quadratic Programming) optimization algorithm. The SLSQP algorithm is a numerical optimization method used to solve constrained nonlinear optimization problems [195]. The basic principle of SLSQP is to iteratively minimize a quadratic approximation of the objective function subject to a set of linear and nonlinear equality and inequality constraints. The optimization process of SLSQP involves iteratively updating the current solution by solving a sequence of quadratic programming subproblems. At each iteration, the algorithm computes the gradient and Hessian of the objective function and constraints, and then uses these to update the current solution in the direction that minimizes the quadratic approximation of the objective function while satisfying the constraints [196].

In the first-step optimization, the objective can be formulated as the maximization of the life-cycle energy efficiency of the cooling system under the number of cooling units of 3-7. The decision variables to be optimized include the capacity of each cooling unit and the part load ratio of each cooling unit. The optimization objective of maximizing the energy efficiency (EE) of cooling system units over their lifetime is presented in Eq. (5.1).

$$\text{Maximize EE} = \sum_{1 \le s \le 5} \tau_s * TCOSP_s \tag{5.1}$$

$$TCOSP_s = \sum_{\substack{i,j \\ 1\le i\le N \\ 1+(s-1)N\le j\le sN}} c_i * p_j \Big/ \sum_{\substack{i,j \\ 1\le i\le N \\ 1+(s-1)N\le j\le sN}} (c_i * p_j / COSP_j) \quad (5.2)$$

$$COSP_j = f(p_j) \quad (5.3)$$

where, $TCOSP_s$ is the total coefficient of the system performance under the stage *s*, *s* = *1,2,3,4,5*. τ_s is the time weighting factor of stage *s*. N is the number of cooling units. c_i is the capacity of the cooling unit *i*. p_j is the part load ratio of each cooling unit under different stages. $COSP_j$ is the coefficient of the system performance at a part load ratio of p_j.

In actual operation, there are constraints for the design and operation of the cooling system and equipment. The primary constraint is to meet design standards for data center cooling systems. The total design capacity of the cooling system is 1.1 times the 100% IT load (L_{design}). In addition, chillers with two capacities are preferred for convenient maintenance and control in practice [197].

$$\sum_{1\le i\le N} c_i = 1.1 * L_{design} \quad (5.4)$$

subject to: $c_i \in c_1, c_2$

In each stage, the total cooling capacity of cooling units should meet actual loadings. L is the actual loading at different stages.

$$\sum_{\substack{1\le i\le N \\ 1+(s-1)N\le j\le sN}} c_i * p_j = L_s \quad (5.5)$$

The range of values for decision variables, the capacity of each cooling unit(i.e., chiller), is assumed as 500-7000kW.

$$500 \text{ kW} \le c_i \le 7000 \text{ kW} \quad (5.6)$$

The range of values for decision variables, the chiller PLR, is assumed as 0.3-1 if the chiller is in operation. The minimum value of PLR is 0.3 because a surge may occur when the load of chillers is as low as 30% of the rated value [198].

$$p_j = 0 \ or \ 0.3 < p_j \le 1 \quad (5.7)$$

In the second-step optimization, the objective is to find the optimal number of cooling units that have the lowest life-cycle total cost (LTC), including capital cost (CC) and operating cost (OC). The objective can be given in the following Eq. (5.8),

$$\text{Minimize LTC} = \text{CC} + \text{OC} \quad (5.8)$$

5.2 Description of baseline scenario and economic analysis

5.2.1 Description of the baseline scenario

The specifications of the cooling equipment (Table 4.1) are sourced from the referenced data center presented in Chapter 4. The baseline design for the centralized cooling system with multiple chillers consists of four same-size water-cooled chillers, four water-side economizers, four-speed variable cooling towers, four cooling water pumps, and four chilled water pumps. Fig. 5.4 shows the schematic of the data center cooling system in the baseline scenario. Each cooling unit is shown in Fig. 3.1.

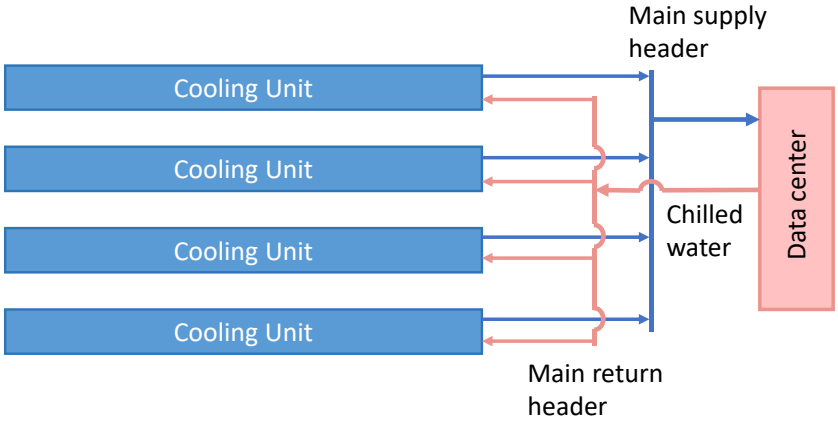


Fig. 5.4 Schematic of a data center cooling system in the baseline scenario.

It is worth noting that the cooling system COP discussed in this chapter focuses solely on the optimization of the cooling plant in the data center. Consequently, the energy consumption of the CRAH (Computer Room Air Handler) is not taken into account.

5.2.2 Description of economic analysis

According to the literature [199] and the data from the chiller manufacturer, the capital cost of chillers versus capacity is shown in Fig. 5.5 and Eq. (5.9). As the capacity of chillers increases, the capital cost per kilowatt gradually decreases and eventually stabilizes at approximately 150 $/kW. The capital costs for pumps [200] relative to their rated power (kW) are provided in Eq. (5.10).

The capital costs for cooling towers [200] and heat exchangers versus their capacities are expressed in Eq. (5.11) and Eq. (5.12). It is important to note that the capacity of the cooling tower, CP_{CT}, is rated air flow, m^3/h.

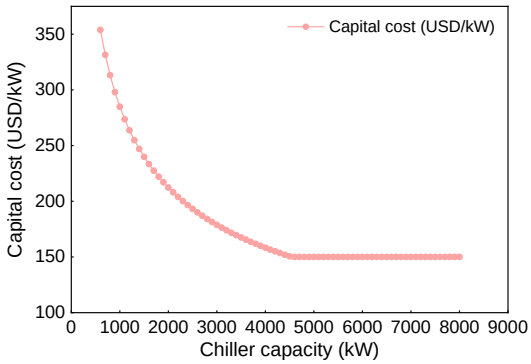


Fig. 5.5 The capital cost of chillers versus capacity

$$CC_{chiller} = -10^{-9}\, CP_{chiller}^{3} + 2 \times 10^{-5} CP_{chiller}^{2} - 0.132 CP_{chiller} + 400 \quad (5.9)$$

$$CC_{pump} = 1.63 \times CP_{pump}^{2} + 133.72 \times CP_{pump} + 500.37 \quad (5.10)$$

$$CC_{CT} = 29.85 \times CP_{CT} \quad (5.11)$$

$$CC_{HX} = 2 \times CP_{HX} \quad (5.12)$$

5.3 Results for optimized designs of cooling systems in different climate conditions

5.3.1 Energy and cost benefits under different numbers of cooling units (CPS control strategy)

Table 5.2 presents the optimal capacity combinations for 3, 4, 5, 6 and 7 cooling units and corresponding optimal PLRs under each IT loading stage when adopting the CPS control strategy in Climate Zone 4A.

Table 5.2 Optimal capacity combinations under different numbers of cooling units and optimal PLRs under different IT loadings (CPS)

	N = 3	N = 4	N = 5	N = 6	N = 7
			Optimal capacity combinations ((kW)		
	[5400, 6500, 6500]	[3400, 5000, 5000, 5000]	[3200, 3200, 3200, 3200, 5600]	[2400, 3200, 3200, 3200, 3200, 3200]	[2400, 2400, 2400, 2400, 2400, 3200, 3200]
			Optimal PLRs at different stages		
20% IT loading	[0.59, 0, 0]	[0.94, 0, 0, 0]	[1, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0]	[0, 0, 0, 0, 0, 1, 0]
30% IT loading	[0.89, 0, 0]	[0, 0.96, 0, 0]	[0, 0, 0, 0, 0.86]	[0.86, 0.86, 0, 0, 0, 0]	[1, 1, 0, 0, 0, 0, 0]
40% IT loading	[0, 0.98, 0]	[0.76, 0.76, 0, 0]	[1, 1, 0, 0, 0]	[0, 1, 1, 0, 0, 0]	[0, 0, 0, 0, 0, 1, 1]
50% IT loading	[0.67, 0.67, 0]	[0.95, 0.95, 0, 0]	[0.91, 0, 0, 0, 0.91]	[0.91, 0.91, 0.91, 0, 0, 0]	[1, 1, 0, 0, 0, 1, 0]

60% IT loading	[0.81, 0.81, 0]	[0, 0.96, 0.96, 0]	[1, 1, 1, 0, 0]	[0, 1, 1, 1, 0, 0]	[1, 1, 1, 1, 0, 0, 0]

Fig. 5.6(A) illustrates the cooling energy savings under different numbers of cooling units when adopting the CPS control strategy in Climate Zone 4A. The highest cooling energy saving, up to 24.7%, is observed when the number of cooling units is 7. Moreover, the optimized systems with 4, 5, and 6 cooling units all demonstrate more than 20% cooling energy savings.

On the other hand, Fig. 5.6(B) shows total cost savings under different numbers of cooling units. The most substantial reduction in life-cycle cost, up to 13.5%, is achieved when the number of cooling units is 4. The highest cooling energy savings are observed when the number of cooling units is 7. This is due to the fact that, for more cooling units, it is easier to accurately match a combination of cooling units in which each unit operates at its optimum efficiency, under different IT loading conditions. However, the capital cost for purchasing additional cooling equipment, particularly chillers, is also higher. Consequently, from the perspective of life-cycle total cost, the optimal number of cooling units is 4.

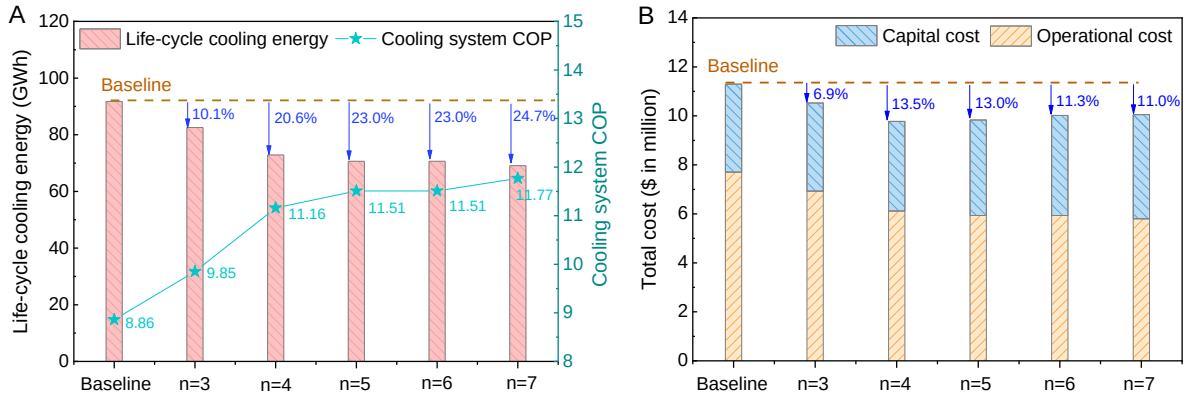


Fig. 5.6 Cooling energy savings (A) and total cost savings (B) under different numbers of cooling units when adopting the CPS control strategy in Climate Zone 4A

5.3.2 Energy and cost benefits under different numbers of cooling units (OPR control strategy)

Table 5.3 presents the optimal capacity combinations for 3, 4, 5, 6 and 7 cooling units when adopting the OPR control strategy in Climate Zone 4A. The optimal capacity combinations under different numbers of cooling units when adopting the OPR control strategy are nearly identical to the results obtained when adopting the CPS control strategy. The only difference is the capacity combination and PLRs when the number of cooling units is 4, where the optimized capacity combination is [4000, 4800, 4800, 4800] kW.

Fig. 5.7 illustrates the cooling energy savings and total cost savings under different numbers of cooling units when adopting the OPR control strategy in Climate Zone 4A. Similarly, the highest life-cycle cooling energy saving (i.e., 5.0%) is observed when the number of cooling units is 7, and the greatest reduction in life-cycle cost (i.e., 2.5%) is achieved when the number of cooling units is 4. The cooling energy savings when adopting the OPR control strategy are notably lower than those of the CPS control strategy. Moreover, the total costs increase instead by 1.2% and 4.2% when the number of cooling units of 6 and 7, respectively. This increase can be attributed to the additional capital cost incurred from purchasing more cooling equipment, which cannot be offset by energy-saving costs.

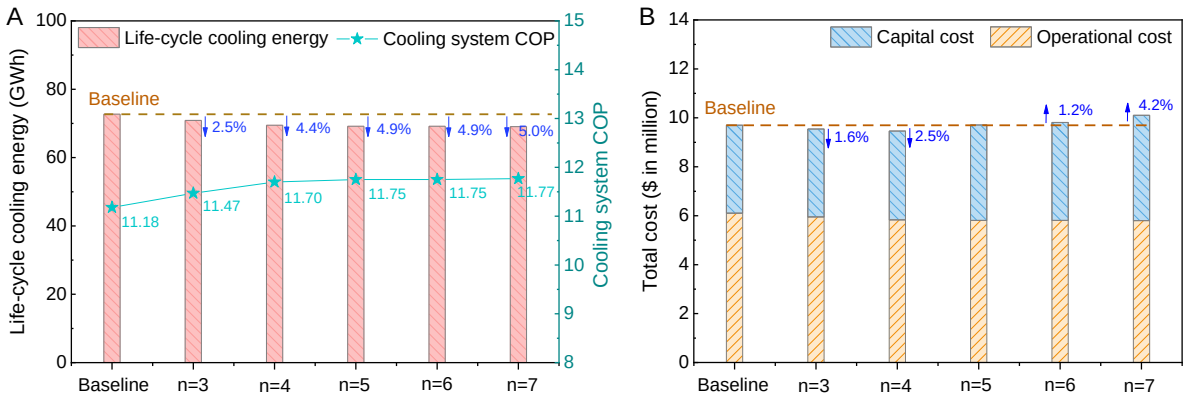


Fig. 5.7 Cooling energy savings (A) and total cost savings (B) under different numbers of cooling units when adopting the OPR control strategy in Climate Zone 4A

Table 5.3 Optimal capacity combinations under different numbers of cooling units and optimal PLRs under different IT loadings (OPR strategy)

	N = 3	N = 4	N = 5	N = 6	N = 7
	Optimal capacity combinations ((kW)				
	[5400, 6500, 6500]	[4000, 4800, 4800, 4800]	[3200, 3200, 3200, 3200, 5600]	[2400, 3200, 3200, 3200, 3200, 3200]	[2400, 2400, 2400, 2400, 2400, 3200, 3200]
	Optimal PLRs at different stages				
20% IT loading	[0.59, 0, 0]	[0.8, 0, 0, 0]	[1, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0]	[0, 0, 0, 0, 0, 1, 0]
30% IT loading	[0.89, 0, 0]	[0, 1, 0, 0]	[0, 0, 0, 0, 0.86]	[0.86, 0.86, 0, 0, 0, 0]	[1, 1, 0, 0, 0, 0, 0]
40% IT loading	[0, 0.98, 0]	[0.72, 0.72, 0, 0]	[1, 1, 0, 0, 0]	[0, 1, 1, 0, 0, 0]	[0, 0, 0, 0, 0, 1, 1]
50% IT loading	[0.67, 0.67, 0]	[0.91, 0.91, 0, 0]	[0.91, 0, 0, 0, 0.91]	[0.91, 0.91, 0.91, 0, 0, 0]	[1, 1, 0, 0, 0, 1, 0]
60% IT loading	[0.81, 0.81, 0]	[0, 1, 1, 0]	[1, 1, 1, 0, 0]	[0, 1, 1, 1, 0, 0]	[1, 1, 1, 1, 0, 0, 0]

5.3.3 Optimal capacity combinations in different climate conditions

In this case study, the optimal number of cooling units of 4 is also applicable in other climate zones. Table 5.4 and Table 5.5 present the optimal capacity combinations and corresponding PLRs in different climate zones. The optimal capacity combination under the CPS control strategy in different climate zones is consistent, i.e., [3400, 5000, 5000, 5000]. However, the optimal capacity under the OPR control strategy varies across different climate zones. Specifically, the optimal capacity combination under OPR control is [3400, 5000, 5000, 5000] in Zones 0A, 0B, 1A, 1B and 2A, [4000, 4800, 4800, 4800] in Zones 2B, 3A, 3B, 3C, 4A, 4B and 4C, [3700, 3700, 5500, 5500] in Zones 5A, 5B, 5C, 6A, 6B, 7 and 8. In Table 5.5, it can be observed that PLRs all fall within the range of 0.7-1.

Table 5.4 Optimized designs in different climate conditions

Zones	**Capacity combination (CPS)**	**Capacity combination (OPR)**
0A	[3400, 5000, 5000, 5000] kW	[3400, 5000, 5000, 5000] kW
0B	[3400, 5000, 5000, 5000] kW	[3400, 5000, 5000, 5000] kW
1A	[3400, 5000, 5000, 5000] kW	[3400, 5000, 5000, 5000] kW
1B	[3400, 5000, 5000, 5000] kW	[3400, 5000, 5000, 5000] kW
2A	[3400, 5000, 5000, 5000] kW	[3400, 5000, 5000, 5000] kW
2B	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
3A	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
3B	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
3C	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
4A	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
4B	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
4C	[3400, 5000, 5000, 5000] kW	[4000, 4800, 4800, 4800] kW
5A	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW
5B	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW
5C	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW
6A	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW
6B	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW
7	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW
8	[3400, 5000, 5000, 5000] kW	[3700, 3700, 5500, 5500] kW

Table 5.5 Optimal combinations and corresponding PLRs under progressive loading

Optimal results (kW)	[3400, 5000, 5000, 5000]	[4000, 4800, 4800, 4800]	[3700, 3700, 5500, 5500]
PLRs (20% loading)	[0.94, 0, 0, 0]	[0.8, 0, 0, 0]	[0.86, 0, 0, 0]
PLRs (30% loading)	[0, 0.96, 0, 0]	[0, 1.0, 0, 0]	[0, 0, 0.87, 0]
PLRs (40% loading)	[0.76, 0.76, 0, 0]	[0.72, 0.72, 0, 0]	[0.86, 0.86, 0, 0]
PLRs (50% loading)	[0.95, 0.95, 0, 0]	[0.91, 0.91, 0, 0]	[0.87, 0, 0.87, 0]
PLRs (60% loading)	[0, 0.96, 0.96, 0]	[0, 1.0, 1.0, 0]	[0, 0, 0.87, 0.87]

5.4 Energy and cost benefits of optimized designs in different climate conditions

5.4.1 Energy performance of the cooling unit under full-range loads and ambient temperature

The operation modes under full-range cooling loads and ambient air temperatures are shown in Fig. 5.8. It can be observed that the range of ambient air temperatures suitable for free cooling mode decreases as the part load ratio (PLR) increases. This is attributed to the "oversized" cooling towers designed for low PLRs, which have the capacity to handle low cooling loads even at higher wet-bulb temperatures than those of the design condition. Additionally, the range of ambient air temperatures suitable for partial free cooling mode decreases as the part load ratio (PLR) decreases. This is due to the fact that the chilled water return temperature under low PLRs is lower than the design condition, resulting in a reduced temperature difference between the chilled water return temperature and the cooling water supply temperature in heat exchangers. Consequently, the range of ambient air temperatures suitable for partial free cooling mode decreases.

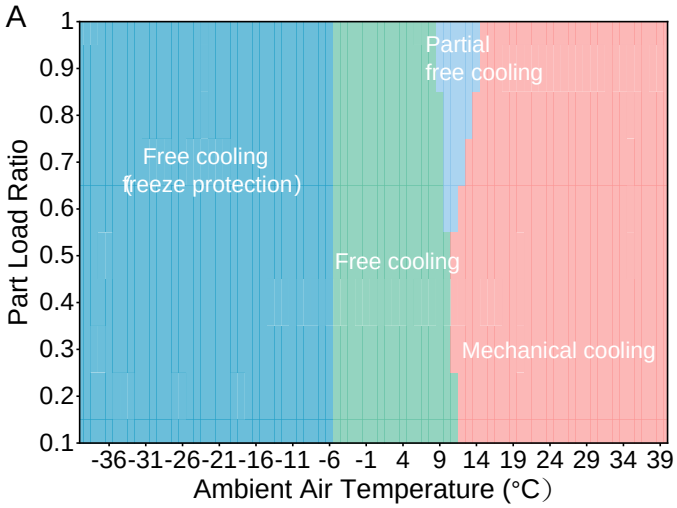


Fig. 5.8 Operation modes under full-range cooling loads and ambient air temperatures

The energy performance of the cooling system unit under full-range cooling loads and ambient air temperatures is shown in Fig. 5.9. Overall, the cooling system COP increases as the PLR increases. Higher COPs are observed at PLRs between 0.8 and 1. In addition, the cooling system COP is closely linked to the operation modes shown in Fig. 5.8. Furthermore, the cooling system COP, when adopting the OPR control strategy (Fig. 5.9(B)), outperforms that of the CPS control strategy (Fig. 5.9(A)) under off-design conditions. This is attributed to the more energy-efficient operation of chilled water pumps under the OPR control strategy, achieved by adjusting the flow rates of chilled water pumps in response to the cooling load.

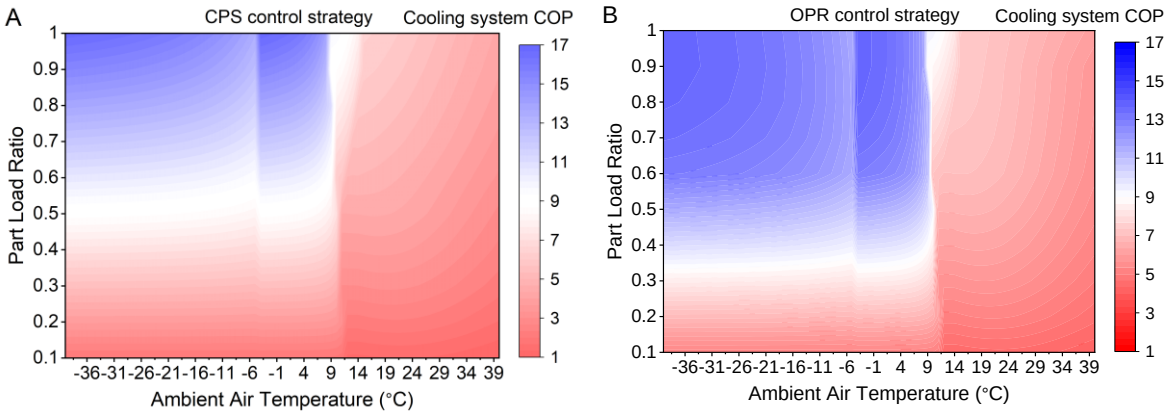


Fig. 5.9 Energy performance of the cooling unit under different ambient temperatures and part load ratios under CPS control strategy (A) and OPR control strategy (B)

5.4.2 Free cooling hour

Fig. 5.10 shows the impact of the optimal design on average free cooling hours annually worldwide. The optimal designs under both strategies show a decrease in free cooling hours. The average free cooling time is reduced by 13-860 hours annually under the CPS control strategy, and 13-735 hours annually under the OPR control strategy. The reduction can be attributed to the fact that the cooling units switch from low-PLR operation (i.e. 0.5-0.7) to high PLR operation (i.e. 0.8-1). As shown in Fig. 5.8, the range of ambient air temperatures suitable for free cooling mode decreases as the part load ratio (PLR) increases. Therefore, there is a decrease in free cooling hours when adopting optimal designs. When the CPS control strategy is adopted, a greater reduction in free cooling hours is observed compared to the OPR control strategy.

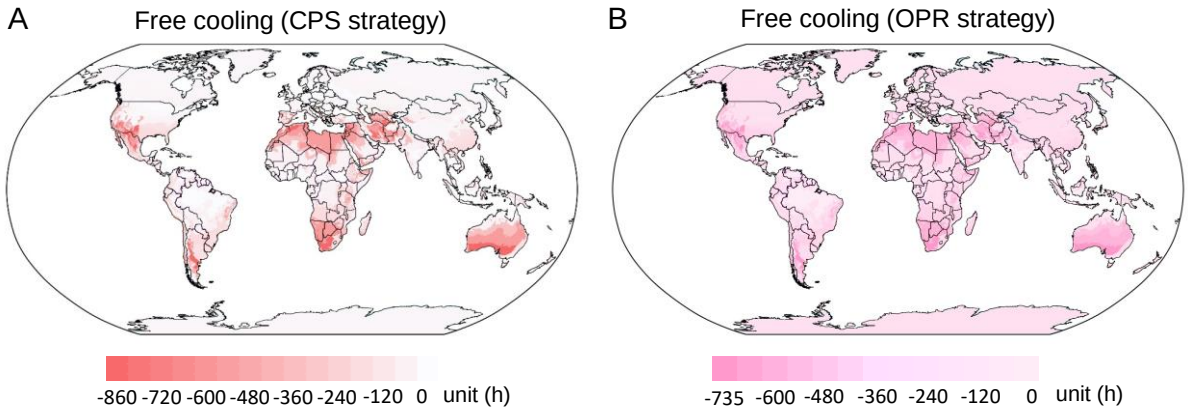
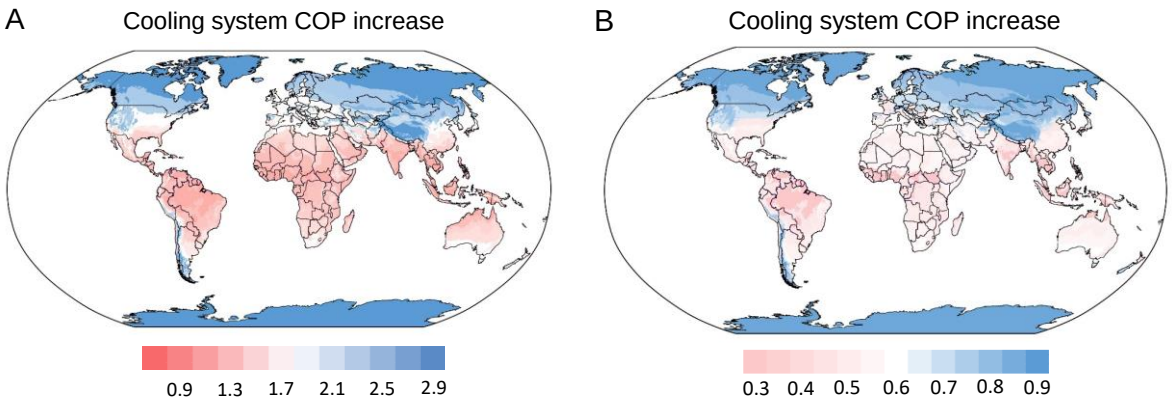


Fig. 5.10 Impact on free cooling hours worldwide under CPS (A) and OPR (B) control strategies

5.4.3 Cooling energy saving

Fig. 5.11 shows that worldwide cooling system COP increases when adopting the CPS control strategy (A) and OPR control strategy (B). The cooling system COP can be enhanced by 0.7-2.9 for the CPS control strategy and by 0.3-0.9 for the OPR control strategy, depending on climate conditions. Despite the decrease in free cooling hours, the overall cooling system can operate at higher system COPs over the data center lifespan for both control strategies. The reduced free cooling hours are transformed into partial free cooling hours. Although the system COP decreases during these reduced free cooling hours, the system COPs during other periods (i.e., other free cooling hours, mechanical cooling hours, and partial free cooling hours) increase. In addition, Fig. 5.11 shows that the increase in system COP in cold regions is higher than that in hot regions when adopting optimal design. Cold regions have more free cooling time year-round. Therefore, a higher increase in system COP in cold regions can be attributed to the fact that there is a higher increase in system COP in free cooling mode than in other modes.

