Energy Engineering

Tech Science Press

Doi:10.32604/ee.2025.072631

ARTICLE

Multi-Time Scale Optimization Scheduling of Data Center Considering Workload Shift and Refrigeration Regulation

Luyao Liu*, Xiao Liao, Yiqian Li and Shaofeng Zhang

China Energy Engineering Group Guangdong Electric Power Design Institute Co., Ltd., Guangzhou, 510700, China

*Corresponding Author: Luyao Liu. Email: liuluyao@gedi.com.cn

Received: 31 August 2025; Accepted: 06 November 2025; Published: 27 January 2026

ABSTRACT: Data center industries have been facing huge energy challenges due to escalating power consumption and associated carbon emissions. In the context of carbon neutrality, the integration of data centers with renewable energy has become a prevailing trend. To advance the renewable energy integration in data centers, it is imperative to thoroughly explore the data centers' operational flexibility. Computing workloads and refrigeration systems are recognized as two promising flexible resources for power regulation within data center micro-grids. This paper identifies and categorizes delay-tolerant computing workloads into three types (long-running non-interruptible, long-running interruptible, and short-running) and develops mathematical time-shifting models for each. Additionally, this paper examines the thermal dynamics of the computer room and derives a time-varying temperature model coupled to refrigeration power. Building on these models, this paper proposes a two-stage, multi-time scale optimization scheduling framework that jointly coordinates computing workloads time-shift in day-ahead scheduling and refrigeration power control in intra-day dispatch to mitigate renewable variability. A case study demonstrates that the framework effectively enhances the renewable-energy utilization, improves the operational economy of the data center microgrid, and mitigates the impact of renewable power uncertainty. The results highlight the potential of coordinated computing workloads and thermal system flexibility to support greener, more cost-effective data center operation.

KEYWORDS: Data center; renewable energy; load shift; multi-time scale optimization

1 Background and Motivation

Artificial Intelligence (AI) has experienced explosive growth in recent years and plays a pivotal role in improving productivity and driving economic development. Under this background, data has increasingly emerged as a primary factor of production, and society is rapidly transitioning into the digital economy era that stimulates unprecedented demand for computational capacity. Data centers, functioning as the backbone for data processing and storage, are thus emerging as vital socio-economic infrastructure during this transition [1].

The data center industries are facing huge energy challenges due to their high power consumption and carbon emissions. Worldwide, data centers consume approximately 460 TWh of electricity annually, comprising nearly 2% of the global electricity use [2]. With world's over 8000 data centers, the United States, Europe, and China host a significant portion of these facilities, which are 33%, 16% and 10%, respectively. In 2022, US data centers consumed 200 TWh of electricity, accounting for about 4% of the national electricity consumption. China's data center electricity consumption reached 216.6 TWh, roughly 2.6% of the societal electricity consumption in 2021. The CO^2 emissions from data centers reached 135 million tons,

approximately 1.14% of the national emissions [3]. Emerging AI models with billions of parameters demand enormous energy during their training, and as AI technology evolves, data centers' power consumption and carbon emissions will continue to grow [4]. In the pursuit of carbon neutrality, the integration of data centers with renewable energy sources has emerged as an inevitable trend [5]. China's 'Channeling computing resources from the east to the west' policy was introduced against this backdrop, aiming to optimize the data-center layout and promote green energy utilization by leveraging the abundant wind and solar resources of the western regions [6].

To effectively integrate the intermittent renewable energy, it is essential to extensively analyze the operation flexibility of data centers [7]. Fig. 1 illustrates the structural overview of a green data center micro-grid. The information technology (IT) equipment mainly includes the servers that provide computing resource for workloads. The refrigeration system provides cooling resources to extract the heat generated from the IT equipment. The power infrastructure connects the wind, photovoltaic (PV) power system and main-grid in supply side and IT equipment, refrigeration system in power demand side.

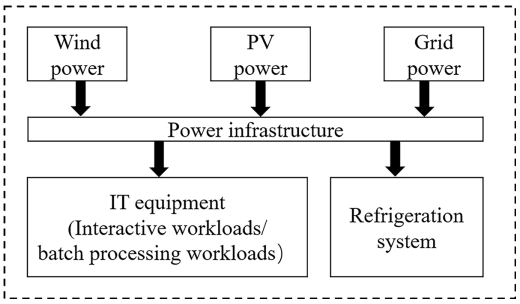

Figure 1: Structure of green data center

Data centers manage an exponentially increasing volume of computational workloads driven by continuous enhancements in processing capabilities. As computing workloads migrate across the network in temporal and spatial domains, the associated power consumption of the data center correspondingly varies. Thus, computing workloads exhibit significant spatiotemporal adjustable characteristics [8]. Additionally, computer room temperature can fluctuate within predefined limits, and the temperature change exhibits a temporal delay relative to the cooling power change. By dynamically modulating the refrigeration power within acceptable temperature bounds, fluctuations in renewable energy generation can be effectively managed [9]. Consequently, the refrigeration system functions as an auxiliary flexible resource for data center micro-grid operations.