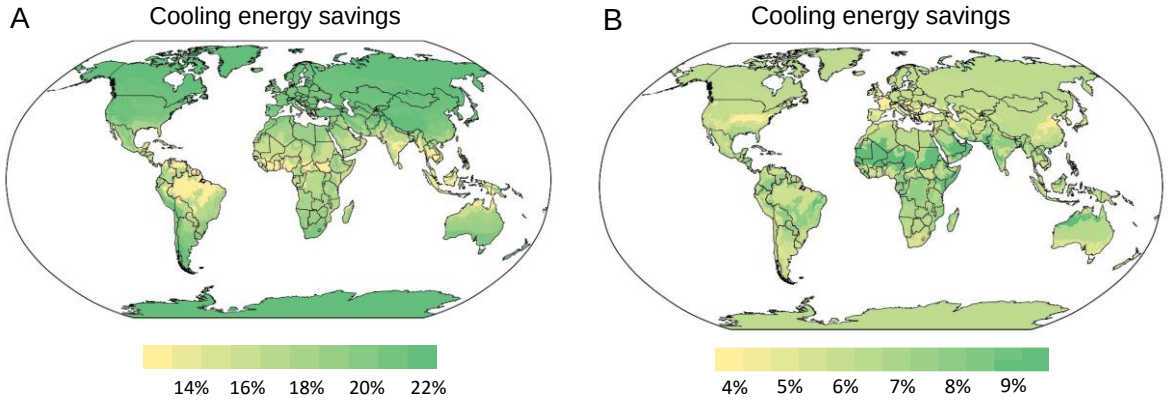


CPS control strategy OPR control strategy

Fig. 5.11 Worldwide system COP increase under CPS and OPR control strategies

Fig. 5.12 shows worldwide cooling energy savings when adopting the CPS control strategy (A) and OPR control strategy (B). The cooling energy can be saved by 13-22% for the CPS control strategy and 4-9% for the OPR control strategy, depending on climate conditions. It can be observed that cooling energy savings under the OPR control strategy are lower than those under the CPS control strategy. This is attributed to the fact that, while the cooling system operates efficiently under off-design conditions when the OPR control strategy is adopted, the improvement in cooling system COP is not as substantial as that achieved under the CPS control strategy.

Furthermore, more cooling energy savings are observed in cold regions when adopting the CPS control strategy, such as Climate Zones 5A (Fig. 5.12(A)). Whereas, more cooling energy savings are observed in hot regions when adopting the OPR control strategy, such as Climate Zones 0B and 1A (Fig. 5.12(B)). This can be attributed to the similar increases in cooling system COP in most climate zones. For example, the increase in cooling system COP is 0.6 in Zone 0B and 0.7 in Zone 5A. However, the cooling system COP in the baseline scenario in climate Zone 0B is 6.19, and 11.03 in Zone 5A. Therefore, an increase of 0.6 over 6.19 (Zone 0B) could show a higher percentage of cooling energy savings, compared to an increase of 0.7 over 11.03 (Zone 5A).



CPS control strategy OPR control strategy

Fig. 5.12 Worldwide cooling energy savings under CPS and OPR control strategies

5.4.4 Data Center PUE

Fig. 5.13 shows the worldwide PUE reduction under the CPS (A) and OPR (B) control strategies. The datacenter PUE can be reduced by 0.06-0.1 under the CPS control strategy, and 0.022-0.044 under the OPR control strategy (assume the baseline PUE of 1.58, the average in 2023 [201]). The reduction in data center PUE is closely related to cooling energy savings. Similar to the cooling energy savings, there is a more significant reduction in PUE in cold regions when adopting the CPS control strategy, while there is a more significant reduction in PUE in hot regions when adopting the OPR control strategy.

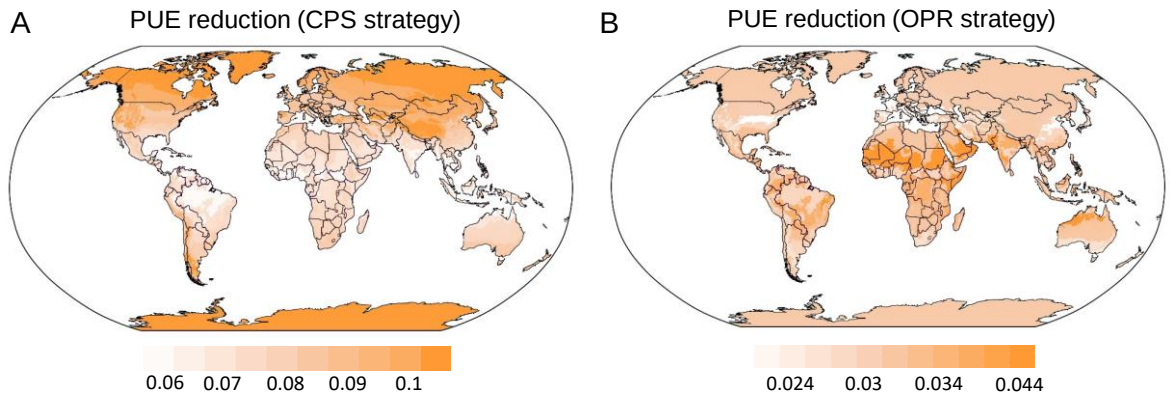


Fig. 5.13 PUE reduction worldwide under CPS (A) and OPR (B) control strategies

5.4.5 Total cost

Fig. 5.14 shows the impact of the optimal design on total cost savings worldwide. Under the CPS control strategy (Fig. 5.14(A)), total cost savings of 9.7-13.8% can be achieved, while under the OPR control strategy, the savings range from 2.5-6.4%, depending on climate conditions. These cost savings are primarily attributed to the reduction in cooling energy savings, but not capital cost as the optimized number of cooling units remains the same as in the baseline design. Furthermore, the total cost savings are relatively consistent across climate zones in both strategies, as the capital accounts for nearly one-third of the total cost.

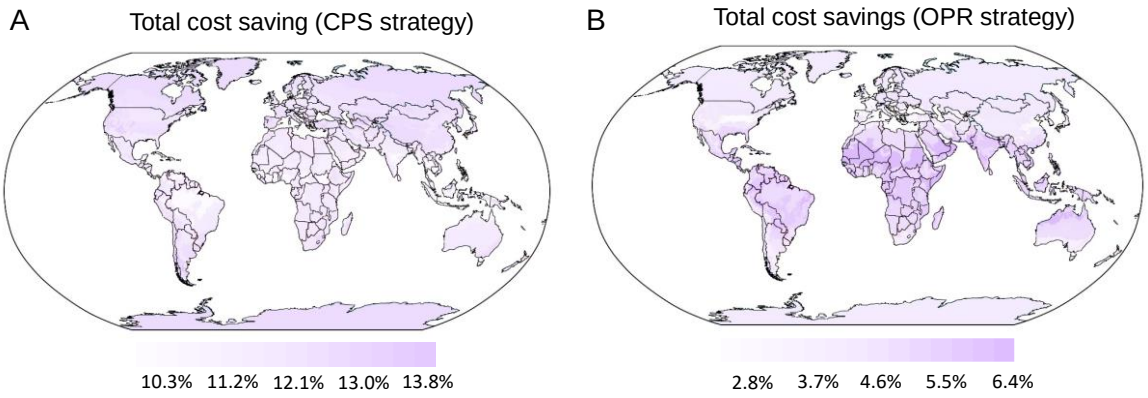


Fig. 5.14 Total cost savings worldwide under CPS (A) and OPR (B) control strategies

5.5 Summary

This chapter focuses on the enhancement of the energy efficiency of data center cooling systems from the life-cycle perspective. The main contributions of this chapter include: *i)*, proposing an optimal design method for centralized cooling systems suitable for any type of progressive loading profile. *ii)*, presenting and analyzing the energy performance of the cooling system under two typical control strategies under full-range cooling loads and ambient air temperatures. *iii)*, identifying the optimal designs in different climate zones according to

energy performance results. *iv)*, analyzing and comparing comprehensively the free cooling hours, cooling energy, life-cycle total cost and PUE reductions of the optimized designs with conventional designs. *v)*, quantifying the energy and cost benefits in different climate conditions. The key findings of this study are as follows.

Although the highest cooling energy savings are achieved when the number of cooling units is 7, the capital cost for purchasing additional cooling equipment, particularly chillers, is higher. From the perspective of life-cycle cost, the optimal number of cooling units is determined to be 4.

Under the CPS control strategy, the worldwide cooling system COP can be enhanced by 0.7-2.9, and the corresponding cooling energy can be saved by 13-22%, depending on climate conditions. The total cost savings worldwide are estimated to be 9.7-13.8% and data center PUE can be reduced by 0.06-0.1. Despite a decrease in free cooling hours (i.e., 13-860), the cooling system operates more energy-efficiently over its lifespan when adopting the optimized design.

For the near-optimal control strategy OPR, there are still 4-9% cooling energy savings, a 0.3-0.9 increase in cooling system COP, and 2.5-6.4% cost savings over the data center lifetime. This highlights the importance of both optimal design and optimal control strategy for efficient operation over the data center life cycle.

The results provide valuable insights into the energy efficiency of centralized cooling systems with multiple chillers for air-cooled data centers under life-cycle operations and guide the development of next-generation high-efficiency cooling systems for data centers.

CHAPTER 6 UNLOCKING THE FLEXIBILITIES OF DATA CENTERS IN SMART GRID MARKETS: OPTIMAL DISPATCH AND DESIGN OF ENERGY STORAGE SYSTEMS CONSIDERING PROGRESSIVE LOADING

This chapter presents a pioneering approach that leverages the surplus capacity of energy storage systems for emergencies in data centers to participate in flexible grid services, considering progressive loading throughout their lifecycle. Chapter 6.1 presents the formulation of optimization problems considering the effective use of surplus capacity of storage systems in flexibility markets. Chapter 6.2 introduces the electricity markets, proposed design scenarios and the mathematical formulas for economic analysis. Chapter 6.3 presents the specifications of energy storage and cooling systems in the referenced data center. Chapter 6.4 presents optimal dispatch results utilizing surplus capacity in data centers. Chapter 6.5 quantifies and discusses the life-cycle economic benefits under typical electricity markets. Chapter 6.6 analyzes the impacts of discount rate and battery price on life-cycle economic benefits.

6.1 Formulation of optimization problems considering effective use of surplus capacity of storage systems in flexibility markets

6.1.1 Typical progressive IT loading and surplus capacity

The IT load within data centers is not static but rather experiences a progressive increase over their lifetime, as illustrated in Fig. 6.1 [192]. Initially, the initial load is estimated to be approximately 20% of the design capacity, progressively increasing in stages to ultimately reach around 60% of the design capacity. The load increase follows several stages: 20% in the initial stage, 30% in the second stage, 40% in the third stage, 50% in the fourth stage, and ultimately 60% in the fifth stage.

Typical emergency designs for data centers include 15-minute cold energy storage and 15-minute battery storage systems, designed to support 100% of the IT load [202]. These systems are essential for ensuring an uninterrupted power supply in the event of an emergency or power outage. In the context of progressive loading, there is remaining energy storage capacity for emergencies in data centers that can be flexibly scheduled at each stage throughout the data center's lifecycle, without compromising the reliability of the data center. Data centers can

leverage the surplus energy storage capacity to provide grid services without requiring additional investments to stabilize the grid and generate revenues, creating a win-win situation.

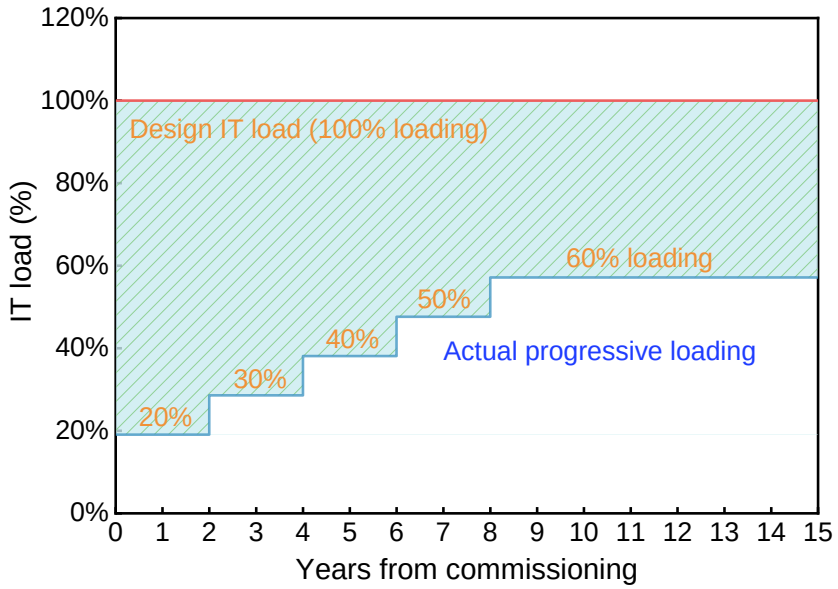


Fig. 6.1 Typical progressive IT loading of data centers

6.1.2 Typical electricity markets and flexibility services

Pricing mechanisms vary across countries or regions that are managed by different independent system operators. For instance, the electricity market in the United States, specifically CAISO, is managed by an independent system operator and operates using a joint market model, allowing for the simultaneous trading of energy and ancillary services. In contrast, China's Time of Use (ToU) electricity market utilizes a time-based pricing system, which leverages differential electricity prices during different periods to encourage users to shift their electricity consumption from peak hours, thereby alleviating peak load pressure on the power system.

This study selects these two electricity markets as backgrounds to analyze the potential benefits of the effective use of the surplus capacity in energy storage systems in data centers. One is the CAISO (California Independent System Operator) electricity market in the United States, which offers various services for demand-side participants. The historical data in 2020 from CAISO's energy markets is obtained from the official website [203].

The other electricity market is the Time of Use (ToU) electricity market in Guangdong Province, China. In a typical three-period ToU tariff, electricity prices are divided into peak, flat, and valley prices to incentivize users to shift their demand from peak hours to off-peak hours, thereby alleviating the burden of peak power generation for the power system [204]. The data for the Time of Use (ToU) electricity market in Guangdong Province is obtained from

the official website [205], and the rewards of grid services in China's electricity market are obtained from [206].

Demand-side flexibility involves the capacity to modify consumption patterns while engaging with the power grid. Energy flexibility in data centers for the grid includes energy arbitrage and ancillary services, such as frequency regulation and spinning reserve. Energy arbitrage involves leveraging flexible resources to transfer energy use from peak-price times to off-peak times, resulting in economic benefits. Ancillary services require flexible resources to respond rapidly to ensure a short-term balance in power distribution. For instance, in the process of frequency regulation, system authorities measure the power imbalance as the area control error (ACE), which is then converted into normalized automatic generation control (AGC) signals that vary from -1 to 1. These signals are continuously transmitted to service providers at second intervals. The spinning reserve requires resources to be synchronized to the grid and respond within 10 minutes. Participants on the demand side offering spinning reserve service must be capable of swiftly curtailing their load for a short period upon an urgent request from the power grid.

6.1.3 Formulation for optimizing dispatching energy flexibilities

Fig. 6.2 illustrates the flowchart of dispatch optimization of surplus energy storage in data centers. The optimization objective is to minimize operational costs, which involve energy arbitrage and revenues from frequency regulation and spinning reserve services. Each dispatch optimization has a 24-hour horizon with a time step of 1 hour. Optimization variables include the hourly rates of discharging and charging for the battery and TES tank, along with the hourly capacities allocated for ancillary services. The optimization objective of dispatch optimization of surplus energy storage in data centers is presented as follows.

$$\text{Min} \sum_{k\in T}(\pi_k^E P_k^E h - \pi_k^{SR} Cap_k^{SR} - \pi_k^{FR} Cap_k^{FR}) \forall\ k \in [1,24] \tag{6.1}$$

where, $\pi_k^E, \pi_k^{SR}, \pi_k^{FR}$ are the energy price, the revenues for providing spinning reserve service and the frequency regulation service at the hour k. P_k^E is the total energy consumption at the hour k. Cap_k^{SR} is the provided spinning reserve capacity at the hour k. Cap_k^{FR} is the provided frequency regulation capacity at the hour k.

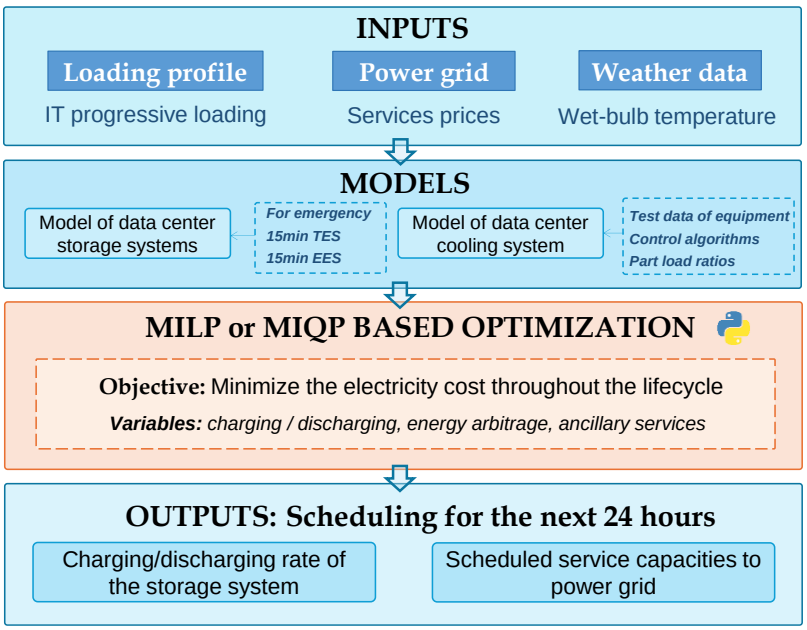


Fig. 6.2 Flowchart of dispatch optimization of surplus energy storage in data centers

Optimal dispatch strategy of electrical energy storage (EES)

In addition to leveraging energy arbitrage, batteries are perfectly suited for providing frequency regulation and spinning reserve services due to their fast response speed. The optimization objective for dispatching surplus capacity in EES is presented in Eq. (6.2). Formulations (6.3)-(6.4) show operating constraints associated with scheduling capacity in market services. Notably, it is assumed that frequency regulation has an ideally neutral effect on hourly energy consumption (kWh). This assumption is based on the design of the regulation signal, which is intended to average out to zero over a specified timeframe (e.g., 1 hour), following the guidelines set by the Independent System Operator (ISO) [207]. The defined minimum and maximum values of the state of charge of batteries are presented in Formulation (6.7). Establishing appropriate upper and lower boundaries can significantly mitigate batteries degradation.

The battery's operational depth of discharge (DoD) has a significant impact on its lifecycle. Here, we limit the DoD to below 60%, as exceeding this threshold can lead to a shorter lifespan for the battery. When DoD is set at 60%, the losses incurred from participating in grid regulation services throughout the life cycle are nearly equivalent to the losses experienced when leaving the energy storage systems unused [208, 209].

$$P_k^E = P_{EES,k}^{\wedge} - P_{EES,k}^{\vee} \tag{6.2}$$

$$P_{EES,k}^{\wedge} + Cap_{EES,k}^{FR} \leq P_{EES,rated} \tag{6.3}$$

$$-P^{\vee}_{EES,k} - Cap^{FR}_{EES,k} - Cap^{SP}_{EES,k} \geq -P_{EES,rated} \tag{6.4}$$

$$P_{EES,rated} = E_{EES,rated} / r_{ep} \tag{6.5}$$

$$SOC_{EES,k+1} = SOC_{EES,k} + (P^{\wedge}_{EES,k}\eta_{EES} - P^{\vee}_{EES,k}/\eta_{EES})/ E_{EES,rated} \tag{6.6}$$

$$\underline{SOC}_{EES} \leq SOC_{EES,k} \leq \overline{SOC}_{EES} \tag{6.7}$$

where, $P^{\wedge}_{EES,k}$ and $P^{\vee}_{EES,k}$ are the charging and discharging states of batteries at the hour k, respectively. When the storage is being discharged at $P^{\vee}_{EES,k}$, the charging power $P^{\wedge}_{EES,k}$ is equal to 0 and vice versa. $P_{EES,rated}$ is the rated charging/discharging rate. $E_{EES,rated}$ is rated battery capacity. r_{ep} is energy to power ratio. $SOC_{EES,k}$ is the state of charge at the hour k. η_{EES} is the battery charging/discharging efficiency. $\underline{SOC}_{EES}$ and $\overline{SOC}_{EES}$ are lower and upper of the battery, 0.2 and 0.8, respectively.

Optimal dispatch strategy of thermal energy storage (TES)

Cooling systems integrated with TES typically have a slower response speed in terms of power adjustment (minutes to hours). Therefore, TES cannot provide frequency regulation services to precisely track the grid control signals, which require faster response speed in the timescale of seconds. However, TES has the capability to offer significant power reduction and shifting capacity within hours. In the optimization of the dispatch of TES capacity, only spinning reserve service and energy arbitrage are taken into account. Eq. (6.8) presents the optimization objective for the dispatching surplus capacity in TES. Formulations (6.12)-(6.14) show the operating constraints associated with the flexible scheduling of TES capacity.

$$P^{E}_{k} = \frac{Q_{dem,k} + Q_{TES,k}}{COSP_k} \tag{6.8}$$

$$COSP_k = f(PLR_k, T_{wet,k}) \tag{6.9}$$

$$PLR_k = \frac{Q_{dem,k} + Q_{TES,k}}{N_{op} * Q_{EC,rated}} \tag{6.10}$$

$$Q_{store,k} = \eta_{TES} Q_{store,k-1} + Q_{TES,k} \tag{6.11}$$

$$0.3 \leq PLR_k \leq 1 \tag{6.12}$$

$$\left|Q_{TES,k}\right| \leq Q_{TES,rated} \tag{6.13}$$

$$0 \leq Q_{store,k} \leq Q_{TES,rated} \tag{6.14}$$

$$0 \leq Cap^{SP}_{TES,k} \leq \min(Q_{store,k}, Q_{EC,k}, Q_{TES,rated} + Q_{TES,k}) /COSP_k \tag{6.15}$$

where, $Q_{dem,k}$ is the cooling demand at the hour k. $Q_{TES,k}$ is the charging/discharging of TES at the hour k. $COSP_k$ is the coefficient of performance of the cooling system at the hour k. PLR_k is the part load ratio of the cooling system. $T_{wet,k}$ is the wet-bulb temperature. $Q_{store,k}$ is the thermal energy storage at the hour k. $Q_{EC,rated}$ is the rated cooling capacity of the electric chiller. η_{TES} is TES tank's storage efficiency. N_{op} is the quantity of chillers in operation.

6.1.4 Description of design scenarios for energy storage systems

Fig. 6.3 presents the energy storage profiles of the baseline scenario and design scenarios 1-4 throughout their lifetimes. In the baseline scenario, the energy storage system for emergencies is a one-time investment and does not participate in the grid. *In Scenario 1*, the emergency storage system is a one-time investment and provides auxiliary services to the grid throughout the data center's lifecycle, utilizing surplus energy storage capacity. *In Scenario 2*, the energy storage system for emergencies is a phased investment based on progressive IT loading. There is no additional dispatchable capacity to provide auxiliary services to the grid. *In Scenario 3*, the energy storage system for emergencies is a phased investment based on progressive IT loading, with an additional 20% capacity corresponding to the progressive loading. The extra 20% capacity at each stage can be flexibly scheduled to provide auxiliary services. *In Scenario 4*, the energy storage system for emergencies is a phased investment based on progressive IT loading, with an additional 40% capacity corresponding to the progressive loading. The extra 40% capacity at each stage can be flexibly scheduled to provide auxiliary services.

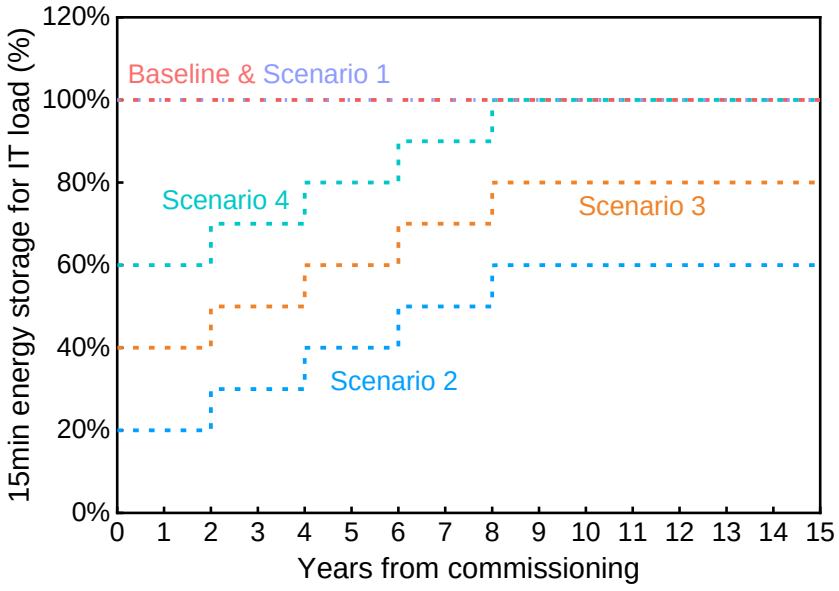


Fig. 6.3 Surplus capacities of different scenarios throughout the lifecycle

6.1.5 Formulation for optimizing storage system design in data centers

Typically, life-cycle economic analysis is used as an effective means to determine the most cost-effective solution from multiple options by comparing their economic benefits over their lifetime. The objective of optimizing storage system design is to minimize the life-cycle cost.

$$LCC = C_{INV} + C_{REP} + \sum_{i=1}^{L} \frac{(C_{O\&M,i} - Rev_i)}{(1+r)^i} \tag{6.16}$$

where, LCC is the life-cycle cost, C_{INV} is the total investment cost throughout the lifetime, C_{REP} is the total replacement cost throughout the lifetime, $C_{O\&M,i}$ is the maintenance cost at ith year, Rev_i is the total revenue at ith year, r is the discount rate. L is the lifetime. i is the ith year after the investment.

For electrical energy storage, the investment and replacement costs at ith year are given in Eq. (6.17)-(6.18).

$$C_{INV,i} = \frac{C_{INV,0}}{(1+r)^i} * d \tag{6.17}$$

$$C_{REP,i} = \frac{C_{INV,0}}{(1+r)^i} * d \tag{6.18}$$

where, d is the annual decline rate of battery price.

For thermal energy storage, the investment cost at ith year is shown as follows.

$$C_{INV,i} = \frac{C_{INV,0}}{(1+r)^i} \tag{6.19}$$

Fig. 6.4 illustrates the schematic of the economic performance analysis of energy storage systems. The economic performance is assessed using the life-cycle cost (LCC) indicator, as shown in Eq (6.16). The investment costs include initial costs, operation and maintenance costs, and replacement costs. Economic benefits are maximized by leveraging multiple revenue streams through optimized dispatch strategies. Investment, operational costs and revenues are discounted over the data center's lifetime using an assumed discount rate. Additionally, the annual decline rate of battery price is considered in the economic analysis.

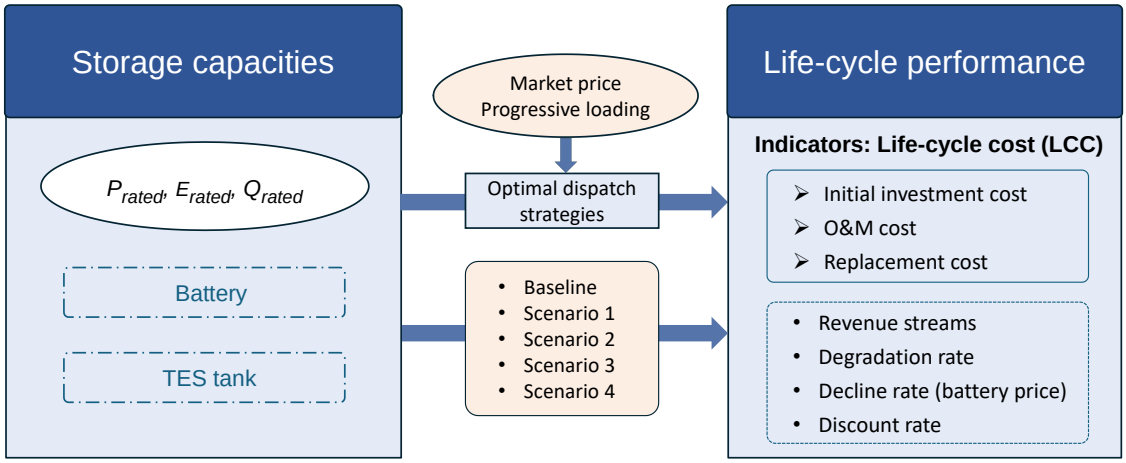


Fig. 6.4 The framework of life-cycle analysis

6.2 Description of energy storage systems in the referenced data center

In the referenced data center [129], the design IT load is 16800 kW. The standard design for both EES and TES is to provide enough electrical power or cooling energy to cover the design IT load for 15 minutes in case of an emergency [202]. Table 6.1 shows the specifications of EES and TES in the referenced data center. The cooling system is a centralized cooling plant with multiple chillers, water-side economizers and TES tanks. The specification of the cooling system is shown in Table 4.1.

Table 6.1 Specifications of energy storage systems in the referenced data center

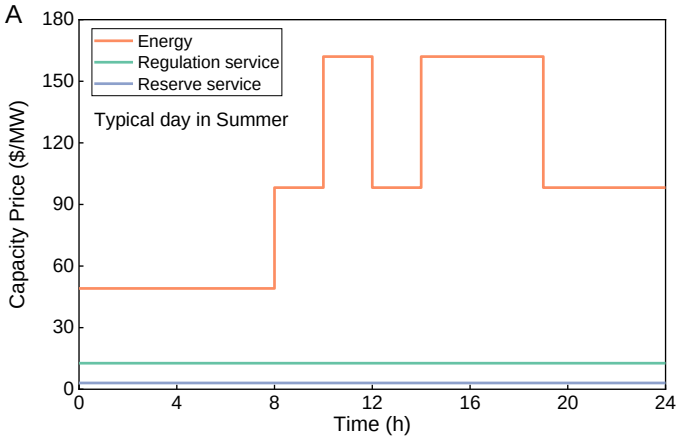
TES/EES	Parameter
TES capacity	600 m^3
TES capacity cost	31.8 $/kWh [210]
TES O&M cost/year	0.7% of investment cost [211]
TES lifespan	20 years [210]
TES energy storage efficiency	0.995 [212]
Battery capacity	4200 kWh
Battery capacity cost	200 $/kWh [213]
Battery O&M cost/year	0.5% of investment cost [211]
Battery cycle (to end-of-life)	9000 [208]
Battery float lifespan	10 years [214]
Battery charging/discharging efficiency	0.95 [208]
Battery Energy-to-Power ratio	2.8 [208]
r (discount rate)	4% [215]
d (annual decline rate of battery price)	5% [213]

6.3 Optimal dispatch results utilizing surplus capacity in data centers

The dispatch optimizations for both EES and TES are programmed in Python, and the optimization models are solved using the Gurobi optimization solver. The dispatch optimization for EES is formulated as a mixed-integer linear programming problem. For the cooling system, the power consumption and cooling supply are represented as a piecewise-quadratic function based on regression analysis. The dispatch optimization for TES is addressed as a mixed-integer quadratic programming problem.

6.3.1 Optimal dispatch results under the Guangdong electricity market in China

Fig. 6.5 shows the Time of Use (ToU) electricity market in Guangdong Province, China, where electricity prices are categorized into peak, flat, and valley rates daily. Fig. 6.5(B) shows the optimized hourly dispatch results of electrical energy storage (EES), including charging/discharging as well as capacities provided for frequency regulation and spinning reserve. Fig. 6.5(C) shows the optimized hourly dispatch results of thermal energy storage (TES), including charging/discharging as well as the capacity provided for spinning reserve. It can be observed that the charging and discharging occur at low and high tariffs, respectively, thereby facilitating energy arbitrage. Furthermore, since the revenues generated from regulation services typically exceed those from operating reserves, the surplus capacity of EES is fully allocated to provide frequency regulation services. Similarly, for Thermal Energy Storage (TES), charging takes place during periods of low tariffs, while discharging is timed with high tariffs to achieve energy arbitrage. In addition, TES is capable of providing reserve capacity as long as the stored energy in the TES tank remains above zero.



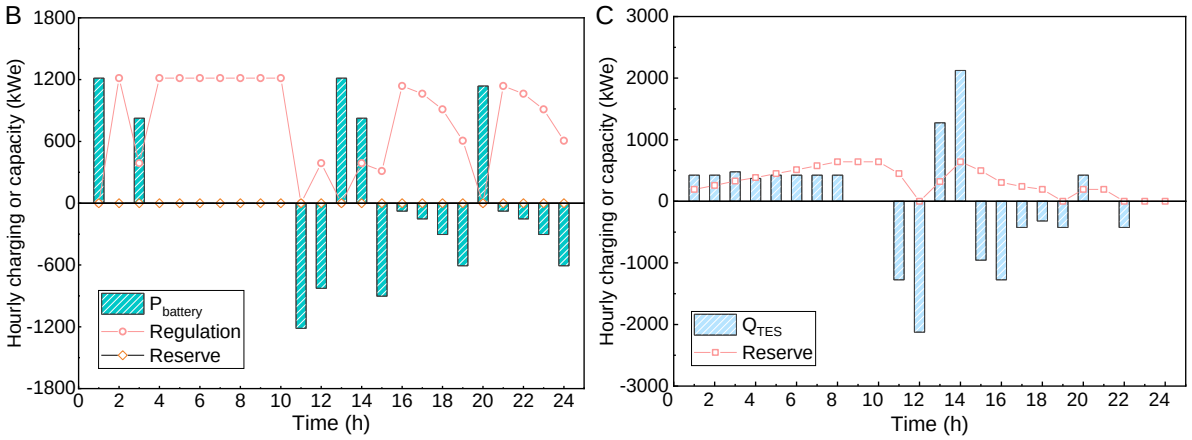
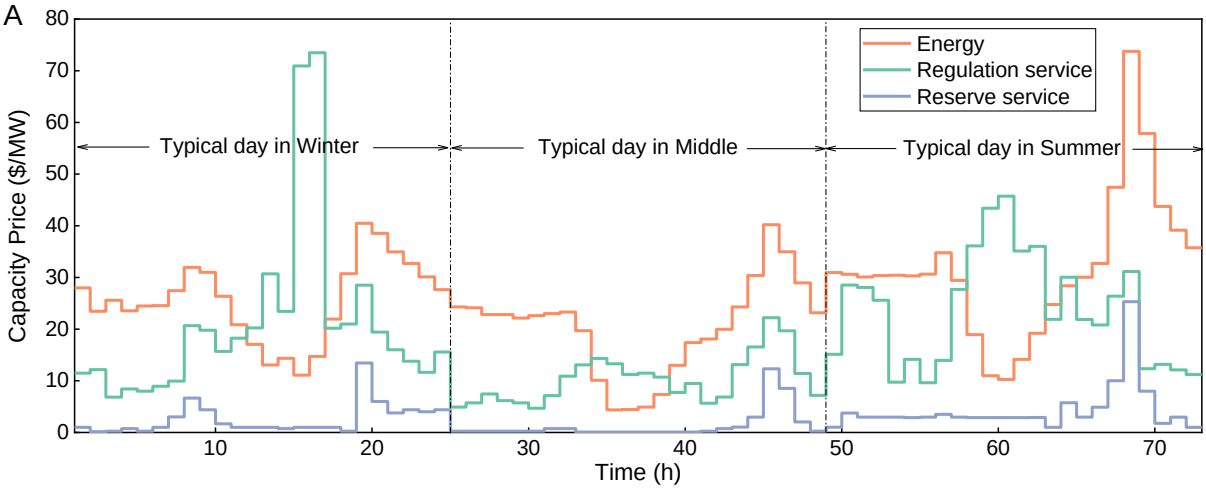


Fig. 6.5 (A) Time of Use (ToU) energy, regulation and operating reserve prices in Guangdong, China; Optimal dispatch results of (B) EES (C) TES.

6.3.2 Optimal dispatch results under the CAISO electricity market in the US

Fig. 6.6(A) shows hourly energy, frequency regulation and operating reserve prices of three typical days from the CAISO electricity market in the US. Fig. 6.6(B) shows the optimized hourly dispatch results of electrical energy storage (EES), including charging/discharging as well as capacities provided for frequency regulation and spinning reserve. It can be observed that the majority of surplus battery capacity is used to provide frequency regulation service to generate revenue, with only a minor fraction being allocated for charging and discharging for energy arbitrage. This can be primarily attributed to the higher financial incentives associated with providing frequency regulation compared to those for operating reserve and energy arbitrage. Fig. 6.6(C) shows the optimized hourly dispatch results of thermal energy storage (TES), including charging/discharging and the capacity provided for spinning reserve. It can be observed that the charging and discharging occur at low and high tariffs, respectively, thereby facilitating energy arbitrage. Similarly, TES can provide reserve service as long as the stored energy in the TES tank remains above zero.



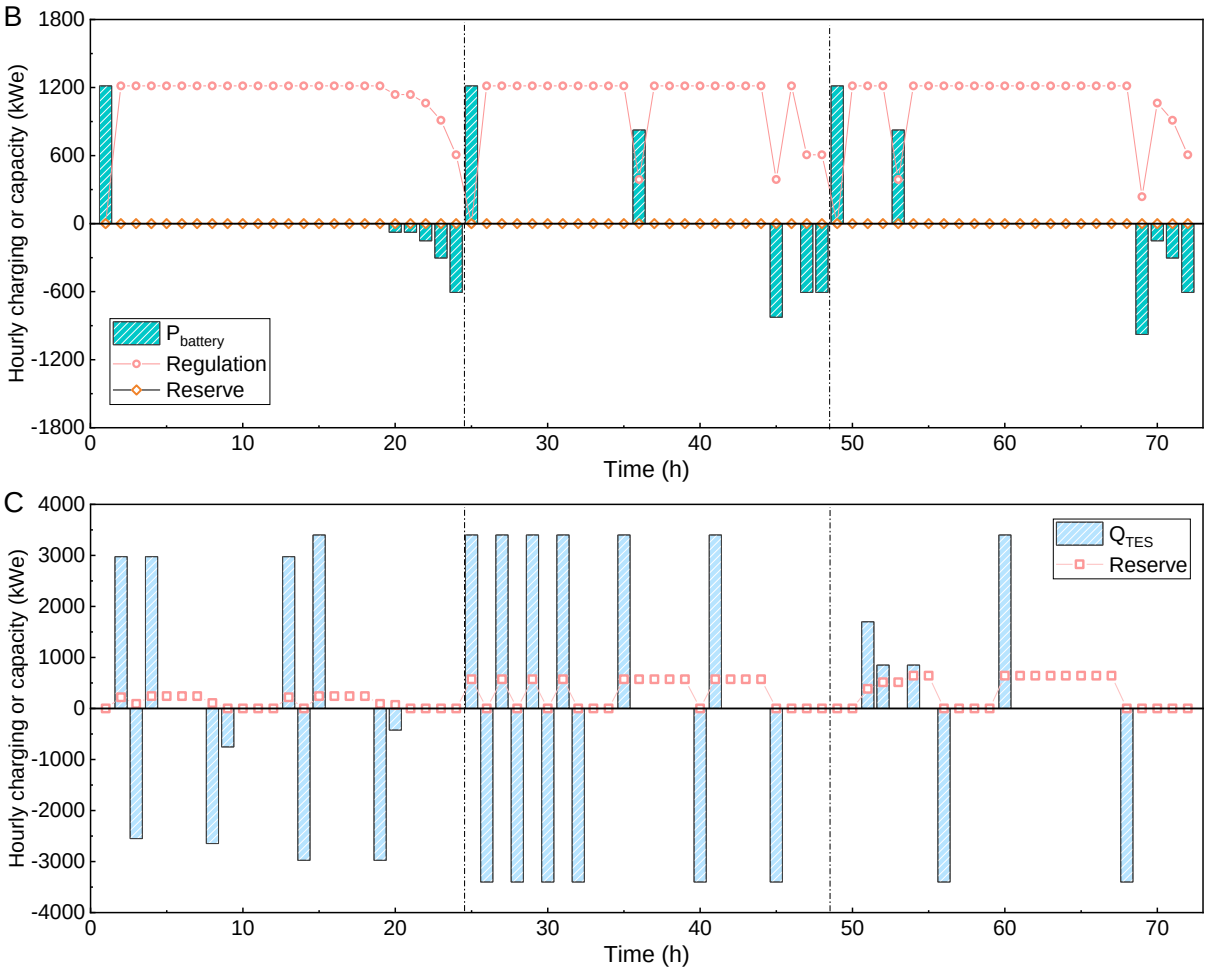


Fig. 6.6 (A) Hourly energy, frequency regulation and operating reserve prices in 01/25, 04/25, and 09/27 from CAISO in the US; Optimal dispatch results in three typical days of (B) EES (C) TES.

6.4 Life-cycle economic benefits under different electricity markets

6.4.1 Life-cycle economic benefits under the Guangdong electricity market in China

Fig. 6.7(A) shows the life-cycle economic benefits of different scenarios of EES under the Guangdong electricity market. Due to multiple revenues from energy arbitrage and the provision of auxiliary services to the electricity market, the life-cycle economic benefits of Scenarios 1-4 all show better economic performance compared to the baseline scenario. Notably, the revenue from energy arbitrage accounts for a major portion, while the revenue from the provision of frequency regulation to the electricity market accounts for a minor portion of the Guangdong electricity market. Furthermore, Scenario 1 emerges as the most favorable, achieving higher profits ($-86,418) compared to the other scenarios (Scenario 1: the energy storage system for emergencies is a one-time investment and provides auxiliary services to the grid throughout the data center's lifecycle). In addition, it can be observed that there are

significant differences in capital costs across scenarios, primarily attributed to the discount rate and annual decline of battery price. This indicates that a staged investment results in a lower total investment amount when compared to a one-time investment.

Fig. 6.7(B) shows the life-cycle economic benefits of different scenarios of TES under the Guangdong electricity market. It can be observed that Scenarios 1, 3 and 4 all yield positive profits over the lifetime, attributed to the revenues from energy arbitrage, the provision of reserve capacity and discount rates. Notably, Scenario 1 is still the most advantageous design option, yielding the highest profits, $205,213. This can be attributed to more revenue streams from energy arbitrage and the provision of reserve capacity to the grid. These revenue streams from the grid are higher than the economic benefits from a staged investment adopted in Scenarios 2-4.

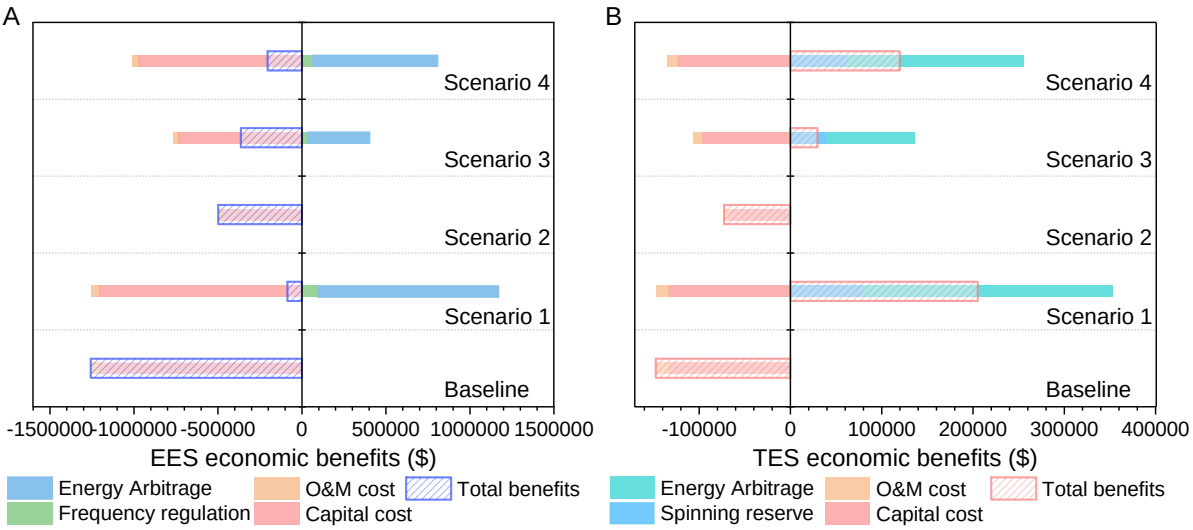


Fig. 6.7 Life-cycle economic benefits of different scenarios of (A) EES, (B) TES under the Guangdong electricity market, China

6.4.2 Life-cycle economic benefits under the CAISO electricity market in the US

Fig. 6.8(A) shows the life-cycle economic benefits of different scenarios of EES under the CAISO electricity market. Unlike the results under the Guangdong electricity market, both Scenario 1 and Scenario 4 yield positive profits over the lifetimes under the CAISO electricity market. Similarly, Scenario 1 emerges as the most profitable design option for EES, yielding the highest economic benefit, $361,453. This indicates that the revenues from energy arbitrage and the provision of frequency regulation services to the grid are significantly higher compared to the Guangdong electricity market. Contrary to the results in the Guangdong electricity market, under the CAISO market, the majority of the revenue comes from providing frequency

regulation services, while energy arbitrage contributes a smaller portion. The results differ significantly from those under the Guangdong electricity market. The reward associated with frequency regulation services under the CAISO electricity market is significantly higher than under the Guangdong electricity market.

Fig. 6.8(B) shows the life-cycle economic benefits of different scenarios of TES under the CAISO electricity market. Unlike the results under the Guangdong electricity market, only Scenario 1 yields a positive profit, of $36,985. This profitability is attributed to multiple revenue streams, including energy arbitrage and the provision of reserve capacity to the grid. Despite the presence of similar revenue streams in Scenarios 2-4, such as energy arbitrage, the provision of reserve capacity, and considerations of discount rates, these scenarios fail to achieve positive profits. The inability to generate positive profits in Scenarios 2-4 is primarily due to the fact that the revenue streams from the grid do not outweigh the economic benefits of a staged investment strategy implemented in these scenarios.

Overall, the results indicate that energy storage systems (ESS) designed for emergencies can yield positive profits through participation in grid interactions. Under both electricity markets, Scenario 1 emerges as the optimal design option for deploying EES and TES for emergency use in data centers.

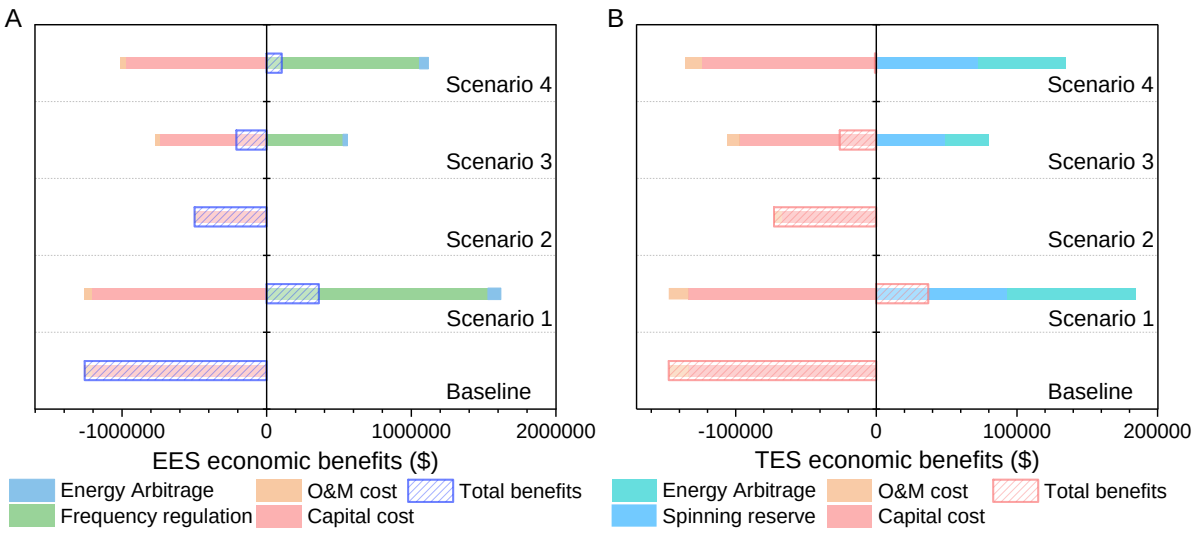


Fig. 6.8 Life-cycle economic benefits of different scenarios of (A) EES, (B) TES under CAISO electricity market

6.5 Impacts of discount rate and battery price on life-cycle economic benefits

6.5.1 Impacts of discount rate and battery price on life-cycle economic benefits of EES

When the discount rate is set at 4% and the annual decline rate of battery price is 5%, Scenario 1 is identified as the optimal design option for deploying Energy Storage Systems (EES) and Thermal Energy Storage (TES) for emergency use in data centers. However, the results might be significantly different when these two critical factors vary.

Fig. 6.9 illustrates the impacts of discount rate and annual decline rate of battery price on the life-cycle economic benefits of four design scenarios under the Guangdong electricity market. It can be observed that both Scenario 2 (Fig. 6.9B) and Scenario 3 (Fig. 6.9C) do not exhibit economic benefits across a range of discount rates (2%-10%) and battery price decline rates (5%-15%). For Scenario 1, profits become positive when the discount rate falls below 2.7% (Fig. 6.9A). Similarly, there is also a dividing curve for Scenario 4. In the lower right of the dividing curve in Fig. 6.9D (the discount rate of 9.5% and the annual decline rate of battery price of 5.6%), there are positive economic benefits. Generally, staged investments accompanied by revenues from grid services are more likely to yield positive returns with lower discount rates and higher annual decline rates of battery price.