This study concentrates on formulating power regulation models for delay-tolerant computing workloads and refrigeration system within data centers, and incorporating these flexible regulation models into power scheduling framework of data centers to increase the operational economy and maximize the utilization of renewable energy. Section 2 reviews pertinent literature to delineate the existing state of research and to identify critical knowledge gaps.

2 Literature Review

In recent studies, leveraging the spatiotemporal flexibility of computing workloads in data centers to promote renewable energy utilization has emerged as a new research focus. Chen et al. [10] analyzed the topology of computing power networks from an energy perspective, and used price leverage to guide the spatial shift of workloads from interconnected data centers to promote the renewable energy utilization and improve the grid resilience. Yang et al. [11] proposed a spatial migration mechanism of computing workloads

based on the spatiotemporal distribution complementarity of renewable energy in multiple regions around the world, and conducted simulations using Google's interconnected multiple data centers to minimize carbon emissions. However, the aforementioned research focuses mainly on the spatial redistribution of computing workloads across multiple data centers, with limited attention to the temporal shift of workloads within individual data centers.

The flexibility of computing workloads within an individual data center primarily manifests in their temporal transferability. According to the length of response delay, workloads can be categorized into delay-sensitive and delay-tolerant types [12]. Delay-sensitive workloads primarily refer to 'online tasks', or 'interactive workloads' that provide real-time feedback to users. Delay-tolerant workloads, by contrast, refer to 'offline tasks' or 'batch processing workloads' that execute a series of jobs based on a computer program without human intervention; Their maximum response time can range from several minutes to days [13]. By dynamically adjusting the delay-tolerant workloads' start time and the amount of computation load for each time interval within the response deadline, the power consumption of workloads can be controlled. This study concentrates on the temporal flexibility of workloads within an individual data center.

Studies regarding the temporal flexibility modeling of delay-tolerant workloads within an individual data center have been conducted. Liu et al. [14] proposed a power consumption model for the delay-tolerant workloads and integrated it with a renewable power supply model. The optimal plan of generator units' outputs and workloads scheduling was obtained with the objective of minimizing operation cost of data centers. Cupelli et al. [15] simulated the power consumption of delay-tolerant workloads considering their arrival, queuing, and execution process. By coordinating workloads scheduling with cooling systems and energy-storage devices, they demonstrated reductions in overall data-center energy costs. Kwon [16] analyzed the time-shift characteristics of a simple delay-tolerant workload and developed a flexible mechanism model to shift workloads power consumption towards periods of high renewable generation. The studies above highlight the potential of delay-tolerant workloads to provide flexible regulation in individual data centers.

However, existing studies have primarily focused on modeling single, simple time-shiftable workload and haven't comprehensively accounted for multiple time-shiftable workload types with distinct operational characteristics and shifting behaviors. Comprehensive modeling of multiple delay-tolerant workloads, and systematic integration of these models into data center power dispatch frameworks, remain insufficiently explored.

Furthermore, the refrigeration system serves as an additional flexible resource in data center operations. Data centers commonly employ refrigeration systems to maintain suitable thermal conditions. With advancements in technology, the temperature requirements for server clusters have become less stringent. The ASHRAE introduced the 'Data Center Environmental Thermal Guidelines' in 2014, extending the design ranges for maximum allowable computer room temperatures to 32°C, 35°C, 40°C, and 45°C across A1–A4 classifications. Considering the temperature fluctuation range and thermal inertia, existing research investigated the feasibility of utilizing refrigeration modulation to optimize data center operation. Tang et al. [17] established the first-order equivalent thermal parameter model to analyze the thermal load dynamics of the air condition system in base stations. By incorporating temperature variation boundaries, they employed day-ahead demand response strategies utilizing start-stop control to manage cooling and heating loads. Zhu et al. [18] formulated a temperature response model considering thermal inertia within data center environments and proposed dynamic temperature regulation strategies to optimize real-time renewable energy utilization. These investigations demonstrate the potential of using temperature and refrigeration modulation to enhance power scheduling efficiency in data centers. However, they don't address the coordination of computing workloads time-shift and refrigeration modulation in data center operational management.

In light of the deficiencies, this research aims to provide a power scheduling framework for data center by incorporating the computing workloads shift and refrigeration regulation. To this aim, this work first analyzes the time-shift characteristics of three kinds of typical delay-tolerant computing workloads and establishes the corresponding time-shift models. Then, this work quantitatively examines the heat transfer process of data center and builds the time-variant model of computer room temperature. Last, this work proposes a two-stage multi-time scale optimization scheduling framework for the green data center, where the time-shift models of workloads are incorporated into the day-ahead optimization scheduling, and the refrigeration power is set as the control variable in the intraday scheduling. The proposed multi-time scale energy management method is applied to a data center to verify its effectiveness.

3 Methods

3.1 Time-Shift Models of the Delay-Tolerant Workloads

This section introduces three kinds of delay-tolerant workloads, i.e., Long-running non-interruptible, long-running interruptible, and short-running workloads, and models the time-shifting processes of the workloads.