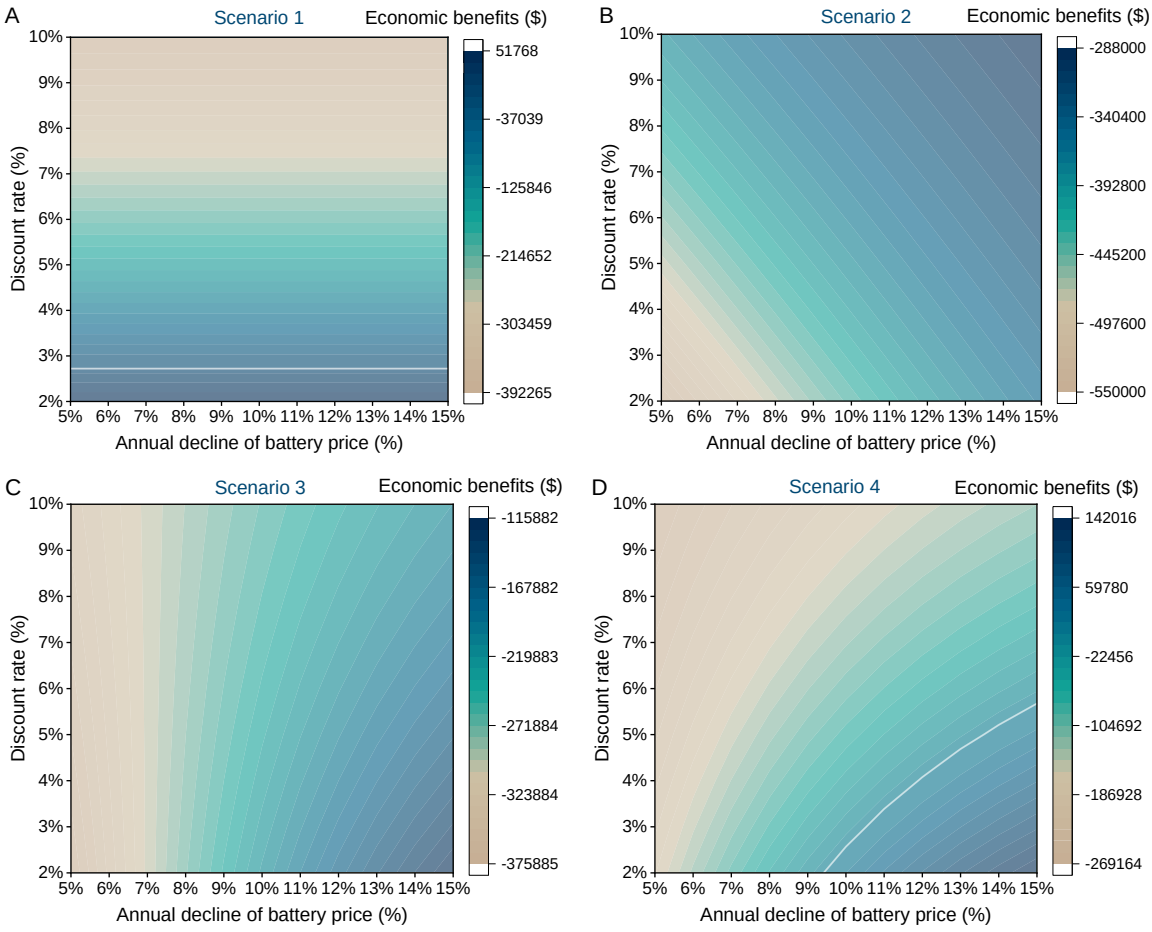


Fig. 6.9 Life-cycle economic benefits of EES versus discount rates and battery price decline rates under the Guangdong electricity market

Fig. 6.10 illustrates the impacts of discount rate and annual decline rate of battery price on the life-cycle economic benefits of four design scenarios under the CAISO electricity market. It can be observed that Scenarios 1, 3 and 4 each show dividing lines/curves where positive and negative economic benefits are represented on either side of the line/curve. For Scenario 1, profits are positive when the discount rate is less than 9% (Fig. 6.10A). For Scenario 3 (Fig. 6.10C) and Scenario 4 (Fig. 6.10D), there are positive economic benefits in the lower right of the dividing curves. Additionally, similar to the results under the Guangdong electricity market, Scenario 2 (Fig. 6.10B) does not exhibit economic benefits across a range of discount rates (2%-10%) and battery price decline rates (5%-15%).

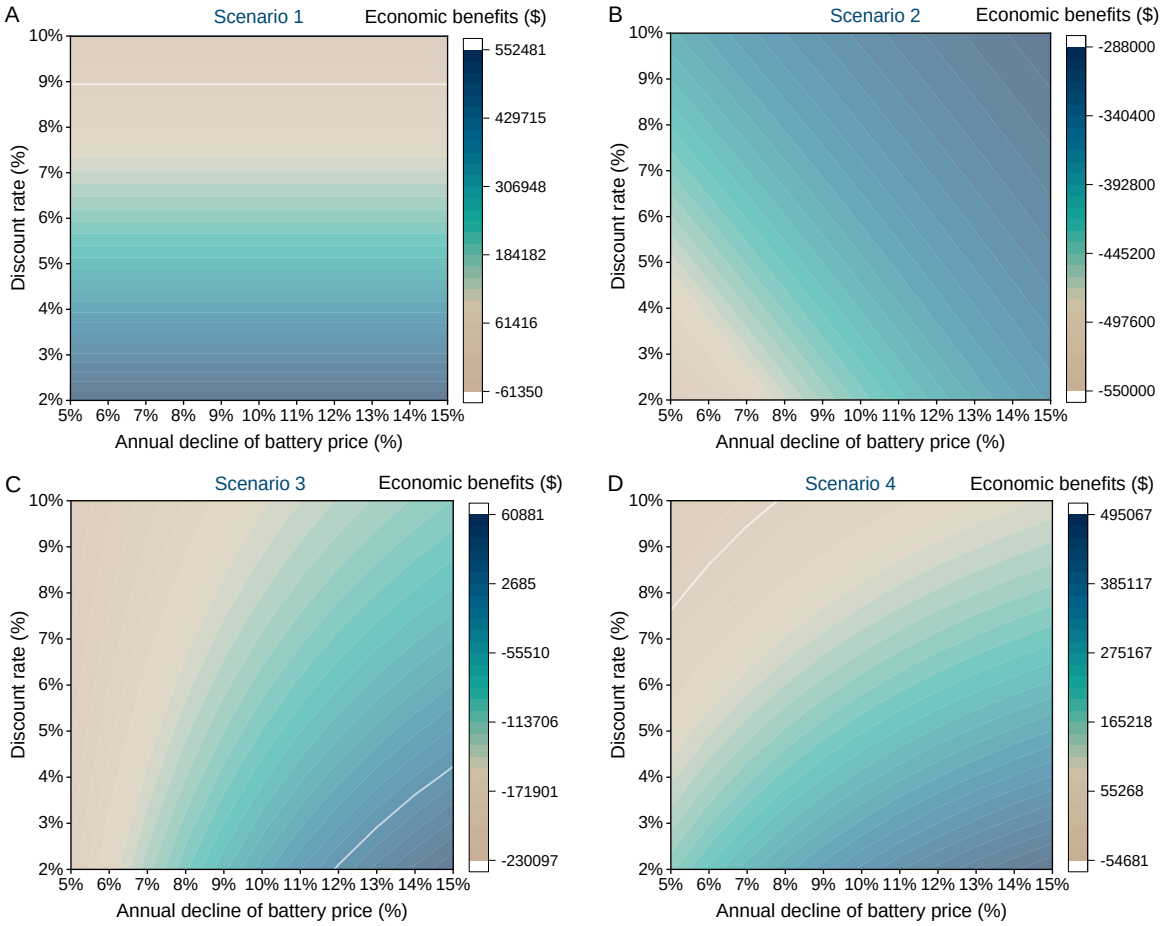


Fig. 6.10 Life-cycle economic benefits of EES versus discount rates and battery price decline rates under the CAISO electricity market

6.5.2 Impacts of discount rate on life-cycle economic benefits of TES

Fig. 6.11 shows the impact of the discount rate on the life-cycle economic benefits of TES under the Guangdong electricity market. It can be observed that Scenarios 1, 3 and 4 all exhibit

positive economic benefits, which decrease as the discount rate increases from 2% to 10%. This trend can be attributed to the fact that a higher discount rate devalues future money, resulting in a lower present value of future earnings. In contrast, Scenario 2 shows opposite results, exhibiting negative economic benefits and an uptrend as the discount rate increases. In Scenario 2, the energy storage system is designed for emergencies with a staged investment aligned with progressive IT loading, without providing additional dispatchable capacity for auxiliary services to the grid. As a result, future investments, when discounted back to their present value, are less when the discount rate is higher.

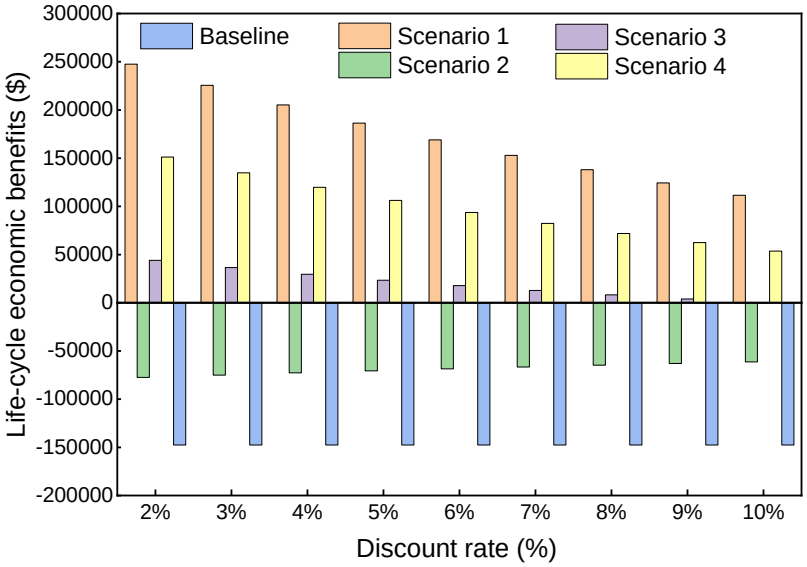


Fig. 6.11 Life-cycle economic benefits of TES under the Guangdong electricity market at discount ranging from 2% to 10%

Fig. 6.12 shows the impact of the discount rate on the life-cycle economic benefits of TES under the CAISO electricity market. It can be observed that only Scenario 1 shows positive economic benefits, and this is limited to instances where the discount rate is below 9%. In this scenario, the economic benefits decrease as the discount rate increases from 2% to 10%. This can be attributed to the fact that a higher discount rate devalues future money and results in a lower present value of future earnings.

The other scenarios all show negative economic benefits across a range of discount rates (2%-10%). For Scenario 2, the trend of economic benefits as the discount rate increases is similar to that under the Guangdong electricity market. However, in Scenarios 3 and 4, there is a noticeable decline in economic benefits as the discount rate rises from 2% to 10%. Both scenarios involve staged investments and revenues generated from providing services to the

grid. As the discount rate increases, the present value of future earnings decreases more significantly, resulting in a downward trend.

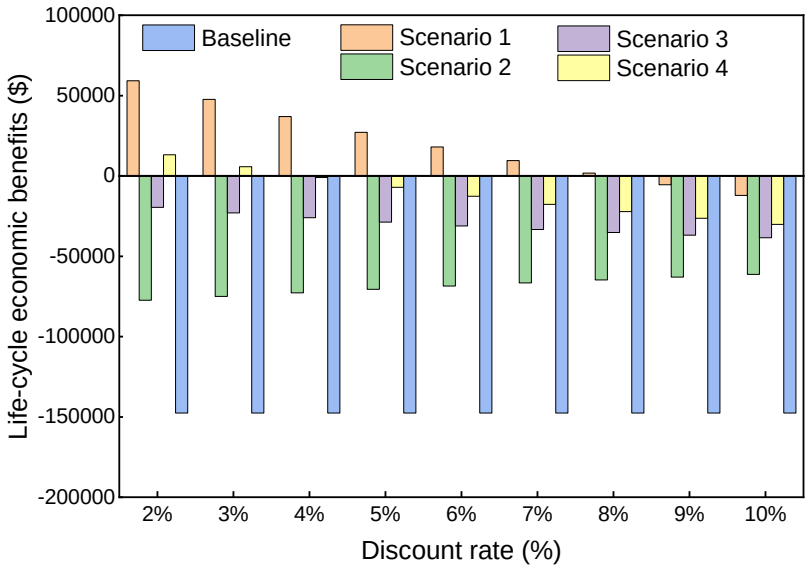


Fig. 6.12 Life-cycle economic benefits of TES under the CAISO electricity market at discount ranging from 2% to 10%

6.6 Summary

This chapter focuses on the optimal dispatch and design of energy storage systems for emergencies in data centers, particularly in the context of progressive IT loading. The main contributions of this chapter include: i) identifying the surplus capacities of energy storage systems in data centers concerning progressive loading. ii) formulating the optimization problem to minimize the life-cycle electricity costs, by allocating surplus capacities to provide flexible services to the grid. iii) proposing four design scenarios for energy storage systems in data centers considering progressive loading over the data center's lifecycle. iv) quantifying and analyzing the economic benefits of each scenario based on the optimized dispatch strategies. v) examining the impacts of discount rates and battery prices on the life-cycle economic benefits of EES and TES. The key findings of this study are as follows.

Under both the Guangdong electricity market and the CAISO electricity market, the charging and discharging occur at low and high tariffs, respectively, thereby facilitating energy arbitrage. Furthermore, under the CAISO electricity market, a significant portion of the surplus battery capacity is allocated to provide frequency regulation services. In contrast, under the Guangdong electricity market, the majority of the surplus battery capacity is dedicated to energy arbitrage. This can be attributed to higher rewards for providing regulation services in the CAISO market than in the Guangdong market.

Under both electricity markets, Scenario 1 emerges as the optimal option with the highest life-cycle economic benefits. Specifically, under the Guangdong electricity market, the life-cycle economic benefits of EES and TES are $-86,418 and $205,213, respectively. Under the CAISO electricity market, the life-cycle economic benefits of EES and TES are $361,453 and $36,985, respectively. Notably, the results reveal that EES yields greater economic benefits under the CAISO market, while TES yields greater economic benefits under the Guangdong market.

However, the results might be significantly different when the discount rate and annual decline rate of battery price vary. Typically, staged investments accompanied by revenues from grid services are more likely to yield positive returns with lower discount rates and higher annual decline rates of battery price. These results can be elucidated by examining the present value of future cash flows; as the discount rate escalates, the present value of both future earnings and investments correspondingly decreases.

The results provide valuable insights into the optimal dispatch and design of energy storage systems in data centers and guide the development of next-generation data centers that can engage in dynamic interactions with energy systems. Data centers have the potential to play a significant role in the future energy landscape by actively participating in grid interactions and supporting the transition to a more sustainable and resilient power system.

CHAPTER 7 "EASTERN DATA, WESTERN COMPUTING": THE ENERGY, ECONOMIC AND CARBON BENEFITS FOR CHINA'S DATA CENTERS

This chapter presents a comprehensive and quantitative assessment of the energy-saving potentials, economic benefits and carbon emission reductions of the national initiative 'Eastern Data, Western Computing'. Chapter 7.1 presents the motivation for assessing the impact of the national initiative 'Eastern Data, Western Computing'. Chapter 7.2 presents the modeling of cooling systems, data transmission and power transmission. Chapter 7.3 introduces the mathematical formulas associated with the energy saving, economic benefits and carbon emission reductions of the national initiative. Chapter 7.4 presents the energy performance of cooling systems and data-transmission energy. Chapter 7.5 analyzes the impacts on energy, economy and carbon emissions of the initiative, and discusses the energy/emissions trade-offs associated with each route. Chapter 7.6 and Chapter 7.7 propose future perspectives and challenges on carbon emission reduction, as well as potential policy suggestions and actionable insights.

7.1 Motivations for "Eastern Data, Western Computing"

China's digital economy is expanding rapidly, leading to a significant surge in demand for computing capacity and the proliferation of data centers. In 2021, data centers in China consumed over 200 billion kWh, accounting for approximately 2.7% of total electricity consumption [216]. The ever-increasing energy consumption of data centers has been a great challenge to China's commitment to achieving carbon neutrality by 2060.

The majority of China's data centers are located in economically developed yet electricity-deficient Eastern regions, such as the Beijing-Tianjin-Hebei Region and the Yangtze River Delta Region, where the demand for cloud services is higher. To mitigate power shortages in these regions, the Western regions with abundant electricity will transmit power to the Eastern regions. This is known as 'West-to-East Power Transmission (WEPT)', proposed by the Chinese government in 2000 [217]. However, power transmission lines are inevitably accompanied by power loss, averaging 5-6% [218].

In 2021, the Chinese government introduced an ambitious national initiative called 'Eastern Data, Western Computing' (EDWC) [219], namely 'moving bits rather than moving watts'. The

national initiative aims to migrate computing workloads from data centers in the East (electricity-deficient) to data centers in the West (renewable-rich) within China. Then, the abundant electricity in the Western regions can be utilized locally by the data centers, thereby eliminating the power losses incurred by long-distance transmission. Furthermore, the Western regions of China have favourable climates that could significantly reduce cooling energy use by more effective utilization of 'free cooling' [220]. This could substantially reduce Power Usage Effectiveness (PUE) of data centers. Moreover, from the perspective of cost, the construction of fibre-optic networks is much cheaper, at $70-150K per mile [221, 222], compared to the construction of power transmission infrastructure, typically averaging $1.5-$2.0 million per mile [223, 224].

Fiber-optic communication technology plays a key role in data transmission [152], which enables effective 'communication' among geographically distributed data centers. Over the past four decades, significant progress has been made in this field [153, 154], including the development of wavelength-division multiplexing [155] and space-division multiplexing [156]. Despite these advancements, network delay remains a challenge in long-distance data transmission [157, 158]. In EDWC initiative, workloads that can tolerate delay, such as storage and backup, will be migrated. Whereas workloads requiring timely responses, such as web search and videoconferencing, will continue to be processed at local data centers in the East. It is noteworthy that data centers typically have a significant proportion of delay-tolerant workloads, more than 50% of total workloads [159]. This provides substantial temporal and spatial flexibility for geographically distributed data centers [160]. Existing studies have shown an increasing interest in the flexible scheduling of workloads in geo-distributed data centers, with a focus on optimizing energy use and cost [161, 162] and maximizing renewable utilization [163]. A study shows that up to 40% of the operational cost can be reduced through load distribution and scheduling in geographically distributed data centers [164]. By migrating data center workloads from the fossil-fuel-heavy regions to the renewable-heavy regions, up to 239 $KtCO_2e$ can be reduced per year [1].

Despite these potential benefits, the implementation of the EDWC project involves three key trade-offs. The first is the energy trade-off. While potential energy benefits can be achieved by eliminating power transmission loss and reducing cooling energy, these advantages need to be weighed against the increased energy consumption of massive data transmission [225]. The second is the economic trade-off. The construction of new data centers in the Western regions and the dedicated data transmission lines require significant capital investments. The costs of

recent data center constructions typically range from $7 to $9 per watt [226]. Therefore, a balance needs to be struck between the capital cost and the potential energy cost saving. The third is the carbon emission trade-off. The current carbon emission factors associated with power consumption and generation vary across different regions [227], which significantly impacts the overall carbon emission reduction. This trade-off involves weighing the carbon emission reduction associated with energy saving against the carbon emission associated with power consumption.

Therefore, a comprehensive assessment is conducted concerning the energy, economic and carbon impacts of the national initiative. The analysis focuses on the three major routes, considering two scenarios for data transmission: on existing backbone networks and newly built dedicated fiber-optic lines. To conduct the quantitative assessment, a typical cooling system model is developed using fundamental mathematical formulas and test data of cooling equipment. Meanwhile, the power required for data transmission and the power transmission losses for each route are determined using relevant models and data. Based on the energy performance assessment, the economic benefits and carbon emission reductions of each route are evaluated. The quantitative results offer valuable insights for policy-makers, industry stakeholders, and researchers to make further policy targets together with technological innovations to facilitate the decarbonization of the data center industry.

7.2 Development of data transmission model

7.2.1 Data transmission on the existing backbone networks

The historical trends of the energy intensity (EI) of data transmission on the existing backbone networks from 2006 to 2021 are shown in Fig. 7.1(D) and Table 7.1. In 2021, data-transmission EI on the backbone network is estimated as 0.001 kWh/GB at an average distance of 700 km [228]. The power consumption of the backbone network is the sum of the energy consumption of network equipment (i.e., routers, amplifiers, etc.) [229]. Notably, if the two data centers are relatively close to each other, there will be a shorter ‘communication’ distance and the migrated workloads will be through fewer routers (fewer hops) [230]. Therefore, it is assumed that the data-transmission EIs are proportional to the distance between the two regions involved.

Table 7.1 Historical development of energy intensity of data transmission from 2004 to 2021

Studies	Year (data)	EI (kWh/GB)	Source
[1] Taylor et al. (2004) [231]	2000	0.107	[232]
[2] Schien et al. (2012) [233]	2008	0.057	[234]

[3] Schien et al. (2013) [235]	2010	0.038	Original text
[4] Malmodin et al. (2012) [236]	2010	0.08	SI in [232]
[5] Schien et al. (2014) [237]	2011	0.02	Original text
[6] Krug et al. (2014) [238]	2012	0.04	[232]
[7] Schien et al. (2015) [239]	2014	0.052	Original text
[8] Malmodin et al. (2016) [240]	2015	0.008	Estimate based on data in [232]&[240]

7.2.2 Data transmission on newly built dedicated fiber-optic lines

In the scenario 'dedicated data-transmission lines', we use the power consumption model developed by Heddeghem et al. [241, 242], given by Eq. (7.1)-(7.2).

$$P_{BACKBONE} = P_{IP} + P_{WDM} \tag{7.1}$$

$$P_{WDM} = P_{OXC} + P_{TXP} + P_{OLA} \tag{7.2}$$

The power consumption of each network equipment is further given as Eq. (7.3)-(7.6) [241, 242]

$$P_{IP} = \eta_{eo} \cdot \frac{\eta_{pr}}{2} \cdot \eta_{op} \cdot T \cdot (H+1) \cdot (\frac{P_{ip}}{C_{ip}} \cdot 2) \tag{7.3}$$

$$P_{OXC} = \eta_{eo} \cdot \eta_{pr} \cdot \eta_{op} \cdot T \cdot H \cdot (\frac{P_{oxc}}{C_{oxc}} \cdot 2) \tag{7.4}$$

$$P_{TXP} = \eta_{eo} \cdot \eta_{pr} \cdot \eta_{op} \cdot T \cdot H \cdot (\frac{P_{txp}}{C_{txp}} \cdot 2) \tag{7.5}$$

$$P_{OLA} = \frac{\eta_{eo}}{2} \cdot \frac{\eta_{pr}}{2} \cdot \eta_{op} \cdot T \cdot H \cdot (\frac{P_{ola}}{C_{ola}} \cdot \frac{\text{link length}}{80\ \text{km}}) \tag{7.6}$$

where P_x is the power consumption of network equipment *x*, Internet protocol (IP), Wavelength division multiplexing (WDM), Optical cross-connects (OXC), Transponder (TXP) and Optical cross-connects (OLA) (W); η_{eo} is the external overhead factor; η_{pr} is the protection factor; η_{op} is the overprovisioning factor; T is the total traffic in the network (Gbps); H is the average hop count in the respective network layer; $\frac{P_x}{C_x}$ expresses the average power per capacity (in W/Gbps) for a given equipment *x*. Note that telecommunication equipment becomes more power efficient each year largely driven by Moore's law. In this study, we assume there is an 11% per year reduction in energy-per-bit [241, 243]. The analytical parameters (Table 7.2) are all based on the latest-generation equipment in this study.

Table 7.2 The analytical parameters of the data-transmission model [241, 242]

Type	Symbol	Value
IP router efficiency	$\frac{P_{ip}}{C_{ip}}$	3.1 W/Gbps
Optical cross-connects (OXC)	$\frac{P_{oxc}}{C_{oxc}}$	0.14 W/Gbps
Transponder (TXP)	$\frac{P_{txp}}{C_{txp}}$	1.6 W/Gbps
Optical cross-connects (OLA)	$\frac{P_{ola}}{C_{ola}}$	0.13 W/Gbps
Hop count	H	1
Protection factor	η_{pr}	2
External overhead factor	η_{eo}	2
Overprovisioning factor	η_{op}	1

The data-transmission EI on the new dedicated communication lines is calculated to be 0.00003 kWh/GB. In this scenario, the distance has very little impact on power consumption due to there being only one hop. The substantial difference between the EI values of the backbone network and the new dedicated lines can be attributed to two factors. Firstly, the backbone network incorporates multiple hops and routers, requiring additional energy consumption compared to the dedicated data transmission line with only one hop. Secondly, Moore's Law states that the number of transistors in an integrated circuit doubles approximately every two years [244]. The rapid development in communication equipment has resulted in the latest-generation equipment being highly power-efficient. However, it is unrealistic to replace all network equipment with the latest and most efficient generations continuously [241]. There is a time lag between the introduction of new technology and its widespread adoption in the backbone network.

7.3 Analysis of energy saving, economic benefits and carbon emission reductions

7.3.1 Energy-Saving Potential

The potential energy benefits can be achieved by *i*. the reduction in cooling energy due to the different climate conditions in two regions involved in each route and *ii*. the eliminated power transmission loss involved in long-distance power transmission.

Difference in cooling energy consumption of data centers in the East and West

In this study, it is assumed that the electricity used by data centers in the East is transmitted from Western regions. The power transmission losses of these power transmission lines are summarized in Table 7.3. A typical 10MW data center, which has a peak power of 21 MW and

consumes approximately 114 GWh of electricity annually, is used as the fundamental unit for analysis [1]. The detailed specifications and characteristics of the typical data center are given in Table 7.4. The energy consumption breakdown in a typical data center is summarized in Table 7.5. Additionally, it is assumed that the energy consumption of IT equipment is proportional to computing workloads [245]. The cooling energy required for processing the same computing workloads in the West is calculated by Eq. (7.7). The energy consumption for the same computing workloads in the East and West is given by Eq. (7.8)-(7.9).

$$Ele_{coolwest} = \frac{Ele_{cooleast} \times COP_{West}}{COP_{East}} \tag{7.7}$$

$$TotEle_{Eest} = Ele_{IT} + Ele_{cooleast} + Ele_{others} \tag{7.8}$$

$$TotEle_{West} = Ele_{IT} + Ele_{coolwest} + Ele_{others} \tag{7.9}$$

Table 7.3 Transmission loss for cross-region power transmission in typical cases

Engineering	Voltage (kV)	Type	Transmission loss
Inner Mongolia-Beijing	±800	UHVDC	7.00% [246]
Sichuan-Shanghai	±800	UHVDC	6.54% [247]
Guizhou-Guangdong	±500	HVDC	7.05% [248]

Table 7.4 Characteristics of the 10 MW* typical data center used for analysis [1]

Metric	Value
Idling power per server	120 W
Maximum power per server	250 W
Annual average server utilization rate	40%
Peak total IT load	10 MW
Peak total load	20.7 MW
Estimated annual total energy consumption	114,234 MWh
Delay tolerant workload	10-50%

Table 7.5 Energy consumption breakdown in a typical data center [3, 249]

Component	Proportion
Servers	44%
Storage	15%
Network	3%
Cooling	30%

Others	8%

Consumption of Data Transmission vs Power Transmission Loss

The amount of data in China's data centers is estimated by [250, 251], shown in Table 7.6. It is assumed that data volume is proportional to the computing workloads from the perspective of overall trends and general patterns. The energy consumption of data transmission is calculated by Eq. (7.10).

$$Ele_{Datatrans} = EI_{Datatrans} \times V_{Datatrans} \quad (7.10)$$

Three existing typical lines in the WEPT project are selected as baseline scenarios. The energy loss of power transmission is calculated by Eq. (7.11).

$$Ele_{Ploss} = \lambda_{Ptrans} \times Ele_{Ptrans} \quad (7.11)$$

Table 7.6 The amount of data in data centers [250, 251]

Data type	Volume (ZB)
The data stored in data centers globally	1.3 [250]
Percentage of data volume in China	10% [251]
The data stored in data centers in China	0.13

7.3.2 Carbon emission reductions

The carbon emission reductions are achieved by *i*. the difference in carbon emissions of power consumption between Eastern and Western regions. *ii*. the difference in carbon emissions of data transmission and power transmission losses.

In China, cross-regional electricity trade is an important measure to alleviate the unbalanced spatial distribution of power consumption and generation. It entails the transfer of carbon emissions, thereby resulting in unequal CO_2 emission factors associated with power generation and power consumption in specific provinces and cities. The carbon emission factors associated with power generation and power consumption in different provinces and cities are summarized in Table 7.7. In addition, carbon emissions factors associated with the raw material and production of data transmission and power transmission are summarized in Table 7.8

In the power-transmission scenario, total carbon emissions include the carbon emissions associated with power consumption in Eastern regions. In the data-transmission scenario, total carbon emissions include the carbon emissions associated with power consumption in Western regions, the increased carbon emissions of data transmission, and the reduced carbon emissions

associated with eliminated power transmission loss. The equations for calculating total carbon emission reductions are presented in Eq. (7.12)-(7.16).

$$\Delta CE_{Total} = \Delta CE_{Elecons} + \Delta CE_{Other} \quad (7.12)$$

$$\Delta CE_{Elecons} = f_{Westcons} \times TotEle_{West} - f_{Eastcons} \times TotEle_{East} \quad (7.13)$$

$$\Delta CE_{Other} = CE_{Datatrans} - CE_{Ploss} \quad (7.14)$$

$$CE_{Datatrans} = \beta_{Datatrans} \times V_{Datatrans} + f_{Avecons} \times Ele_{Datatrans} \quad (7.15)$$

$$CE_{Ploss} = \beta_{Ptrans} \times Ele_{Ploss} + f_{Westgene} \times Ele_{Ploss} \quad (7.16)$$

Table 7.7 CO_2 emissions factors of major cities involved in this study [227]

City	Emissions factor of power generation* (g/kWh)	Emissions factor of power consumption* (g/kWh)
Beijing	344.51	661.37
Inner Mongolia	918.24	916.59
Shanghai	750.8	532.36
Sichuan	95.47	112.04
Guangzhou	558.35	442.25
Guizhou	451.87	451.55

* The emissions factor of power generation is the CO_2 emissions of local electricity generation. The emissions factor of power consumption is the CO_2 emissions of local electricity consumption, considering the CO_2 emissions embodied in cross-province electricity trade (the mix of local electricity generation and electricity transmitted from other regions)

Table 7.8 CO_2 Emissions factors associated with raw material and production of data transmission and power transmission

Category	CO_2 Emissions of production
Data transmission	0.03 (gCO_2/GB) [252]
Power transmission	1.12 (gCO_2/kWh) [253]

7.3.3 Economic analysis

In the economic analysis, we consider several key factors. *i*, the capital investment required for building duplicated data centers in the Western regions and for building dedicated fiber-optic lines. *ii*, the difference in electricity prices between Eastern regions and Western regions. *iii*, the difference in the energy costs of data transmission and the eliminated power transmission loss.

In the data-transmission scenario, the amortized cost of traditional data centers is $1.56/W per year (at an interest rate of 8% and depreciating data centers over 12 years) [226]. Therefore, a 10 MW traditional data center has an amortized facility cost of $15.6 million. The amortized capital cost for newly built fiber-optic lines is estimated according to the literature [254, 255], shown in Table 7.9. The electricity cost in different regions involved in this study is summarized in Table 7.10. The economic analysis is calculated by Eq. (7.17)-(7.21). Nomenclature for Eq. (7.1)-(7.21) is shown in Table 7.11.

$$\Delta Cost_{Total} = \Delta Cost_{DC} + \Delta Cost_{Ele} + \Delta Cost_{Other} + (\Delta Cost_{DL}) \quad (7.17)$$

$$\Delta Cost_{Ele} = Eprice_{West} \times TotEle_{West} - Eprice_{East} \times TotEle_{East} \quad (7.18)$$

$$\Delta Cost_{Other} = Cost_{Datatrans} - Cost_{Ploss} \quad (7.19)$$

$$Cost_{Datatrans} = AveEprice \times Ele_{Datatrans} \quad (7.20)$$

$$Cost_{Ploss} = AveEprice \times Ele_{Ploss} \quad (7.21)$$

Table 7.9 The amortized cost for data centers and dedicated fiber-optic lines [226]

Category	Amortized cost
Data centers ($/W *)	1.56
Route: Inner Mongolia-Beijing (million $)	7 **
Route: Sichuan-Shanghai (million $)	28 **
Route: Guizhou-Guangdong (million $)	14 [254, 255]

* The unit $/W is dollar per watt of IT power

** Estimated value mainly considering the distance between two regions based on the special value in the case of Guizhou-Guangzhou.

Table 7.10 Electricity prices in major cities involved in this study

City	Electricity price (yuan/kWh)	Electricity price ($/kWh)
Beijing	0.62 [256]	0.09*
Ulanqab, Inner Mongolia	0.26 [257]	0.04
Shanghai	0.69 [256]	0.1
Ya'an, Sichuan	0.34 [258]	0.05
Guangzhou	0.64 [256]	0.09
Guiyang, Guizhou	0.35 [257]	0.05
Average electricity cost in China	0.61 [259]	0.08

*The exchange rate between RMB and USD: 1RMB = 0.14 USD (2023)

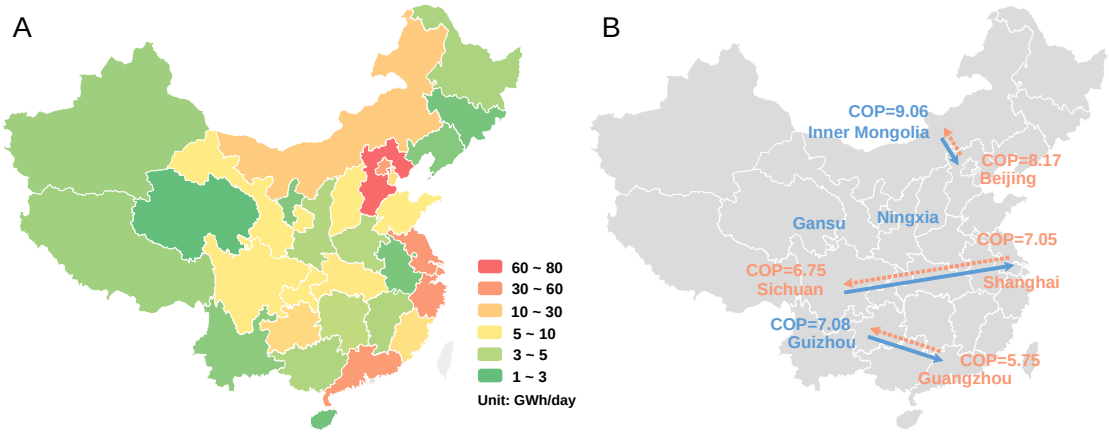
Table 7.11 Nomenclature for Equations

Symbol	Unit	Description
$Ele_{coolwest}$	kWh	Cooling energy consumption in the Western region
$Ele_{cooleast}$	kWh	Cooling energy consumption in the Eastern region
COP_{West}	-	Cooling efficiency in the Western region
COP_{East}	-	Cooling efficiency in the Eastern region
$TotEle_{East}$	kWh	Total electricity consumption in the Eastern region
$TotEle_{West}$	kWh	Total electricity consumption in the Western region
Ele_{IT}	kWh	Energy consumption of IT equipment (servers, storage…)
Ele_{others}	kWh	Energy consumption of other equipment (lighting…)
$Ele_{Datatrans}$	kWh	Energy consumption of data transmission
$EI_{Datatrans}$	kWh/GB	Energy intensity of data transmission
$V_{Datatrans}$	GB	Total volume of data transmission
Ele_{Ploss}	kWh	Energy loss of power transmission
Ele_{Ptrans}	kWh	Total power transmission
λ_{Ptrans}	%	Power loss rates of existing power transmission lines
$\Delta Cost_{Total}$	$	Total cost saving
$\Delta Cost_{DC}$	$	Amortized capital cost for newly built data centers
$\Delta Cost_{Ele}$	$	The reduced electricity cost due to workload migration
$\Delta Cost_{Other}$	$	The difference in the energy costs of data transmission and eliminated power transmission loss
$\Delta Cost_{DL}$	$	Amortized capital cost for building dedicated fiber-optic lines
$Eprice_{East}$	$/kWh	Electricity prices in the Eastern regions
$Eprice_{West}$	$/kWh	Electricity prices in the Western regions
$Cost_{Datatrans}$	$	Energy cost of data transmission
$Cost_{Ploss}$	$	Energy cost of power transmission loss
$AveEprice$	$/kWh	Average electricity price in China
ΔCE_{Total}	tCO_2e	Total carbon emissions reduction
$\Delta CE_{Elecons}$	tCO_2e	The difference in carbon emissions (electricity consumption)
ΔCE_{Other}	tCO_2e	The difference in carbon emissions (data transmission *vs* power transmission)
$CE_{Datatrans}$	tCO_2e	Carbon emissions of data transmission
CE_{Ploss}	tCO_2e	Carbon emissions of power transmission loss
$\beta_{Datatrans}$	gCO_2e/GB	CO_2 emission factors of raw material and production for data transmission

β_{Ptrans}	gCO2e/kWh	CO_2 emission factors of raw material and production for power transmission
$f_{Eastcons}$	gCO2e/kWh	CO_2 emission factors of power consumption in the Eastern regions
$f_{Westcons}$	gCO2e/kWh	CO_2 emission factors of power consumption in the Western regions
$f_{Westgene}$	gCO2e/kWh	CO_2 emission factors of power generation in the Western regions
$f_{Avecons}$	gCO2e/kWh	Average CO_2 emission factors of power consumption in the regions involved in each route

7.4 Energy performance of cooling systems and data-transmission energy

We collect the spatial distribution of the energy consumption of data centers across China in 2020 (Fig. 7.1(A)) [260]. Beijing-Tianjin-Hebei region, Yangtze River Delta Region, and Pearl River Delta Region show the highest energy consumption. These regions are also the destinations of power delivery in the 'West-East Power Transmission (WEPT)' project. Three existing major long-distance power transmission lines, from Inner Mongolia to Beijing, from Sichuan Province to Shanghai, and from Guizhou Province to Guangzhou Province are shown in Fig. 7.1(B) (in bold blue lines). In this study, the three existing power transmission lines are considered the baseline scenarios for power transmission scenarios. "Eastern Data, Western Computing (EDWC)" aims to migrate computing workloads from data centers in the East to data centers in the West for storage and processing, as shown in Fig. 7.1(B) (bold orange lines). This migration can eliminate the energy loss associated with long-distance power transmission and utilize the cold weather in the Western regions for more energy-efficient cooling, which is considered in the data-transmission scenario.



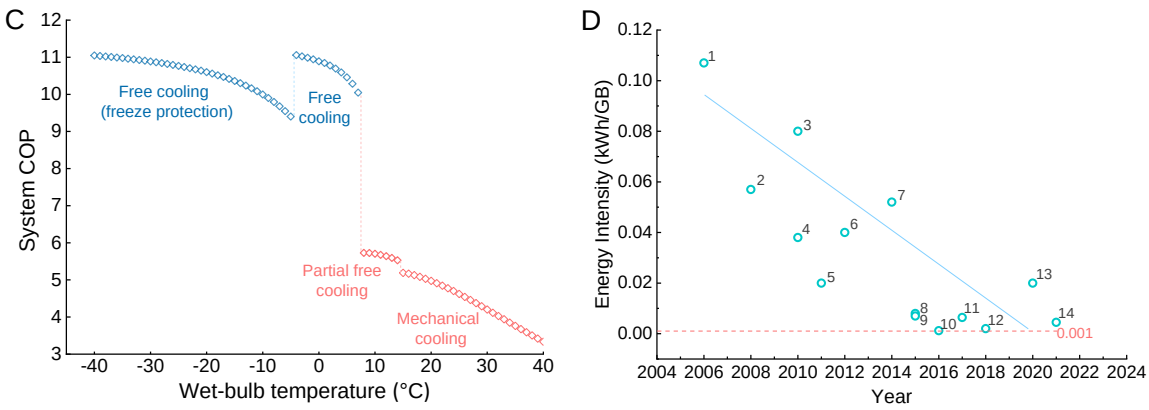


Fig. 7.1 Data center energy consumption, major migration routes, performance of cooling systems, and data-transmission energy intensity

(A) Spatial distribution of the energy consumption of data centers across China [260].

(B) Cooling system COP in different locations.

(C) Cooling system COP as a function of wet-bulb temperature.

(D) Historical development of energy intensity of data transmission between 2004 and 2021. In (D), y-axis is electricity intensity (kWh/GB). x-axis is the year of each estimate. Data points: (1) [231]; (2) [233]; (3) [235]; (4) [236]; (5) [237]; (6) [238]; (7) [239]; (8) [240]. (9) [232]; (10) [261]; (11) (2019) [230] (12) [252]; (13) [234]; (14) [228]

A basic cooling system model is developed based on fundamental principles and mathematical formulas combined with the test data of cooling equipment performance. Using weather data from typical years, we simulated the cooling system under the three operation modes corresponding to outdoor conditions, and compared their coefficients of performance (COPs). The mode that satisfies the cooling load and consumes the lowest cooling energy is selected for each outdoor condition.

The cooling system COP in the wet-bulb temperatures range of −40°C to 40°C is shown in Fig. 7.1(C). In both partial-free-cooling and mechanical-cooling modes (with chillers in operation), the cooling system COP is around 3–5, whereas in free-cooling mode the COP is around 11 (without chillers in operation). According to the annual wet-bulb temperature distribution of different cities and the cooling system COP as a function of wet-bulb temperature, the annual average COP of cooling systems in each city can be obtained. Fig. 7.1(B) shows the annual average COP of cooling systems in six major cities involved in three routes. The annual average COP of cooling systems improves from 8.17 to 9.06 for Beijing – Inner Mongolia route, and from 5.75 to 7.08 for Guangzhou – Guizhou route, respectively. However, it is reduced from 7.05 to 6.75 for Shanghai – Sichuan route.

Two scenarios for data transmission are considered in this study. One is data transmission on existing backbone networks, the other is on dedicated fiber-optic lines newly built. A common-used metric, electricity intensity, representing the energy efficiency of Internet data transmission is adopted in this study. It is defined as the "electrical energy consumed per amount of data transmitted", and is measured as kilowatt hour per gigabyte of data transferred (kWh/GB) [5]. The historical evolution of the energy intensity (EI) of data transmission on the backbone networks from 2004 to 2021 is summarized in Fig. 7.1(D). There is a consistent downward trend in data-transmission EI over this period. In 2021, data-transmission EI on the backbone networks is estimated as 0.001 kWh/GB at an average distance of 700 km [228]. We also calculate the data-transmission EI on newly-built dedicated fiber-optic lines for three routes based on the power model developed by Heddeghem et al [241, 242], given by Eq. (7.1)-(7.6). Furthermore, we consider the improvement in energy efficiency of equipment due to technological updates according to Moore's law. The data-transmission EI on newly built dedicated fiber-optic lines is estimated as 0.00003 kWh/GB in 2021.

7.5 Impacts on energy, economy and carbon emissions

7.5.1 Data transmission on the existing backbone networks

All migration routes show significant energy-saving potential, indicating that 'moving bits' is much more energy efficient than 'moving watts' (Fig. 7.2(A)). Both Beijing-Inner Mongolia and Guangzhou-Guizhou routes show substantial energy savings, amounting to 695 GWh (9.8%) and 942 GWh (12.5%) annually respectively. For these two routes, the energy benefits can be primarily attributed to cooling energy saving and the elimination of power transmission loss. This is mainly due to the fact that both routes transfer a portion of the cooling load of data centers to regions where cooling systems are more energy efficient. It is also observed that the energy consumption associated with 'moving bits' is much lower than the power loss incurred by 'moving watts'.

Unlike the above two routes, the Shanghai-Sichuan route shows a notable increase in cooling energy consumption. This increase is due to the fact that this route transfers a portion of the cooling load of data centers to the region where cooling systems are less energy efficient. However, this route still shows high energy savings, amounting to 320 GWh (4.8%) annually, primarily due to the elimination of long-distance power transmission loss. Similarly, the energy consumption associated with 'moving bits' is considerably lower than the eliminated power loss. Moreover, all these three routes show that more computing workload migration results in more overall energy saving (Fig. 7.2(B)).

However, no economic benefit is observed for all migration routes if considering the high capital costs for constructing duplicated data centers in the Western regions (Fig. 7.2(C)). The energy cost savings associated with reduced cooling energy and the eliminated of power transmission loss are quite low, compared with the the capital cost of constructing new data centers. We also assess the impact of government subsidies on overall economic benefits, and the results are shown in Fig. 7.2(D). Currently, the local governments of Ulanqab in Inner Mongolia Province, Ya'an in Sichuan Province, and Guiyang in Guizhou Province provide substantial subsidies on their electricity bills for data centers to encourage the development of data centers in the Western regions. The actual electricity prices in these regions are almost half of those in the Eastern regions. Nevertheless, the three routes still do not yield economic benefits although a certain amount of capital costs are offset by government subsidies.

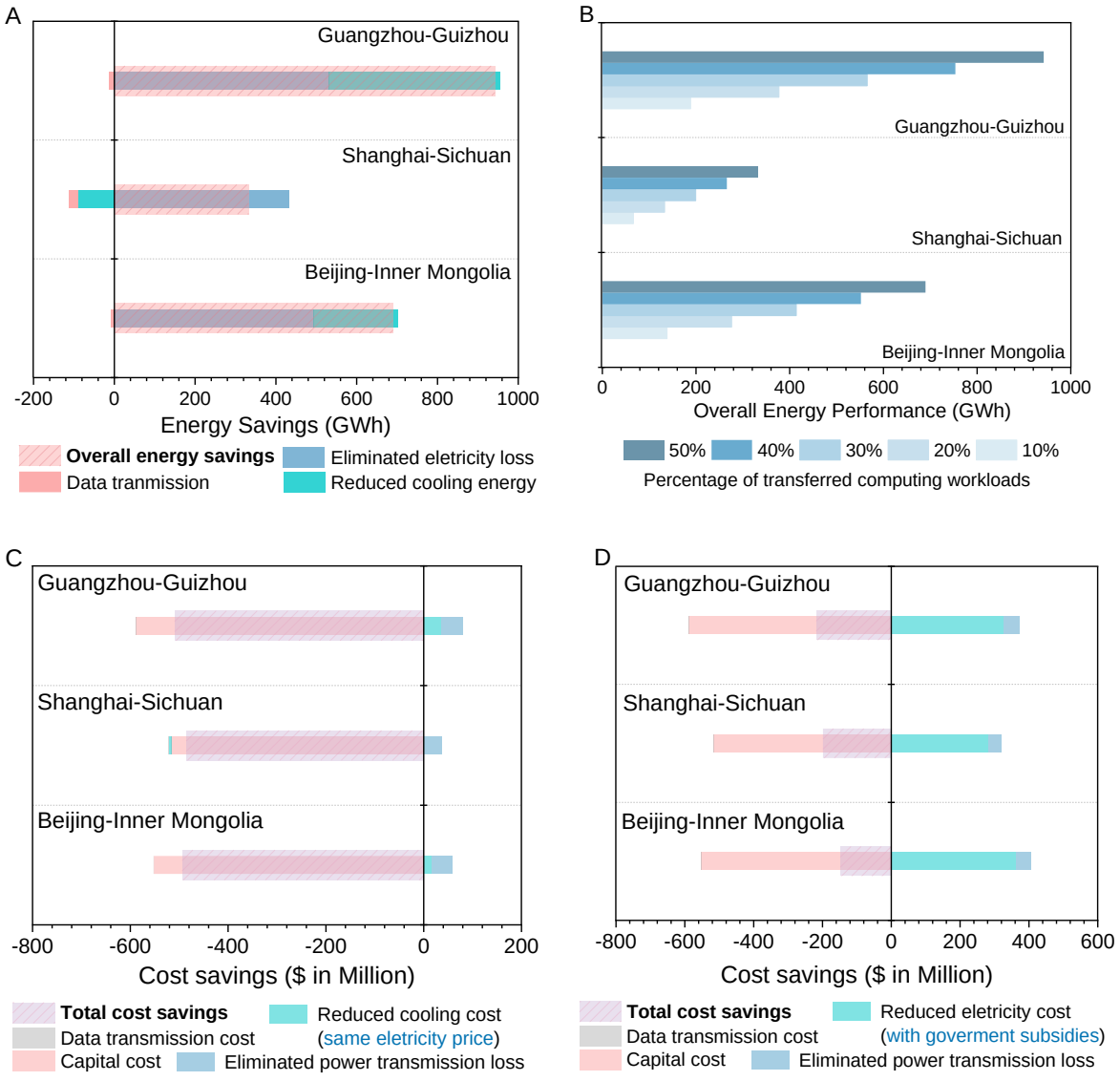


Fig. 7.2 Energy and economic benefits when data transmission on the backbone networks

(A) Overall energy savings and the detailed breakdown with 50% workload migration.
(B) Overall energy performance with 10%-50% workload migration.
(C) Total cost savings and detailed breakdown with 50% workload migration.
(D) Total cost savings considering government subsidies with 50% workload migration.

The benefits of carbon emission reduction in different routes are significantly different (Fig. 7.3(A)). Both Shanghai-Sichuan and Guangzhou-Guizhou routes show a reduction in total carbon emissions, i.e., 2803 KtCO2e (79.6%) and 356 $KtCO_2e$ (10.7%) respectively. However, the Beijing-Inner Mongolia route shows an increase in total carbon emissions, i.e., 1164 $KtCO_2e$ (24.9%). There are different causes for the carbon emissions changes in the three routes as elaborated in the paragraphs below, with some common observations. The eliminated transmission power loss contributes to total carbon reduction for all three routes but it is not the major reason. In addition, the energy consumption of data transmission is extremely low, thereby having a negligible influence on the overall magnitude of total carbon emissions for all three routes.

For the route unfavourable to carbon reduction (Beijing-Inner Mongolia), the increase in CO_2 emissions is primarily attributed to the difference in the CO_2 emission associated with electricity consumption at the departure city (Beijing) and destination city (Inner Mongolia). As reported by a published comprehensive study of regional CO_2 emissions in China, summarized in Table 7.7, the CO_2e factor in Inner Mongolia (917 g/kWh, where thermal power dominates local power generation nowadays) is much higher than that in Beijing (661 g/kWh). There is an increase in carbon emission by 1612 $KtCO_2e$ as the Beijing-Inner Mongolia route transfers a total electricity consumption of 7064.71 GWh computing workloads from Beijing to Inner Mongolia (the total electricity consumption of these computing workloads in Inner Mongolia is 6856.52 GWh, due to the cooling energy savings). In addition, the increase in carbon emission incurred by 'moving bits' is extremely low, only 6 $KtCO_2e$. In contrast, there is a reduction in carbon emissions of 454 $KtCO_2e$ due to eliminated power transmission loss.

For the route favourable to carbon reduction (Shanghai-Sichuan), the reduction in CO_2 emissions is also primarily attributed to the difference in the CO_2 emission associated with electricity consumption at the departure city (Shanghai) and destination city (Sichuan). The CO_2e factor in Sichuan Province (112 g/kWh, a large proportion of hydropower [262])) is much lower than that in Shanghai (532 g/kWh). There is a reduction in carbon emission by 2769 $KtCO_2e$ as the Shanghai-Sichuan route transfers a total electricity consumption of 6610.15 GWh computing workloads from Shanghai to Sichuan (the total electricity

consumption of these computing workloads in Sichuan is 6698.85 GWh, due to an increase in cooling energy). Furthermore, there is a reduction in carbon emissions of 42 $KtCO_2e$ due to eliminated power transmission loss.

For another route favourable to carbon reduction (Guangzhou-Guizhou), the reason for reduced CO_2 emissions is significantly different from the above two routes. In this route, the CO_2e factors at the departure city (Guangzhou) and destination city (Guizhou) are similar. The reduction in CO_2 emissions in this route is primarily attributed to energy benefits, i.e., eliminated power transmission loss and reduced cooling energy. There is a reduction in CO_2 emissions of 240 $KtCO_2e$ and 122 $KtCO_2e$ for eliminating power transmission loss and reduced cooling energy, respectively. In addition, both Guangzhou-Guizhou and Shanghai-Sichuan routes show that more computing workload migration results in more carbon emission reductions (Fig. 7.3(B)), whereas Beijing-Inner Mongolia shows the opposite results.

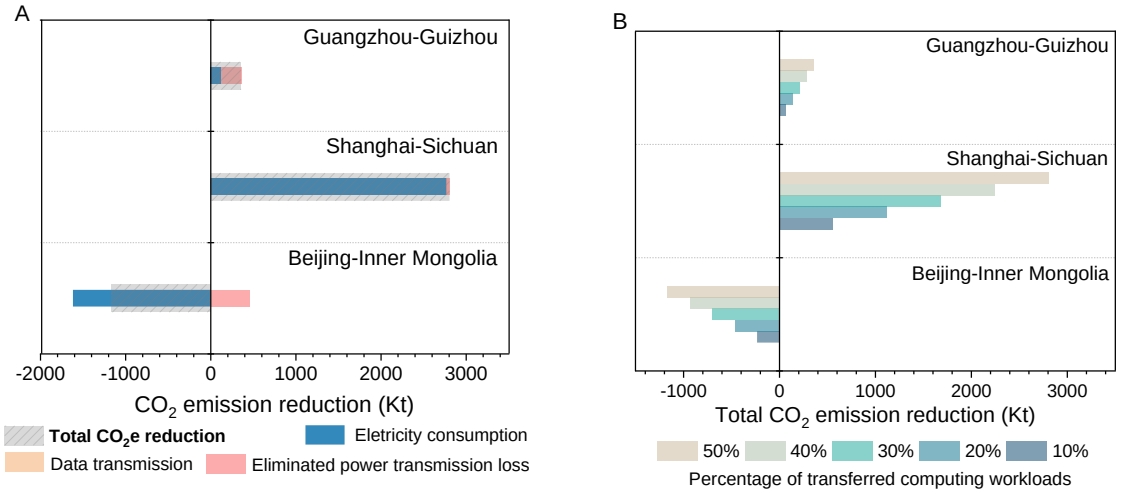


Fig. 7.3 Carbon emission reductions when data transmission on the backbone networks (A) Total carbon emission reductions and detailed breakdown with 50% workload migration. (B) Total carbon emission reductions with 10%-50% workload migration.

7.5.2 Data transmission on newly built dedicated fiber-optic lines

Currently, the bandwidth of existing backbone networks connecting the Eastern and Western regions is insufficient for the implementation of EDWC. Low network latency and higher network capacity are required for the timely transmission of batch computing workloads without data loss. Therefore, establishing new dedicated fiber-optic communication lines with higher bandwidth is necessary for the implementation of EDWC. Fig. 7.4 (A)-(D) shows the overall energy savings, economic benefits, and carbon emission reductions of the national project EDWC in the scenario of 'data transmission on newly built dedicated lines'.

The energy consumption of data transmission on newly built dedicated lines is lower than that on the existing backbone networks, and can even be considered negligible (Fig. 7.4(A)). The amortized capital cost of constructing fiber-optic lines is relatively low, compared to that of newly built data centers. All three representative routes show higher energy-saving potential, compared to data transmission on the existing backbone networks, ranging from 344 GWh to 955 GWh. Similarly, all routes are not economically beneficial if they involve high capital costs for constructing new data centers in two scenarios, without considering government subsidies (Fig. 7.4(B)) and considering government subsidies (Fig. 7.4(C)). Regarding the benefit of carbon emission reduction (Fig. 7.4(D)), both Shanghai-Sichuan and Guangzhou-Guizhou routes show a notable reduction in carbon emissions, amounting to 2810 $KtCO_2e$ and 362 $KtCO_2e$ respectively. While the Beijing-Inner Mongolia route shows an increase in carbon emissions, amounting to 1158 $KtCO_2e$. The results are also similar to the scenario 'data transmission on existing backbone networks'.

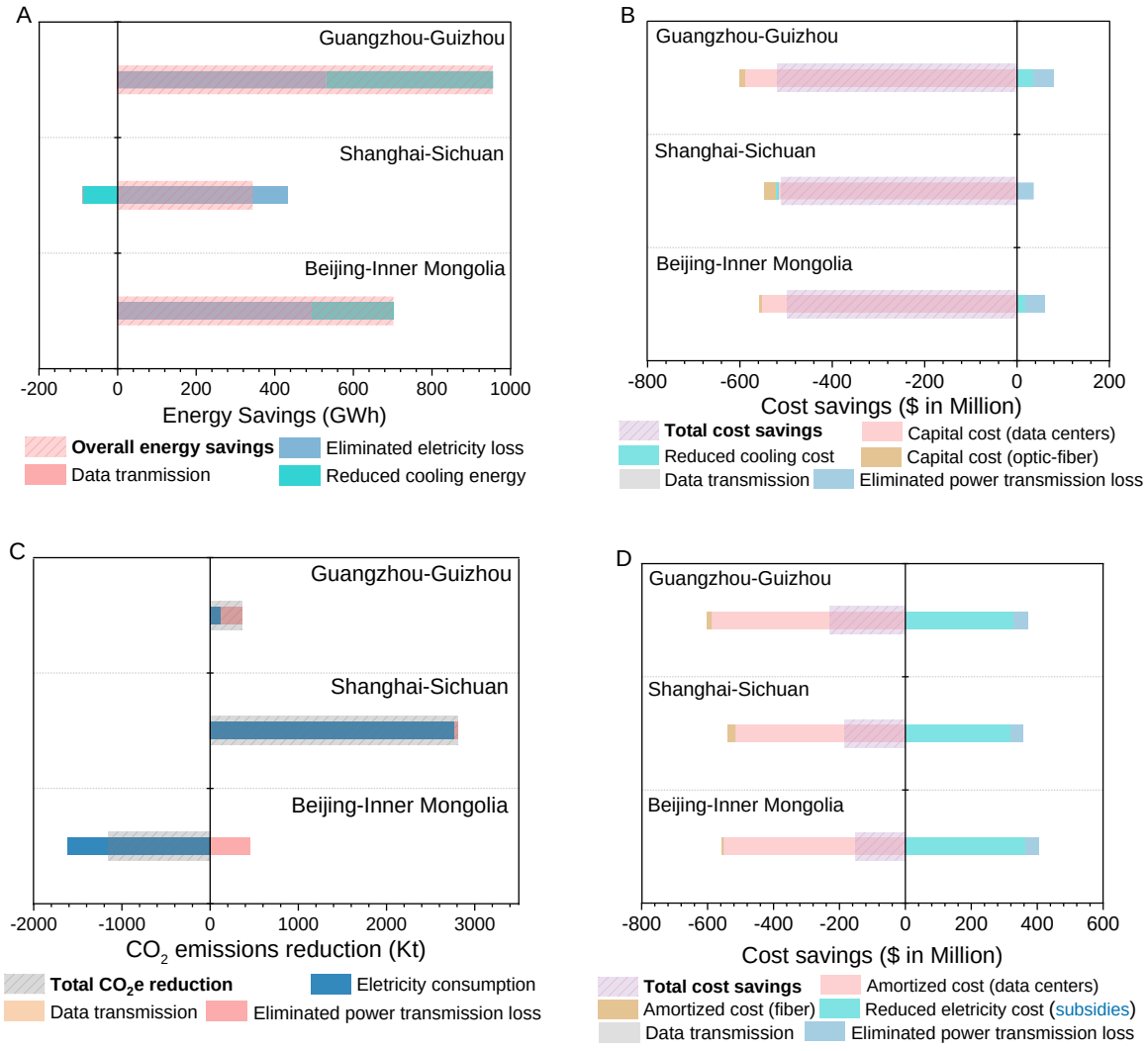


Fig. 7.4 Energy, economic and carbon benefits when data transmission on dedicated lines
(A) Overall energy savings and the detailed breakdown with 50% workload migration.
(B) Total cost savings and detailed breakdown with 50% workload migration.
(C) Total carbon emission reductions and detailed breakdown with 50% workload migration.
(D) Total cost savings considering government subsidies with 50% workload migration.

7.6 Future perspectives on carbon emission reduction and abatement cost

Fig. 7.5 shows the total carbon emission reductions considering future potential changes and uncertainties, including an increase or a decrease in the difference in CO_2e factor of power consumption between the two locations involved in a route.

Currently, the Beijing-Inner Mongolia route exhibits a notable increase in total carbon emissions. However, once the CO_2e factor difference between Inner Mongolia and Beijing is reduced to 85 gCO_2/kWh (the current difference is 255 gCO_2/kWh), the route could yield carbon emission reductions (Fig. 7.5(A)). For the Shanghai-Sichuan route, the current difference in the CO_2e factors is -420 gCO_2/kWh (Fig. 7.5(B)). If we consider the potential decrease in the difference in CO_2e factor between Shanghai (532 g/kWh) and Sichuan (112 g/kWh), the carbon emission reductions for this route will decrease. For the Guangzhou-Guizhou route, the current difference in the CO_2e factors is -9 gCO_2/kWh, and the total carbon emission reduction for this route (Fig. 7.5(C)). However, the CO_2e difference may increase or decrease in the future, primarily due to the penetration of renewable energy in the two regions involved.

In addition, it is worth noting that the Shanghai-Sichuan route shows more carbon emission reductions when more workloads are transferred (Fig. 7.5(B)). However, the other two routes show significantly different trends when CO_2e factor differences are on either side of a critical value. For example, in the Beijing-Inner Mongolia route, when more workloads are transferred, carbon emission reductions increase when the CO_2e difference is greater than 85 gCO_2/kWh, and decrease when the CO_2e difference is less than 85 gCO_2/kWh (Fig. 7.5(A)). Similarly, the critical value for the Guangzhou-Guizhou route is 56 gCO_2/kWh (the current difference is -9 gCO_2/kWh) (Fig. 7.5(C)).

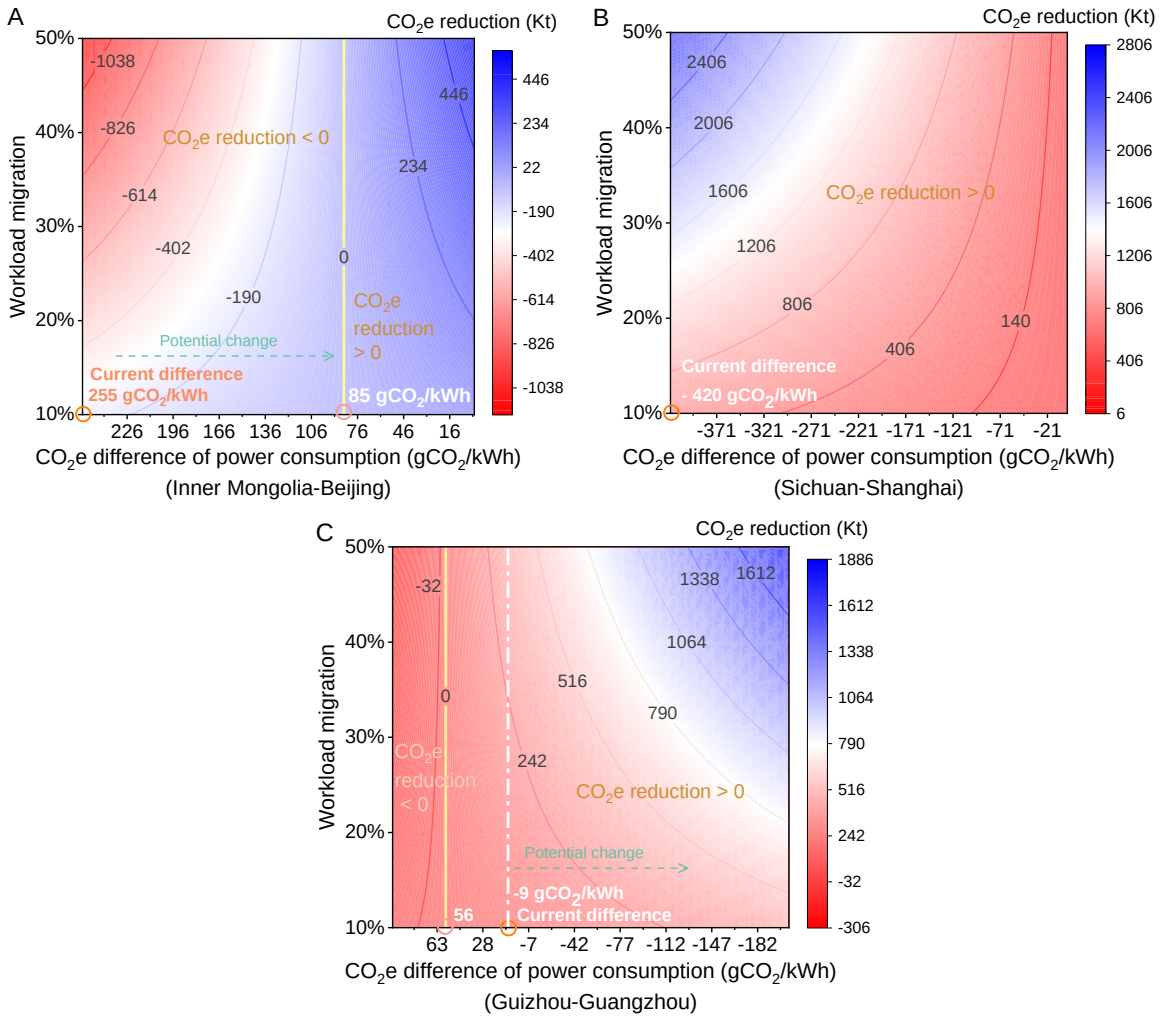


Fig. 7.5 Quantitative carbon emission reductions considering potential increases or decreases in the difference in CO2e factor between the two locations involved in a route.

(A) Beijing-Inner Mongolia route.

(B) Shanghai-Sichuan route.

(C) Guangzhou-Guizhou route.

Fig. 7.6 shows the net carbon abatement costs considering future potential changes and uncertainties in the CO_2e factors at the two locations involved in each route. Currently, the net carbon abatement costs for Shanghai-Sichuan and Guangzhou-Guizhou routes are 174 \$/tCO_2e and 1431 \$/tCO_2e, respectively. In the Shanghai-Sichuan route, the net carbon abatement cost will increase if the CO_2e factor difference decreases in the future. For the Guangzhou-Guizhou route, the net carbon abatement cost will decrease if the CO_2e factor difference increases.

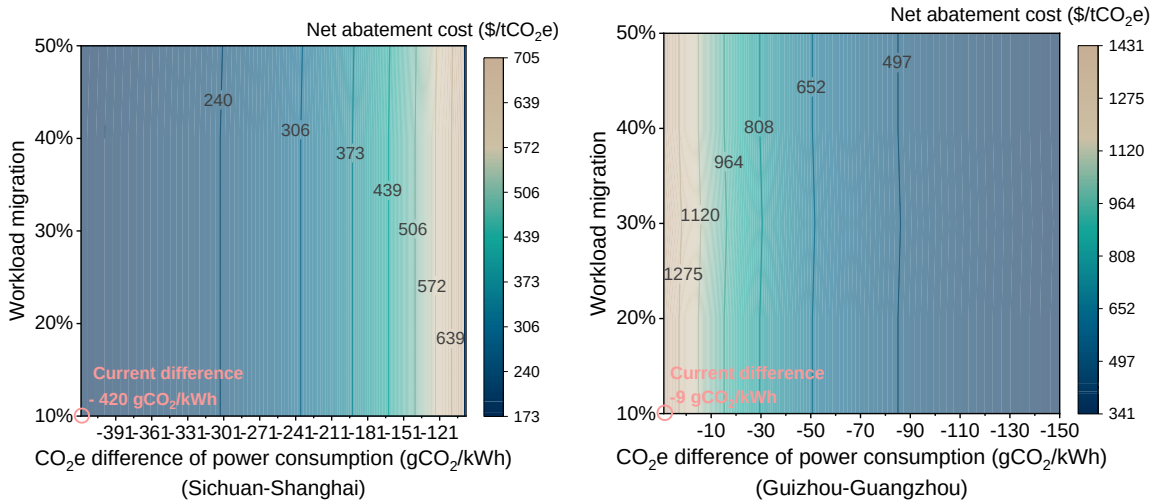


Fig. 7.6 Net carbon abatement cost considering future potential changes in CO2e factor difference of power consumption

(A) Net carbon abatement cost for Shanghai-Sichuan route.

(B) Net carbon abatement cost for Guangzhou-Guizhou route.

7.7 Discussion and policy implications

There are some other challenges and concerns faced in the successful implementation of the EDWC project, such as data confidentiality and security, efficient workload scheduling and timely transmission of workloads. A combination of policy formulation, technological investment, and strategic planning is essential for addressing these challenges to ensure the successful implementation of the EDWC project.

Firstly, ensuring the confidentiality of sensitive data is a major challenge when transferring batch computing workloads across geographically distributed data centers. Addressing data confidentiality necessitates the formulation of stringent data encryption policies. This involves mandating the use of cutting-edge encryption technologies for both data at rest and in transit, such as homomorphic encryption [263]. In addition, aggregation protocols, advanced encryption technology and financial incentives could potentially address confidentiality-related concerns [1].

Secondly, the challenge of efficient workload scheduling in geo-distributed data centers calls for the development and implementation of dynamic scheduling algorithms. These algorithms must be capable of adapting to changing network conditions and workload demands, optimizing for both cost and latency. Policies prioritizing resource allocation based on the criticality and latency-sensitivity of workloads are needed to ensure that high-priority tasks are

completed first. Additionally, a framework for conducting cost-benefit analysis is needed to make informed decisions when balancing the trade-offs between cost, performance, and latency.

Thirdly, current technologies are still far from reaching the level of timely transmission of batch computing workloads without data loss. To fulfill the need for timely transmission of batch computing workloads without data loss, significant investment in high-speed network infrastructure is required. Exploring data transfer optimization techniques, such as data compression and deduplication, can minimize the volume of data needing transmission.

Furthermore, the inconvenience and higher cost of employing IT experts locally is another practical issue of concern. Establishing training and development programs to upskill local IT staff will reduce reliance on external experts. Investing in remote collaboration and support tools would enable remote experts in the East to effectively assist local teams.

7.8 Summary

This chapter presents a comprehensive assessment of the impact of China's national initiative 'Eastern Data, Western Computing' on energy, economic, and carbon emission aspects. By analyzing the three major migration routes, we found that 'moving bits' is much more energy efficient than 'moving watts', but not necessarily beneficial in decarbonization. The results provide valuable insights into the potential environmental and economic implications of this national initiative.

All migration routes show significant energy-saving potential, ranging from 332 GWh to 942 GWh per year. The energy consumption associated with data transmission in both scenarios is significantly lower than the energy savings due to the reduction of cooling energy and the elimination of power transmission losses. Particularly, when dedicated data transmission lines are built, the energy consumption of data transmission can be even ignored.

The benefits of carbon emission reduction in different routes are significantly different. Both Shanghai-Sichuan and Guangzhou-Guizhou routes show a sharp decline in total carbon emissions, up to 2803 $KtCO_2e$ and 356 $KtCO_2e$ respectively. However, the Beijing-Inner Mongolia route shows a significant increase in total carbon emissions, 1164 $KtCO_2e$. Furthermore, we find that, if the CO_2e factor difference between Inner Mongolia and Beijing is reduced to 85 gCO_2/kWh, this route could also achieve carbon emission reduction. In addition, the net carbon abatement costs for Shanghai-Sichuan and Guangzhou-Guizhou routes are 174 $\$/tCO_2e$ and 1431 $\$/tCO_2e$ respectively, when considering extra capital cost for building duplicated data centers in Western regions.

The routes migrating computing workloads to renewable-heavy regions have great potential to reduce carbon emissions, whereas the routes migrating workloads to fossil-fuel-heavy regions show negative impacts. One recommendation for the governments in making policies is to expedite the transition towards low-carbon power generation in these fossil-fuel-heavy regions. Another alternative is to establish new migration routes in which computing workloads are transferred to regions with low carbon emission factors, such as Yunnan province.

Currently, no economic benefit is observed if considering the high capital costs for constructing duplicated data centers in the Western regions unless new data centers need to be built anyway. However, other potential economic benefits are visible in the near future. For example, the development of the data industry could attract other industries and businesses to these regions, thereby boosting overall economic development. Additionally, zero-carbon data centers [1] adopting containers and co-locating at renewable generation sites directly could be a promising alternative for building new data centers. They are more cost-effective and highly customizable compared to traditional data centers.

Our findings imply that the major plans of the national initiative 'EDWC' offer significant opportunities for reducing carbon emissions in the data center industry, whereas the others do not. These results serve as a crucial foundation for policy-makers, industry stakeholders, and researchers to make informed decisions and establish effective strategies for the successful implementation of the EDWC projects. Moreover, our research has broader applicability beyond China, extending to other regions worldwide. As the need for cloud services continues to expand globally, there is substantial potential for reducing carbon emissions through 'moving bits'. The quantitative results offer a promising direction for the global decarbonization of the data center industry.

CHAPTER 8 CONCLUSIONS AND FUTURE WORKS

8.1 Main contributions of this study

This PhD conducted a comprehensive and systematic study on the energy efficiency and sustainability of data center cooling systems over their lifespan, including i) future perspectives and global energy impact of next-generation high-temperature data centers; ii) energy performance analysis of centralized cooling systems for data centers concerning progressive loading throughout the lifecycle; iii) life-cycle optimal design and control strategies of data center cooling systems; iv) optimal dispatch strategies and design of hybrid storage systems in data centers to unlock the flexibilities of data centers; v) energy and carbon impacts of the national initiative 'Eastern Data, Western Computing'. The main contributions of this PhD study can be summarized as follows:

1) A thorough review of research and technologies for next-generation high-temperature data centers is conducted. The main benefits and the major bottlenecks for implementing high-temperature data centers as well as the existing efforts and latest technologies to tackle the bottlenecks are categorized and analyzed comprehensively. Future perspectives for the development and applications of the high-temperature data center are presented.
2) The global energy impacts of high-temperature data centers are quantified. The trade-off between cooling-energy savings and server power rise is critically analyzed. Quantitative guidance and targets are established for developing 'ideal' and 'recommendable' servers, considering the server performance associated with the thermal environment. Quantitative guidance for IT and server professionals to further develop IT equipment that takes the data center cooling energy into account.
3) A comprehensive assessment of the energy performance of centralized cooling systems in data centers is conducted. The energy performance of the cooling system is systematically analyzed under full-range cooling loads and climate conditions. The energy performance of typical cooling systems is quantified under a typical progressive loading experienced throughout the data center's lifecycle.
4) An optimal design method for centralized cooling systems with multiple chillers under progressive loading is developed. The life-cycle optimal designs in different climate zones are determined according to the energy performance under full-range loads and ambient temperatures. Free cooling hours, cooling energy, and life-cycle costs of the optimized

designs and conventional designs in different climate conditions are analyzed and compared comprehensively.

5) A pioneering approach, that leverages the surplus capacity of energy storage systems for emergencies in data centers to participate in flexible grid services under progressive loading, is proposed. Optimal dispatch strategies of energy storage systems are identified by minimizing life-cycle electricity costs. The life-cycle economic benefits of proposed design scenarios are quantified and analyzed under two electricity markets. The impacts of discount rates and battery prices on the life-cycle economic benefits of energy storage systems are comprehensively discussed.
6) A comprehensive assessment of the energy-saving potentials, economic benefits and carbon emission reductions of the national initiative 'Eastern Data, Western Computing' is conducted. The energy/emissions trade-offs associated with each route in the initiative are discussed. Future perspectives and challenges on carbon emission reduction of the initiative are analyzed. Potential policy suggestions and actionable insights are proposed to address these challenges.

8.2 Conclusions

Conclusions on the state-of-the-art research and technologies for next-generation high-temperature data centers and future perspectives

- High-temperature data centers are promising means and a major development direction for cooling energy saving particularly by changing of means of cooling fundamentally, i.e., adopting free cooling effectively. However, the choice of space temperature in most data centers is still conservative due to the concern for server reliability and performance. The major conclusions and observations could be drawn as follows:
- *Main benefits:* High-temperature data centers have great energy-saving potential due to more free cooling hours and even chiller-less/chiller-free cooling plants. Meanwhile, the high-temperature operation can provide more opportunities for waste heat recovery. For example, the waste heat could be used in neighboring buildings or a district heating system.
- *Bottlenecks and limitations:* The most frequently replaced component in servers is the hard disk drive. It needs to be further improved to enhance the reliability of servers placed in high-temperature data centers. As raising the space temperature in data centers involves an increase in initial investment of IT equipment today and a reduction in operating costs, there is a need for a trade-off between them concerning the life cycle operation cost for developing high-temperature data centers.

- *Status quo and existing efforts:* The reliability and performance of servers are more robust than most people might imagine, particularly with the advent of the class A3 and class A4 servers. The new generation of servers equipped with improved heat sinks and advanced chip materials has better performance to withstand high-temperature operation. Current technologies and feasible solutions for high-temperature data centers fall into 4 categories, including room level, rack level, server level and chip level.
- *Future perspectives:* The keys to implementing high-temperature data centers are the development and enhancement of servers and IT equipment for high-temperature operation, the awareness and deployment of new-generation servers, and the optimization of cooling systems in data centers at all levels.

Conclusions on the global energy impacts of high-temperature data centers

- Every 1K 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. 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. Meanwhile, 13%–56% of cooling energy could be saved compared with the baseline space temperature setting of 22°C, with practically no additional cooling-energy savings from further raising of the space temperature.
- Currently, the space temperature settings in most data centers remain conservative, typically at 20–25°C, while ASHRAE recommends a temperature range of 18–27°C. However, we found that the optimal space temperature in each city depends on the types 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 of a particular city and the server performance associated with the thermal environment of the chosen 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 its 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.

Conclusions on the energy performance assessment of centralized cooling systems in data centers under progressive loading

- There is a notable variance in cooling system COP at different PLRs. The difference in cooling system COP at different PLRs can be as high as 6, corresponding to a difference in PUE up to 0.14. Additionally, free cooling time could differ up to 1442 hours at different PLRs in the same location. Furthermore, on average, the cooling system COP throughout the lifecycle with a progressively increasing IT load is 2.9 points lower than the COP under design conditions.
- To address this inefficiency, it is imperative for the future design and operation of data centers to take into account the progressive nature of IT load increases. This involves the optimization of the cooling system design to match the evolving IT load profile and the implementation of adaptive control strategies to adjust cooling capacity in response to changing demands. By considering these factors, data center operators and designers can mitigate the energy inefficiencies associated with progressive loading and achieve greater energy efficiency throughout the data center's lifetime.

Conclusions on the life-cycle optimal design for centralized cooling systems with multiple chillers under progressive loading

- Although the highest cooling energy savings are achieved when the number of cooling units is 7, the capital cost for purchasing additional cooling equipment, particularly chillers, is higher. From the perspective of life-cycle cost, the optimal number of cooling units is determined to be 4.
- Under the CPS control strategy, the worldwide cooling system COP can be enhanced by 0.7-2.9, and the corresponding cooling energy can be saved by 13-22%, depending on climate conditions. The total cost savings worldwide are estimated to be 9.7-13.8% and data center PUE can be reduced by 0.06-0.1. Despite a decrease in free cooling hours (i.e., 13-860), the cooling system operates more energy-efficiently over its lifespan when adopting the optimized design.
- For the near-optimal control strategy OPR, there are still 4-9% cooling energy savings, a 0.3-0.9 increase in cooling system COP, and 2.5-6.4% cost savings over the data center

lifetime. This highlights the importance of both optimal design and optimal control strategy for efficient operation over the data center life cycle.

Conclusions on the optimal dispatch and design of energy storage systems concerning progressive loading in data centers

- Under both the Guangdong electricity market and the CAISO electricity market, the charging and discharging occur at low and high tariffs, respectively, thereby facilitating energy arbitrage. Furthermore, under the CAISO electricity market, a significant portion of the surplus battery capacity is allocated to provide frequency regulation services. In contrast, under the Guangdong electricity market, the majority of the surplus battery capacity is dedicated to energy arbitrage. This can be attributed to higher rewards for providing regulation services in the CAISO market than in the Guangdong market.
- Under both electricity markets, Scenario 1 emerges as the optimal option with the highest life-cycle economic benefits. Specifically, under the Guangdong electricity market, the life-cycle economic benefits of EES and TES are $-86,418 and $205,213, respectively. Under the CAISO electricity market, the life-cycle economic benefits of EES and TES are $361,453 and $36,985, respectively. Notably, the results reveal that EES yields greater economic benefits under the CAISO market, while TES yields greater economic benefits under the Guangdong market.
- However, the results might be significantly different when the discount rate and annual decline rate of battery price vary. Typically, staged investments accompanied by revenues from grid services are more likely to yield positive returns with lower discount rates and higher annual decline rates of battery price. These results can be elucidated by examining the present value of future cash flows; as the discount rate escalates, the present value of both future earnings and investments correspondingly decreases.

Conclusions on the energy, economic and carbon benefits the 'Eastern Data, Western Computing' on China's data centers

- All migration routes show significant energy-saving potential, ranging from 332 GWh to 942 GWh per year. The energy consumption associated with data transmission in both scenarios is significantly lower than the energy savings due to the reduction of cooling energy and the elimination of power transmission losses. Particularly, when dedicated data transmission lines are built, the energy consumption of data transmission can be ignored.

- The benefits of carbon emission reduction in different routes are significantly different. Both Shanghai-Sichuan and Guangzhou-Guizhou routes show a sharp decline in total carbon emissions, up to 2803 $KtCO_2e$ and 356 $KtCO_2e$ respectively. However, the Beijing-Inner Mongolia route shows a significant increase in total carbon emissions, 1164 $KtCO_2e$. Furthermore, we find that, if the CO_2e factor difference between Inner Mongolia and Beijing is reduced to 85 gCO_2/kWh, this route could also achieve carbon emission reduction. In addition, the net carbon abatement costs for Shanghai-Sichuan and Guangzhou-Guizhou routes are 174 $\$/tCO_2e$ and 1431 $\$/tCO_2e$ respectively, when considering extra capital cost for building duplicated data centers in Western regions.
- The routes migrating computing workloads to renewable-heavy regions have great potential to reduce carbon emissions, whereas the routes migrating workloads to fossil-fuel-heavy regions show negative impacts. One recommendation for the governments in making policies is to expedite the transition towards low-carbon power generation in these fossil-fuel-heavy regions. Another alternative is to establish new migration routes in which computing workloads are transferred to regions with low carbon emission factors, such as Yunnan province.
- Currently, no economic benefit is observed if considering the high capital costs for constructing duplicated data centers in the Western regions unless new data centers need to be built anyway. However, other potential economic benefits are visible in the near future. For example, the development of the data industry could attract other industries and businesses to these regions, thereby boosting overall economic development. Additionally, zero-carbon data centers adopting containers and co-locating at renewable generation sites directly could be a promising alternative for building new data centers. They are more cost-effective and highly customizable compared to traditional data centers.

8.3 Recommendation for future work

Despite valuable findings in this thesis, there are still certain limitations and incomplete aspects, which are recommended for future research and improvement.

1) High-temperature data centers offer significant potential for energy savings. However, the current stage of server technology is not yet prepared for large-scale application and deployment in high-temperature environments. Exploring advanced materials, such as GiN (gallium nitride), which belongs to the third-generation semiconductor materials, holds promising potential as a breakthrough for electronic components in high-temperature environments.

2) The energy assessment models adopt the water-cooled chiller cooling systems, the most widely used central chiller plants for data centers today. The conclusions of this study are therefore limited to the implementation of air-cooled data centers using water-cooled chiller plants. Future research efforts can concentrate on liquid-cooled data centers and investigate the potential benefits associated with operating at high temperatures. By exploring the benefits and energy performance of liquid-cooled data centers under high-temperature conditions, researchers can contribute to a more comprehensive understanding of alternative cooling technologies and their potential for improved energy efficiency and sustainability in the data center industry.
3) This thesis focuses on a typical progressive load profile to assess energy performance and determine optimal designs for cooling and energy storage systems in data centers. However, in practical scenarios, data centers experience different progressive loading throughout their lifecycle. There is a necessity to make a comprehensive energy performance assessment to design optimal systems for a specific progressive loading profile. Future research should emphasize the importance of customized energy performance assessments to inform optimal designs and control strategies for specific progressive loading profiles.
4) The efficient scheduling of workloads in geo-distributed data centers presents a significant challenge that necessitates the development of reliable and timely transmission of batch computing workloads. Addressing this challenge requires substantial investment in high-speed network infrastructure to ensure the seamless and timely transfer of data. To optimize data transfer processes, it is essential to explore techniques such as data compression and deduplication, which can effectively reduce the volume of data that needs to be transmitted.

REFERENCE

[1] Zheng J, Chien AA, Suh S. Mitigating curtailment and carbon emissions through load migration between data centers. Joule. 2020;4(10):2208-22.

[2] ASHRAE (2009). Best Practices for Datacom Facility Energy Efficiency, Second Edition.

[3] Masanet E, Shehabi A, Lei N, Smith S, Koomey J. Recalibrating global data center energy-use estimates. Science. 2020;367(6481):984-6.

[4] https://www.iea.org/reports/data-centres-and-data-transmission-networks.

[5] Mytton D, Ashtine M. Sources of data center energy estimates: A comprehensive review. Joule. 2022;6(9):2032-56.

[6] Mitchell-Jackson J, Koomey JG, Nordman B, Blazek M. Data center power requirements: measurements from Silicon Valley. Energy. 2003;28(8):837-50.

[7] Luo Y, Andresen J, Clarke H, Rajendra M, Maroto-Valer M. A decision support system for waste heat recovery and energy efficiency improvement in data centres. Applied Energy. 2019;250:1217-24.

[8] Chen X, Tu R, Li M, Yang X, Jia K. Hot spot temperature prediction and operating parameter estimation of racks in data center using machine learning algorithms based on simulation data. Building Simulation. 2023;16(11):2159-76.

[9] Gupta R, Asgari S, Moazamigoodarzi H, Pal S, Puri IK. Cooling architecture selection for air-cooled Data Centers by minimizing exergy destruction. Energy. 2020;201:117625.

[10] Capozzoli A, Primiceri G. Cooling Systems in Data Centers: State of Art and Emerging Technologies. Energy Procedia. 2015;83:484-93.

[11] Liu Y, Yang X, Li J, Zhao X. Energy savings of hybrid dew-point evaporative cooler and micro-channel separated heat pipe cooling systems for computer data centers. Energy. 2018;163:629-40.

[12] Zhou W, Sun Q, Luo W, Xiao W, Cui P, Wu W, et al. Performance analysis and optimization of free cooling strategies for a liquid-cooled data center. Building Simulation. 2023;16(8):1317-30.

[13] Khalaj AH, Halgamuge SK. A Review on efficient thermal management of air-and liquid-cooled data centers: From chip to the cooling system. Applied energy. 2017;205:1165-88.

[14] ASHRAE. Thermal Guidelines for Data Processing Environments, Fifth Edition2021.

[15] Niemann J, Bean J, Avelar V. Economizer modes of data center cooling systems. Schneider Electric Data Center Science Center Whitepaper. 2011;160.

[16] Taylor ST. How to design & control waterside economizers. ASHRAE Journal. 2014;56(6):30-6.

[17] 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.

[18] Fleischer AS. Cooling our insatiable demand for data. Science. 2020;370(6518):783-4.

[19] Niemann J, Brown K, Avelar V. Impact of hot and cold aisle containment on data center temperature and efficiency. Schneider Electric Data Center Science Center, White Paper. 2011;135:1-14.

[20] Habibi Khalaj A, Scherer T, Siriwardana J, Halgamuge SK. Multi-objective efficiency enhancement using workload spreading in an operational data center. Applied Energy. 2015;138:432-44.

[21] Asgari S, MirhoseiniNejad S, Moazamigoodarzi H, Gupta R, Zheng R, Puri IK. A gray-box model for real-time transient temperature predictions in data centers. Applied Thermal Engineering. 2021;185:116319.

[22] Chu W-X, Wang C-C. A review on airflow management in data centers. Applied Energy. 2019;240:84-119.

[23] Zhang Y, Zhang K, Liu J, Kosonen R, Yuan X. Airflow uniformity optimization for modular data center based on the constructal T-shaped underfloor air ducts. Applied Thermal Engineering. 2019;155:489-500.

[24] Chu W-X, Wang R, Hsu P-H, Wang C-C. Assessment on rack intake flowrate uniformity of data center with cold aisle containment configuration. Journal of Building Engineering. 2020;30:101331.

[25] Tatchell-Evans M, Kapur N, Summers J, Thompson H, Oldham D. An experimental and theoretical investigation of the extent of bypass air within data centres employing aisle containment, and its impact on power consumption. Applied Energy. 2017;186:457-69.

[26] MirhoseiniNejad S, Moazamigoodarzi H, Badawy G, Down DG. Joint data center cooling and workload management: A thermal-aware approach. Future Generation Computer Systems. 2020;104:174-86.

[27] Gupta R, Moazamigoodarzi H, MirhoseiniNejad S, Down DG, Puri IK. Workload management for air-cooled data centers: An energy and exergy based approach. Energy. 2020;209:118485.

[28] Gandhi A, Harchol-Balter M, Das R, Lefurgy C. Optimal power allocation in server farms. ACM SIGMETRICS Performance Evaluation Review. 2009;37:157-68.

[29] ASHRAE. Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance2011.

[30] Miller R. Raise your data center temperature, https://www.datacenterknowledge.com/archives/2008/10/14/google-raise-your-data-center-temperature. 2008.

[31] Breen TJ, Walsh EJ, Punch J, Shah AJ, Bash CE. From chip to cooling tower data center modeling: Part I influence of server inlet temperature and temperature rise across cabinet. Conference From chip to cooling tower data center modeling: Part I influence of server inlet temperature and temperature rise across cabinet. IEEE, p. 1-10.

[32] El-Sayed N, Stefanovici IA, Amvrosiadis G, Hwang AA, Schroeder B. Temperature management in data centers: Why some (might) like it hot. Conference Temperature management in data centers: Why some (might) like it hot. p. 163-74.

[33] Fitch J. Dell's next generation servers: Pushing the limits of data center cooling cost savings, https://studylib.net/doc/18559703/pushing-the-limits-of-data-center-cooling-cost-savings. DELL Enterprise Reliability Engineering; 2012.

[34] Vallury A, Matteson J. Data Center Trends Toward Higher Ambient Inlet Temperatures and the Impact on Server Performance. Conference Data Center Trends Toward Higher Ambient Inlet Temperatures and the Impact on Server Performance, vol. 55768. American Society of Mechanical Engineers, p. V002T09A10.

[35] Beaty DL, Quirk D. De-risking data center temperature increases, Part 1. ASHRAE Journal. 2016;58(1):74-82.

[36] Beaty DL, Quirk D. De-risking data center temperature increases, Part 2. ASHRAE Journal. 2016;58(3):70-5.

[37] He Y, Chen G, Wei W, Liu Q, Zhang J, Zhou T, et al. HTA corrosion resistant technology for free cooling. Conference HTA corrosion resistant technology for free cooling. p. 120-6.

[38] Miller R. Too Hot for Humans, But Google Servers Keep Humming https://www.datacenterknowledge.com/archives/2012/03/23/too-hot-for-humans-but-google-servers-keep-humming. 2012.

[39] Goiri I, Nguyen T, Bianchini R. CoolAir: Temperature- and Variation-Aware Management for Free-Cooled Datacenters. ACM SIGPLAN Notices. 2015;50:253-65.

[40] Sankar S, Shaw M, Vaid K, Gurumurthi S. Datacenter scale evaluation of the impact of temperature on hard disk drive failures. ACM Transactions on Storage (TOS). 2013;9(2):1-24.

[41] Li J, Li Z. Model-based optimization of free cooling switchover temperature and cooling tower approach temperature for data center cooling system with water-side economizer. Energy and Buildings. 2020;227:110407.

[42] Kelley C, Singh CH, Smith V. Data center efficiency and IT equipment reliability at wider operating temperature and humidity ranges (White Paper). The Green Grid; 2012.

[43] Ahuja N. Datacenter power savings through high ambient datacenter operation: CFD modeling study. Conference Datacenter power savings through high ambient datacenter operation: CFD modeling study. p. 104-7.

[44] Dubin FS, Mindell HL, Bloome S. How to save energy and cut costs in existing industrial and commercial buildings: an energy conservation manual. United States: Department of Energy, 1976.

[45] Squillo T. 3 Ways To Increase Chiller Efficiency https://www.facilitiesnet.com/hvac/article/3-Ways-To-Increase-Chiller-Efficiency--17669. 2018.

[46] Liu Y, Wei X, Xiao J, Liu Z, Xu Y, Tian Y. Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers. Global Energy Interconnection. 2020;3(3):272-82.

[47] Daraghmeh HM, Wang C-C. A review of current status of free cooling in datacenters. Applied Thermal Engineering. 2017;114:1224-39.

[48] Fernández-Montes A, Fernández-Cerero D, González-Abril L, Álvarez-García JA, Ortega JA. Energy wasting at internet data centers due to fear. Pattern Recognition Letters. 2015;67:59-65.

[49] Harvey T, Patterson M, Bean J. Updated air-side free cooling maps: The impact of ASHRAE 2011 allowable ranges (White Paper). The Green Grid; 2012.

[50] Ebrahimi K, Jones GF, Fleischer AS. A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. Renewable and Sustainable Energy Reviews. 2014;31:622-38.

[51] Miller R. Server Failures Don't Rise Along With the Heat https://www.datacenterknowledge.com/archives/2012/05/29/study-server-failures-dont-rise-along-with-the-heat. 2012.

[52] Zhang S, Ahuja N, Han Y, Ren H, Chen Y, Guo G. Key considerations to implement high ambient data center. Conference Key considerations to implement high ambient data center. p. 147-54.

[53] Mone G. Redesigning the data center. Communications of the ACM. 2012;55(10):14-6.

[54] Miller R. In Dublin, Cool Climate Fuels Cloud Computing Cluster https://www.datacenterknowledge.com/archives/2013/04/08/dublin-free-cooling. 2013.

[55] Don Atwood JGM. 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. 2008.

[56] He Y, Chen G, Zhang J, Zhou T, Liu T, Zhu P, et al. Consideration for Running Data Center at High Temperatures and Using Free Air Cooling. Conference Consideration for Running Data Center at High Temperatures and Using Free Air Cooling. American Society of Mechanical Engineers, p. V001T09A14.

[57] Schroeder B, Gibson GA. Understanding disk failure rates: What does an MTTF of 1,000,000 hours mean to you? ACM Trans Storage. 2007;3(3):8–es.

[58] Schroeder B, Gibson GA. A Large-Scale Study of Failures in High-Performance Computing Systems. IEEE Transactions on Dependable and Secure Computing. 2010;7(4):337-50.

[59] Sankar S, Shaw M, Vaid K. Impact of temperature on hard disk drive reliability in large datacenters. Conference Impact of temperature on hard disk drive reliability in large datacenters. p. 530-7.

[60] Pinheiro E, Weber W-D, Barroso LA. Failure trends in a large disk drive population. 5th USENIX Conference on File and Storage Technologies: USENIX Association; 2007.

[61] Chan CS, Jin Y, Wu Y-K, Gross K, Vaidyanathan K, Rosing Ti. Fan-speed-aware scheduling of data intensive jobs. Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design. Redondo Beach, California, USA: Association for Computing Machinery; 2012. p. 409–14.

[62] Schroeder B, Merchant A, Lagisetty R. Reliability of nand-Based SSDs: What Field Studies Tell Us. Proceedings of the IEEE. 2017;105(9):1751-69.

[63] Chan CS, Pan B, Gross K, Vaidyanathan K, Rosing TŠ. Correcting vibration-induced performance degradation in enterprise servers. ACM SIGMETRICS Performance Evaluation Review. 2014;41(3):83-8.

[64] Al-Ars Z, van de Goor AJ, Braun J, Richter D. Simulation based analysis of temperature effect on the faulty behavior of embedded drams. Conference Simulation based analysis of temperature effect on the faulty behavior of embedded drams. IEEE, p. 783-92.

[65] Hamamoto T, Sugiura S, Sawada S. On the retention time distribution of dynamic random access memory (DRAM). IEEE Transactions on Electron devices. 1998;45(6):1300-9.

[66] Schroeder B, Pinheiro E, Weber W-D. DRAM errors in the wild: a large-scale field study. ACM SIGMETRICS Performance Evaluation Review. 2009;37:193-204.

[67] Wang Y. Evaluating and modeling the energy impacts of data centers, in terms of hardware/software architecture and associated environment: Ecole nationale supérieure Mines-Télécom Atlantique, 2020.

[68] Zapater M, Tuncer O, Ayala JL, Moya JM, Vaidyanathan K, Gross K, et al. Leakage-Aware Cooling Management for Improving Server Energy Efficiency. IEEE Transactions on Parallel and Distributed Systems. 2015;26(10):2764-77.

[69] Moss D, Bean JH. Energy impact of increased server inlet temperature (APC white paper). American Power Conversion; 2009.

[70] Yeo S, Hossain MM, Huang J-C, Lee H-HS. ATAC: Ambient Temperature-Aware Capping for Power Efficient Datacenters. Proceedings of the ACM Symposium on Cloud Computing. Seattle, WA, USA: Association for Computing Machinery; 2014. p. 1–14.

[71] Patterson MK. The effect of data center temperature on energy efficiency. Conference The effect of data center temperature on energy efficiency. p. 1167-74.

[72] Wan J, Gui X, Zhang R, Fu L. Joint Cooling and Server Control in Data Centers: A Cross-Layer Framework for Holistic Energy Minimization. IEEE Systems Journal. 2018;12(3):2461-72.

[73] Kim NS, Austin T, Baauw D, Mudge T, Flautner K, Hu JS, et al. Leakage current: Moore's law meets static power. Computer. 2003;36(12):68-75.

[74] Shin D, Kim J, Chang N, Choi J, Chung SW, Chung E. Energy-optimal dynamic thermal management for green computing. Conference Energy-optimal dynamic thermal management for green computing. p. 652-7.

[75] Agarwal A, Mukhopadhyay S, Kim C, Raychowdhury A, Roy K. Leakage power analysis and reduction: models, estimation and tools. IEE Proceedings-Computers and Digital Techniques. 2005;152(3):353-68.

[76] Fallah F, Pedram M. Standby and active leakage current control and minimization in CMOS VLSI circuits. IEICE transactions on electronics. 2005;88(4):509-19.

[77] Seaton I. 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/. 2017.

[78] Muroya K, Kinoshita T, Tanaka H, Youro M. Power reduction effect of higher room temperature operation in data centers. Conference Power reduction effect of higher room temperature operation in data centers. p. 661-73.

[79] Seaton I. Airflow Management Considerations for a New Data Center – Part 3: Server Cost vs Inlet Temperature https://www.upsite.com/blog/server-cost-vs-inlet-temperature/. 2017.

[80] Rubenstein B, Faist M. Data center cold aisle set point optimization through total operating cost modeling. Conference Data center cold aisle set point optimization through total operating cost modeling. p. 1111-20.

[81] Nada SA, Elfeky KE. Experimental investigations of thermal managements solutions in data centers buildings for different arrangements of cold aisles containments. Journal of Building Engineering. 2016;5:41-9.

[82] Niemann J, Brown K, Avelar V. Impact of hot and cold aisle containment on data center temperature and efficiency (APC White Paper). Schneider Electric Data Center Science Center; 2011.

[83] Sundaralingam V, Arghode VK, Joshi Y, Phelps W. Experimental characterization of various cold aisle containment configurations for data centers. Journal of electronic packaging. 2015;137(1):011007.

[84] Gondipalli S, Bhopte S, Sammakia B, Iyengar MK, Schmidt R. Effect of isolating cold aisles on rack inlet temperature. Conference Effect of isolating cold aisles on rack inlet temperature. p. 1247-54.

[85] Gao C, Yu Z, Wu J. Investigation of Airflow Pattern of a Typical Data Center by CFD Simulation. Energy Procedia. 2015;78:2687-93.

[86] Cho J, Yang J, Park W. Evaluation of air distribution system's airflow performance for cooling energy savings in high-density data centers. Energy and Buildings. 2014;68:270-9.

[87] Dunlap K, Rasmussen N. Choosing between room, row, and rack-based cooling for data centers (APC White Paper). Schneider Electric Data Center Science Center; 2012.

[88] Nada SA, Said MA. Effect of CRAC units layout on thermal management of data center. Applied Thermal Engineering. 2017;118:339-44.

[89] Schmidt RR, Iyengar M, Vogel M, Pienta B. Comparison between underfloor supply and overhead supply ventilation designs for data center high-density clusters/discussion. ASHRAE Transactions. 2007;113:115.

[90] Nakao M, Hayama H, Nishioka M. Which cooling air supply system is better for a high heat density room: underfloor or overhead? Conference Which cooling air supply system is better for a high heat density room: underfloor or overhead? IEEE, p. 393-400.

[91] Bhopte S, Agonafer D, Schmidt R, Sammakia B. Optimization of Data Center Room Layout to Minimize Rack Inlet Air Temperature. Journal of Electronic Packaging. 2006;128(4):380-7.

[92] Lu H, Zhang Z. Numerical and experimental investigations on the thermal performance of a data center. Applied Thermal Engineering. 2020;180:115759.

[93] Schmidt R, Cruz E. Cluster of High Powered Racks Within a Raised Floor Computer Data Center: Effect of Perforated Tile Flow Distribution on Rack Inlet Air Temperatures. Conference Cluster of High Powered Racks Within a Raised Floor Computer Data Center:

Effect of Perforated Tile Flow Distribution on Rack Inlet Air Temperatures, vol. Heat Transfer, Volume 2. p. 245-62.

[94] VanGilder JW, Schmidt RR. Airflow uniformity through perforated tiles in a raised-floor data center. Conference Airflow uniformity through perforated tiles in a raised-floor data center, vol. 42002. p. 493-501.

[95] Seymour M. Modeling perforated tiles in data centers – What is required? Conference Modeling perforated tiles in data centers – What is required? p. 8-12.

[96] Khalili S, Tradat MI, Nemati K, Seymour M, Sammakia B. Impact of tile design on the thermal performance of open and enclosed aisles. Journal of Electronic Packaging. 2018;140(1):010907.

[97] Mulay V, Agonafer D, Irwin G, Patell D. Effective thermal management of data centers using efficient cabinet designs. Conference Effective thermal management of data centers using efficient cabinet designs, vol. 43604. p. 993-9.

[98] Rolander N, Rambo J, Joshi Y, Allen JK, Mistree F. An Approach to Robust Design of Turbulent Convective Systems. Journal of Mechanical Design. 2006;128(4):844-55.

[99] Pang W, Wang C, Ahuja N, Zhang J, Zhou A, Si P, et al. An advanced energy efficient rack server design. Conference An advanced energy efficient rack server design. IEEE, p. 806-14.

[100] Ling Y, Zhang X, Li S, Han G, Sun X. The numerical simulation and experimental study on rack cooling effect in date center with UFAD system. International Journal of Low-Carbon Technologies. 2015;10(4):446-51.

[101] Wang IN, Tsui Y-Y, Wang C-C. Improvements of Airflow Distribution in a Container Data Center. Energy Procedia. 2015;75:1819-24.

[102] Jin C, Bai X. The study of servers' arrangement and air distribution strategy under partial load in data centers. Sustainable Cities and Society. 2019;49:101617.

[103] Yu Y, Wang L. Solid sorption heat pipe coupled with direct air cooling technology for thermal control of rack level in internet data centers: Design and numerical simulation. International Journal of Heat and Mass Transfer. 2019;145:118714.

[104] Li X, Zhang C, Sun X, Han Z, Wang S. Experimental study on reliable operation strategy for multi-split backplane cooling system in data centers. Applied Thermal Engineering. 2022;211:118494.

[105] Tian H, He Z, Li Z. A combined cooling solution for high heat density data centers using multi-stage heat pipe loops. Energy and Buildings. 2015;94:177-88.

[106] Siddarth AV. Experimental Study on Effects of Segregated Cooling Provisioning on Thermal Performance of Information Technology Servers in Air Cooled Data Centers: The University of Texas, 2015.

[107] Sakanova A, Alimohammadi S, McEvoy J, Battaglioli S, Persoons T. Multi-objective layout optimization of a generic hybrid-cooled data centre blade server. Applied Thermal Engineering. 2019;156:514-23.

[108] Sarma ANSVS, Ambali VD. Cooling solution for computing and storage server. Conference Cooling solution for computing and storage server. p. 840-9.

[109] Iyengar M, David M, Parida P, Kamath V, Kochuparambil B, Graybill D, et al. Server liquid cooling with chiller-less data center design to enable significant energy savings. Conference Server liquid cooling with chiller-less data center design to enable significant energy savings. p. 212-23.

[110] Khalili S, Alissa H, Nemati K, Seymour M, Curtis R, Moss D, et al. Impact of server thermal design on the cooling efficiency: Chassis design. Journal of Electronic Packaging. 2019;141(3):031004.

[111] Lin J, Zheng H, Zhu Z, Gorbatov E, David H, Zhang Z. Software thermal management of DRAM memory for multicore systems. ACM SIGMETRICS Performance Evaluation Review. 2008;36(1):337-48.

[112] Wang L, Wang ZL. Advances in piezotronic transistors and piezotronics. Nano Today. 2021;37:101108.

[113] Habibi Khalaj A, Halgamuge SK. A Review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system. Applied Energy. 2017;205:1165-88.

[114] Feng S, Yan Y, Li H, He Z, Zhang L. Reprint of: Temperature Uniformity Enhancement and Flow Characteristics of Embedded Gradient Distribution Micro Pin Fin Arrays Using Dielectric Coolant for Direct Intra-Chip Cooling. International Journal of Heat and Mass Transfer. 2020;161:120235.

[115] Kheirabadi AC, Groulx D. Cooling of server electronics: A design review of existing technology. Applied Thermal Engineering. 2016;105:622-38.

[116] Sarafraz MM, Arya A, Hormozi F, Nikkhah V. On the convective thermal performance of a CPU cooler working with liquid gallium and CuO/water nanofluid: A comparative study. Applied Thermal Engineering. 2017;112:1373-81.

[117] Zhao N, Guo L, Qi C, Chen T, Cui X. Experimental study on thermo-hydraulic performance of nanofluids in CPU heat sink with rectangular grooves and cylindrical bugles based on exergy efficiency. Energy Conversion and Management. 2019;181:235-46.

[118] Zhao N, Qi C, Chen T, Tang J, Cui X. Experimental study on influences of cylindrical grooves on thermal efficiency, exergy efficiency and entropy generation of CPU cooled by nanofluids. International Journal of Heat and Mass Transfer. 2019;135:16-32.

[119] Choi J, Jeong M, Yoo J, Seo M. A new CPU cooler design based on an active cooling heatsink combined with heat pipes. Applied Thermal Engineering. 2012;44:50-6.

[120] Wang Y, Wang B, Zhu K, Li H, He W, Liu S. Energy saving potential of using heat pipes for CPU cooling. Applied Thermal Engineering. 2018;143:630-8.

[121] Hu HM, Ge TS, Dai YJ, Wang RZ. Experimental study on water-cooled thermoelectric cooler for CPU under severe environment. International Journal of Refrigeration. 2016;62:30-8.

[122] Poachaiyapoom A, Leardkun R, Mounkong J, Wongwises S. Miniature vapor compression refrigeration system for electronics cooling. Case Studies in Thermal Engineering. 2019;13:100365.

[123] Deng Y, Zhang M, Jiang Y, Liu J. Two-stage multichannel liquid–metal cooling system for thermal management of high-heat-flux-density chip array. Energy Conversion and Management. 2022;259:115591.

[124] Liang K, Li Z, Chen M, Jiang H. Comparisons between heat pipe, thermoelectric system, and vapour compression refrigeration system for electronics cooling. Applied Thermal Engineering. 2019;146:260-7.

[125] Chowdhury I, Prasher R, Lofgreen K, Chrysler G, Narasimhan S, Mahajan R, et al. On-chip cooling by superlattice-based thin-film thermoelectrics. Nature Nanotechnology. 2009;4(4):235-8.

[126] Naphon P, Wongwises S, Wiriyasart S. Application of two-phase vapor chamber technique for hard disk drive cooling of PCs. International Communications in Heat and Mass Transfer. 2013;40:32-5.

[127] Lei N, Masanet E. Statistical analysis for predicting location-specific data center PUE and its improvement potential. Energy. 2020;201:117556.

[128] Díaz AJ, Cáceres R, Torres R, Cardemil JM, Silva-Llanca L. Effect of climate conditions on the thermodynamic performance of a data center cooling system under water-side economization. Energy and Buildings. 2020;208:109634.

[129] Cheung H, Wang S. Optimal design of data center cooling systems concerning multi-chiller system configuration and component selection for energy-efficient operation and maximized free-cooling. Renewable Energy. 2019;143:1717-31.

[130] Li Y, Wen Y, Tao D, Guan K. Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning. IEEE Transactions on Cybernetics. 2020;50(5):2002-13.

[131] Han Z, Wei H, Sun X, Bai C, Xue D, Li X. Study on influence of operating parameters of data center air conditioning system based on the concept of on-demand cooling. Renewable Energy. 2020;160:99-111.

[132] Ran Y, Hu H, Zhou X, Wen Y. DeepEE: Joint Optimization of Job Scheduling and Cooling Control for Data Center Energy Efficiency Using Deep Reinforcement Learning. Conference DeepEE: Joint Optimization of Job Scheduling and Cooling Control for Data Center Energy Efficiency Using Deep Reinforcement Learning. p. 645-55.

[133] Yu J, Jiang Y, Yan Y. A simulation study on heat recovery of data center: A case study in Harbin, China. Renewable Energy. 2019;130:154-73.

[134] Chen X, Wang X, Ding T, Li Z. Experimental research and energy saving analysis of an integrated data center cooling and waste heat recovery system. Applied Energy. 2023;352:121875.

[135] Rostirolla G, Grange L, Minh-Thuyen T, Stolf P, Pierson JM, Da Costa G, et al. A survey of challenges and solutions for the integration of renewable energy in datacenters. Renewable and Sustainable Energy Reviews. 2022;155:111787.

[136] He W, Xu Q, Liu S, Wang T, Wang F, Wu X, et al. Analysis on data center power supply system based on multiple renewable power configurations and multi-objective optimization. Renewable Energy. 2024;222:119865.

[137] Han O, Ding T, Zhang X, Mu C, He X, Zhang H, et al. A shared energy storage business model for data center clusters considering renewable energy uncertainties. Renewable Energy. 2023;202:1273-90.

[138] Liang Y, Lin X, Su W, Xing L, Li W, Wang B. Preliminary design and optimization of a solar-driven combined cooling and power system for a data center. Energy Conversion and Management: X. 2023;20:100409.

[139] Zhou Z, Liu F, Chen S, Li Z. A truthful and efficient incentive mechanism for demand response in green datacenters. IEEE Transactions on Parallel and Distributed Systems. 2018;31(1):1-15.

[140] Wang S, Qian Z, Yuan J, You I. A DVFS based energy-efficient tasks scheduling in a data center. Ieee Access. 2017;5:13090-102.

[141] Kontorinis V, Zhang LE, Aksanli B, Sampson J, Homayoun H, Pettis E, et al. Managing distributed ups energy for effective power capping in data centers. ACM SIGARCH Computer Architecture News. 2012;40(3):488-99.

[142] Chen H, Caramanis MC, Coskun AK. Reducing the data center electricity costs through participation in smart grid programs. Conference Reducing the data center electricity costs through participation in smart grid programs. p. 1-10.

[143] Zhang Y, Wilson DC, Paschalidis IC, Coskun AK. A data center demand response policy for real-world workload scenarios in HPC. Conference A data center demand response policy for real-world workload scenarios in HPC. IEEE, p. 282-7.

[144] Mohsenian-Rad AH, Leon-Garcia A. Coordination of Cloud Computing and Smart Power Grids. Conference Coordination of Cloud Computing and Smart Power Grids. p. 368-72.

[145] Hu H, Wen Y, Yin L, Qiu L, Niyato D. Coordinating Workload Scheduling of Geo-Distributed Data Centers and Electricity Generation of Smart Grid. IEEE Transactions on Services Computing. 2020;13(6):1007-20.

[146] Ghatikar G. Demand response opportunities and enabling technologies for data centers: Findings from field studies. 2012.

[147] Fu Y, Han X, Baker K, Zuo W. Assessments of data centers for provision of frequency regulation. Applied Energy. 2020;277:115621.

[148] Mares K. Demand response and open automated demand response opportunities for data centers. 2009.

[149] Zhang Y, Wang Y, Wang X. TEStore: Exploiting thermal and energy storage to cut the electricity bill for datacenter cooling. Conference TEStore: Exploiting thermal and energy storage to cut the electricity bill for datacenter cooling. IEEE, p. 19-27.

[150] Yang T, Zhao Y, Pen H, Wang Z. Data center holistic demand response algorithm to smooth microgrid tie-line power fluctuation. Applied Energy. 2018;231:277-87.

[151] Guo C, Luo F, Cai Z, Dong ZY, Zhang R. Integrated planning of internet data centers and battery energy storage systems in smart grids. Applied Energy. 2021;281:116093.

[152] Al-Amri MD, El-Gomati M, Zubairy MS. Optics in our time: Springer Nature, 2016.

[153] Richardson DJ, Fini JM, Nelson LE. Space-division multiplexing in optical fibres. Nature Photonics. 2013;7(5):354-62.

[154] Agrawal GP. Optical communication: its history and recent progress. Optics in our time. 2016:177-99.

[155] Mukherjee B. WDM optical communication networks: progress and challenges. IEEE Journal on Selected Areas in Communications. 2000;18(10):1810-24.

[156] Kahn JM, Miller DAB. Communications expands its space. Nature Photonics. 2017;11(1):5-8.

[157] Liu Z, Lin M, Wierman A, Low SH, Andrew LL. Geographical load balancing with renewables. ACM SIGMETRICS Performance Evaluation Review. 2011;39(3):62-6.

[158] Kong F, Liu X. A survey on green-energy-aware power management for datacenters. ACM Computing Surveys (CSUR). 2014;47(2):1-38.

[159] Cortez E, Bonde A, Muzio A, Russinovich M, Fontoura M, Bianchini R. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. Conference Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. p. 153-67.

[160] Ahmad I, Khalil MIK, Shah SAA. Optimization - based workload distribution in geographically distributed data centers: A survey. International Journal of Communication Systems. 2020;33(12):e4453.

[161] Rahman S, Gupta A, Tornatore M, Mukherjee B. Dynamic workload migration over backbone network to minimize data center electricity cost. IEEE Transactions on Green Communications and Networking. 2017;2(2):570-9.

[162] Gupta A, Mandal U, Chowdhury P, Tornatore M, Mukherjee B. Cost-efficient live VM migration based on varying electricity cost in optical cloud networks. Photonic Network Communications. 2015;30(3):376-86.

[163] Kong F, Liu X. A Survey on Green-Energy-Aware Power Management for Datacenters. ACM Comput Surv. 2014;47(2):Article 30.

[164] Goudarzi H, Pedram M. Force-directed geographical load balancing and scheduling for batch jobs in distributed datacenters. Conference Force-directed geographical load balancing and scheduling for batch jobs in distributed datacenters. IEEE, p. 1-8.

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

[166] Lui YY. Waterside and Airside Economizers Design Considerations for Data Center Facilities. ASHRAE Transactions. 2010;116(1):98–108.

[167] Braun JE. Methodologies for the design and control of central cooling plants: The University of Wisconsin-Madison, 1988.

[168] ASHRAE. ASHRAE Handbook HVAC Systems and Equipment2020.

[169] TRNSYS 18, A transient systems simulation program, http://sel.me.wisc.edu/trnsys.

[170] Ma Z, Wang S, Xu X, Xiao F. A supervisory control strategy for building cooling water systems for practical and real time applications. Energy Conversion and Management. 2008;49(8):2324-36.

[171] Lu Y, Wang S, Zhao Y, Yan C. Renewable energy system optimization of low/zero energy buildings using single-objective and multi-objective optimization methods. Energy and Buildings. 2015;89:61-75.

[172] Pontes RFF, Yamauchi WM, Silva EKG. Analysis of the effect of seasonal climate changes on cooling tower efficiency, and strategies for reducing cooling tower power consumption. Applied Thermal Engineering. 2019;161:114148.

[173] Taylor ST. Optimizing design & control of chilled water plants: Part 4: Chiller & cooling tower selection. ASHRAE Journal. 2012;54(3):60-70.

[174] https://energyplus.net/.

[175] ASHRAE. Standard 169-2021, Climatic Data for Building Design Standards. 2021.

[176] Zhang Y, Shan K, Li X, Li H, Wang S. Research and Technologies for next-generation high-temperature data centers – State-of-the-arts and future perspectives. Renewable and Sustainable Energy Reviews. 2023;171:112991.

[177] Pan C, Zhai J, Wang ZL. Piezotronics and Piezo-phototronics of Third Generation Semiconductor Nanowires. Chemical Reviews. 2019;119(15):9303-59.

[178] Wang S. Intelligent buildings and building automation: Routledge, 2009.

[179] Gözcü O, Özada B, Carfi MU, Erden HS. Worldwide energy analysis of major free cooling methods for data centers. Conference Worldwide energy analysis of major free cooling methods for data centers. p. 968-76.

[180] Best Practices for Datacome Facility Energy Efficiency, Second Edition. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.; 2009.

[181] Gözcü O, Erden HSJTJoEE, Sciences C. Energy and economic assessment of major free cooling retrofits for data centers in Turkey. 2019;27(3):2097-212.

[183] Ham S-W, Kim M-H, Choi B-N, Jeong J-W. Simplified server model to simulate data center cooling energy consumption. Energy and Buildings. 2015;86:328-39.

[184] Stabat P, Marchio D. Simplified model for indirect-contact evaporative cooling-tower behaviour. Applied Energy. 2004;78(4):433-51.

[185] Joshi Y, Kumar P. Energy efficient thermal management of data centers: Springer Science & Business Media, 2012.

[186] Iyengar M, Schmidt R, Caricari J. Reducing energy usage in data centers through control of Room Air Conditioning units. Conference Reducing energy usage in data centers through control of Room Air Conditioning units. p. 1-11.

[187] Malone C, Belady C. Metrics to characterize data center & IT equipment energy use. Conference Metrics to characterize data center & IT equipment energy use, vol. 35. sn.

[188] Metrics GG. Describing Datacenter Power Efficiency. Green Grid Technical Committee White Paper. 2007.

[189] Brady GA, Kapur N, Summers JL, Thompson HM. A case study and critical assessment in calculating power usage effectiveness for a data centre. Energy Conversion and Management. 2013;76:155-61.

[190] Cho J, Yang J, Lee C, Lee J. Development of an energy evaluation and design tool for dedicated cooling systems of data centers: Sensing data center cooling energy efficiency. Energy and Buildings. 2015;96:357-72.

[191] Ham S-W, Kim M-H, Choi B-N, Jeong J-W. Energy saving potential of various air-side economizers in a modular data center. Applied Energy. 2015;138:258-75.

[192] Rasmussen N. Avoiding costs from oversizing data center and network room infrastructure. Whitepaper. 2011;37:1-9.

[193] Brett Griffin PE. Data center economizer efficiency. ASHRAE JOUNAL. 2015.

[194] Rasmussen N. Calculating total cooling requirements for data centers. White paper. 2007;25:1-8.

[195] Kumar D, Raisee M, Lacor C. Combination of Polynomial Chaos with Adjoint Formulations for Optimization Under Uncertainties. In: Hirsch C, Wunsch D, Szumbarski J, Łaniewski-Wołłk Ł, Pons-Prats J, editors. Uncertainty Management for Robust Industrial Design in Aeronautics : Findings and Best Practice Collected During UMRIDA, a Collaborative Research Project (2013–2016) Funded by the European Union. Cham: Springer International Publishing; 2019. p. 567-82.

[196] Lantoine G, Russell RP. A hybrid differential dynamic programming algorithm for constrained optimal control problems. part 1: Theory. Journal of Optimization Theory and Applications. 2012;154:382-417.

[197] Cheng Q, Wang S, Yan C, Xiao F. Probabilistic approach for uncertainty-based optimal design of chiller plants in buildings. Applied Energy. 2017;185:1613-24.

[198] dos Santos Coelho L, Askarzadeh A. An enhanced bat algorithm approach for reducing electrical power consumption of air conditioning systems based on differential operator. Applied Thermal Engineering. 2016;99:834-40.

[199] Gang W, Wang S, Xiao F, Gao D-c. Robust optimal design of building cooling systems considering cooling load uncertainty and equipment reliability. Applied energy. 2015;159:265-75.

[200] Kang J, Wang S, Yan C. A new distributed energy system configuration for cooling dominated districts and the performance assessment based on real site measurements. Renewable energy. 2019;131:390-403.

[201] https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/.

[202] Ma X, Zhang Q, Zou S. An experimental and numerical study on the thermal performance of a loop thermosyphon integrated with latent thermal energy storage for emergency cooling in a data center. Energy. 2022;253:123946.

[203] http://oasis.caiso.com/mrioasis/logon.do.

[204] Celebi E, Fuller JD. Time-of-use pricing in electricity markets under different market structures. IEEE Transactions on Power Systems. 2012;27(3):1170-81.

[205] https://www.shanwei.gov.cn/swdpb/zxgk/201911/92c15ae1ab904fbfb7e93b8fd19b1694/files/85460ce1f31146acb019e23e6442f644.pdf.

[206] http://120.31.132.37:16001/portal-oa/unauth/readPdf?filePath=/biz/DOC_SEND/888.pdf.

[207] Shi Y, Xu B, Wang D, Zhang B. Using battery storage for peak shaving and frequency regulation: Joint optimization for superlinear gains. IEEE transactions on power systems. 2017;33(3):2882-94.

[208] Tang H, Wang S. Life-cycle economic analysis of thermal energy storage, new and second-life batteries in buildings for providing multiple flexibility services in electricity markets. Energy. 2023;264:126270.

[209] Tang H, Wang S. A model-based predictive dispatch strategy for unlocking and optimizing the building energy flexibilities of multiple resources in electricity markets of multiple services. Applied Energy. 2022;305:117889.

[210] Luerssen C, Gandhi O, Reindl T, Sekhar C, Cheong D. Life cycle cost analysis (LCCA) of PV-powered cooling systems with thermal energy and battery storage for off-grid applications. Applied energy. 2020;273:115145.

[211] Marczinkowski HM, Østergaard PA. Evaluation of electricity storage versus thermal storage as part of two different energy planning approaches for the islands Samsø and Orkney. Energy. 2019;175:505-14.

[212] Li F, Sun B, Zhang C, Liu C. A hybrid optimization-based scheduling strategy for combined cooling, heating, and power system with thermal energy storage. Energy. 2019;188:115948.

[213] Mauler L, Duffner F, Zeier WG, Leker J. Battery cost forecasting: a review of methods and results with an outlook to 2050. Energy & Environmental Science. 2021;14(9):4712-39.

[214] He G, Chen Q, Kang C, Pinson P, Xia Q. Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life. IEEE Transactions on Smart Grid. 2015;7(5):2359-67.

[215] Kumbaroğlu G, Madlener R. Evaluation of economically optimal retrofit investment options for energy savings in buildings. Energy and Buildings. 2012;49:327-34.

[216] Di L, Junwei C, Mingshuang L. Collaborative optimization strategy of information and energy for distributed data centers. Journal of Tsinghua University (Science and Technology). 2022;62(12):1864-74.

[217] Ming Z, Lilin P, Qiannan F, Yingjie Z. Trans-regional electricity transmission in China: Status, issues and strategies. Renewable and Sustainable Energy Reviews. 2016;66:572-83.

[218] https://chinaenergyportal.org/en/2021-electricity-other-energy-statistics-preliminary/.

[219] https://www.ndrc.gov.cn/xwdt/ztzl/dsxs/.

[220] Zhang Y, Li H, Wang S. The global energy impact of raising the space temperature for high-temperature data centers. Cell Reports Physical Science. 2023;4(10).

[221] Lehpamer H. Microwave transmission networks: planning, design, and deployment: McGraw-Hill Education, 2010.

[222] Korotky SK. Price-points for components of multi-core fiber communication systems in backbone optical networks. Journal of Optical Communications and Networking. 2012;4(5):426-35.

[223] Bird S, Achuthan A, Maatallah OA, Hu W, Janoyan K, Kwasinski A, et al. Distributed (green) data centers: A new concept for energy, computing, and telecommunications. Energy for Sustainable Development. 2014;19:83-91.

[224] Fox-Penner PS, Chubka M, Earle RL. Transforming America's Power Industry: The Investment Challenge. Report Forthcoming for The Edison Electric Foundation. 2008.

[225] Li W, Zhou X, Li K, Qi H, Guo D. TrafficShaper: Shaping inter-datacenter traffic to reduce the transmission cost. IEEE/ACM Transactions on Networking. 2018;26(3):1193-206.

[226] Barroso LA, Hölzle U, Ranganathan P. The datacenter as a computer: Designing warehouse-scale machines: Springer Nature, 2019.

[227] Li W, Yang M, Long R, He Z, Zhang L, Chen F. Assessment of greenhouse gasses and air pollutant emissions embodied in cross-province electricity trade in China. Resources, Conservation and Recycling. 2021;171:105623.

[228] Ficher M, Berthoud F, Ligozat A-L, Sigonneau P, Wisslé M, Tebbani B. Assessing the carbon footprint of the data transmission on a backbone network. Conference Assessing the carbon footprint of the data transmission on a backbone network. IEEE, p. 105-9.

[229] Malmodin J. The power consumption of mobile and fixed network data services-The case of streaming video and downloading large files. Conference The power consumption of mobile and fixed network data services-The case of streaming video and downloading large files, vol. 2020.

[230] Wu A, Paul Ryan, and Terence Smith. Intelligent Efficiency for Data Centres and Wide Area Networks. International Energy Agency; 2019.

[231] Taylor C, Koomey J. Estimating energy use and greenhouse gas emissions of internet advertising. Network. 2008.

[232] Aslan J, Mayers K, Koomey JG, France C. Electricity Intensity of Internet Data Transmission: Untangling the Estimates. Journal of Industrial Ecology. 2018;22(4):785-98.

[233] Schien D, Preist C, Yearworth M, Shabajee P. Impact of location on the energy footprint of digital media. Conference Impact of location on the energy footprint of digital media. IEEE, p. 1-6.

[234] Coroama V. Investigating the Inconsistencies among Energy and Energy Intensity Estimates of the Internet. Metrics and Harmonising Values. Bern, Switzerland, Tech Rep. 2021.

[235] Schien D, Shabajee P, Yearworth M, Preist C. Modeling and assessing variability in energy consumption during the use stage of online multimedia services. Journal of Industrial Ecology. 2013;17(6):800-13.

[236] Malmodin J, Lundén D, Nilsson M, Andersson G. LCA of data transmission and IP core networks. Conference LCA of data transmission and IP core networks. p. 1-6.

[237] Schien D, Preist C. Approaches to energy intensity of the internet. IEEE Communications Magazine. 2014;52(11):130-7.

[238] Krug L, Shackleton M, Saffre F. Understanding the environmental costs of fixed line networking. Conference Understanding the environmental costs of fixed line networking. p. 87-95.

[239] Schien D, Coroama VC, Hilty LM, Preist C. The energy intensity of the Internet: edge and core networks. Conference The energy intensity of the Internet: edge and core networks. Springer, p. 157-70.

[240] Malmodin J, Lundén D. The energy and carbon footprint of the ICT and E&M sector in Sweden 1990-2015 and beyond. Conference The energy and carbon footprint of the ICT and E&M sector in Sweden 1990-2015 and beyond. Atlantis Press, p. 209-18.

[241] Van Heddeghem W, Lannoo B, Colle D, Pickavet M, Demeester P. A quantitative survey of the power saving potential in IP-over-WDM backbone networks. IEEE Communications Surveys & Tutorials. 2014;18(1):706-31.

[242] Van Heddeghem W, Idzikowski F, Vereecken W, Colle D, Pickavet M, Demeester P. Power consumption modeling in optical multilayer networks. Photonic Network Communications. 2012;24:86-102.

[243] Elmirghani J, Klein T, Hinton K, Nonde L, Lawey A, El-Gorashi T, et al. GreenTouch GreenMeter core network energy-efficiency improvement measures and optimization. Journal of Optical Communications and Networking. 2018;10(2):A250-A69.

[244] Meng L, Xin N, Hu C, Sabea HA, Zhang M, Jiang H, et al. Dual-gated single-molecule field-effect transistors beyond Moore’s law. Nature Communications. 2022;13(1):1410.

[245] Lu X, Kong F, Liu X, Yin J, Xiang Q, Yu H. Bulk savings for bulk transfers: Minimizing the energy-cost for geo-distributed data centers. IEEE Transactions on Cloud Computing. 2017;8(1):73-85.

[246] https://pmos.sd.sgcc.com.cn/px-settlement-infpubmeex/fileService/preview?fileId=nac1433d2a1ab4f8b9e3f63317dc70a95#toolbar=0.

[247] http://zfxxgk.nea.gov.cn/auto92/201503/t20150330_1896.htm.

[248] https://www.ndrc.gov.cn/xwdt/ztzl/gbmjcbzc/gjfzggw/201807/t20180704_1209052.html.

[249] Hintemann R, Hinterholzer S. Energy consumption of data centers worldwide. Conference Energy consumption of data centers worldwide.

[250] https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2018/m02/global-cloud-index-projects-cloud-traffic-to-represent-95-percent-of-total-data-center-traffic-by-2021.html.

[251] http://www.cac.gov.cn/2022-08/02/c_1661066515613920.htm.

[252] Ullrich N, Piontek FM, Herrmann C, Saraev A, Viere T. Estimating the resource intensity of the Internet: A meta-model to account for cloud-based services in LCA. Procedia CIRP. 2022;105:80-5.

[253] Zhang Y, Liu T, Yao L, Song Q, Gao C. Negligible carbon costs of UHVDC infrastructure delivering renewable electricity. Resources, Conservation and Recycling. 2023;192:106940.

[254] https://www.ndrc.gov.cn/xwdt/ztzl/dsxs/gzdt5/202206/t20220623_1327721.html.

[255] Yang F, Chien AA. Large-scale and extreme-scale computing with stranded green power: Opportunities and costs. IEEE Transactions on Parallel and Distributed Systems. 2017;29(5):1103-16.

[256] https://pdf.dfcfw.com/pdf/H3_AP202301151582026310_1.pdf?1673854741000.pdf.

[257] https://www.hljzx.gov.cn/tagz/tajblxd/2022011927296.htm.

[258]http://cdst.chengdu.gov.cn/cdkxjsj/c108731/2022-05/05/content_0e7adf8555b34137be322b868af10ef6.shtml.

[259] https://t.qianzhan.com/caijing/detail/230627-50f78ce2.html.

[260] The decarbonization path of China's digital infrastructure - data centers and the carbon reduction potential and challenges of 5G (2020-2035). 2021.

[261] Preist C, Schien D, Shabajee P. Evaluating sustainable interaction design of digital services: The case of YouTube. Conference Evaluating sustainable interaction design of digital services: The case of YouTube. p. 1-12.

[262] Tian C, Huang G, Xie Y. Systematic evaluation for hydropower exploitation rationality in hydro-dominant area: A case study of Sichuan Province, China. Renewable Energy. 2021;168:1096-111.

[263] Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys (Csur). 2018;51(4):1-35.