3.1.1 Workloads Classification and Time-Shift Characteristics Analysis

Time-shiftable workloads are categorized into long-running and short-running workloads, distinguished by their processing time length. Long-running workloads include continuous non-interruptible and interruptible workloads based on their interruptibility.

Analysis of large internet data centers shows that processor resource consumption of computing workloads follows a heavy-tailed distribution, indicating that a small portion of long-running workloads consume a significant amount of processor resources [19]. These energy-intensive long-running workloads present opportunities for power regulation. Additionally, analysis reveals that most workloads in Google's cluster last only a few minutes [19], while over 90% of batch jobs in Alibaba's data cluster run for less than 15 min [20]. Short-running computing tasks, when combined, can effectively help mitigate fluctuations in renewable energy.

- Long-running non-interruptible workload

Some workloads, e.g., continuous integration and continuous deployment jobs (CI/CD), test suites, compilation jobs, database migration, and backups, can't be paused or interrupted once execution begins and must run continuously until completion. Workload with this working characteristic are identified as 'long-running non-interruptible workload'.

The temporal mitigation pattern for this kind of workload is to shift the overall workload within a certain time range while keeping the original shape of the computing power load curve unchanged.

- Long-running interruptible workload

Certain workloads, e.g., machine learning training, block-chain mining, protein folding, or long-running scientific simulations, can be paused or interrupted during execution and resumed at an appropriate time. Workload with this working characteristic is regarded as 'long-running interruptible workload'.

According to its time-shift characteristics, long-running interruptible workload can be viewed as a holistic task comprised of multiple sequential subtasks. The shift strategy is to adjust each subtask forward or backward to a certain time point within a specified time frame, while ensuing the sequential order dependency among the subtasks.

- Short-running workloads

Some workloads, including function as a service, scheduled data backups, log cleaning, report generation, email or notification sending, etc., typically require brief execution time and don't demand immediate response. These workloads can tolerate delays ranging from minutes to several hours, classifying them as short-running workloads.

Short-running workloads exhibit two defining characteristics: short execution duration and high task volume. When numerous short-running workloads aggregate, they form a short-running workload cluster, which presents significant potential for power regulation. The time-shifting mechanism for managing these workloads involves rescheduling all or part of the short-running tasks from their original time slots to alternative periods in future time window for execution.

3.1.2 Modelling of Long-Running Non-Interruptible Workload

Set the day-ahead scheduling period to 24 h, and define the scheduling time step as Δt and the number of daily scheduling periods as T. Introduce F_a^{ori} to denote the original computing power demand time series of the long-running non-interruptible workload. F_a^{ori} is a constant vector with dimensions of 1^*T, represented as follows:

$$F_a^{ori} = \left[x_a^1, x_a^2, x_a^3, \ldots x_a^t, \ldots, x_a^{T-1}, x_a^T\right] \tag{1}$$

where x_a^1, x_a^2, x_a^t are the computing power demand values of the long-running non-interruptible workload at the time point of 1, 2, and t. In vector F_a^{ori}, there exists a continuous period of computing power demand values, and the elements before and after these continuous periods are zero. Thus, the vector F_a^{ori} can be more specifically written as:

$$F_a^{ori} = \left[0, 0, \ldots, f_a^1, f_a^2, \ldots f_a^k, \ldots, f_a^K, \ldots, 0, 0\right] \tag{2}$$

where f_a^1 represents the computing power demand value at the starting node of the long-running continuous workload; f_a^2, f_a^k, f_a^K, is the computing power demand value at the 2nd, k^{th}, K^{th} time points following the starting node, respectively. K represents the processing time of the long-running non-interruptible workload execution, which is a constant.

The mitigation pattern for the long-running non-interruptible workload is to shift the entire workload along the time axis. Once the start time of the workload is determined, the computing power demand for each subsequent time point can be uniquely ascertained. Therefore, the pivotal parameter for the time-shift modelling of long-running non-interruptible workload is the time-shift vector of start node.

Use U_a to represent the time-shift vector of start node of the long-running non-interruptible workload. U_a is the variable with a dimension of 1^*T, represented as Eq. (3):

$$U_a = \left[u_a^1, u_a^2, u_a^3 \ldots, u_a^t, \ldots, u_a^{T-1}, u_a^T\right] \tag{3}$$

where the element u_a^t represents whether the start node of the long-running non-interruptible workload is located at a certain time point. When u_a^t is 1/0, it signifies that the start node of the workload is/isn't positioned within the t time point.

For U_a, only one element in the vector can have a value of 1, thus the sum of all elements is constrained to 1, denoted by:

$$\sum_{t=1}^{T} u_a^t = 1 \tag{4}$$

The time at which the start node is positioned is denoted as U_a^{index}, which is expressed as Eq. (5):

$$U_a^{index} = [1, 2, 3, \ldots, T]^{transpose} * U_a \tag{5}$$

The long-running non-interruptible workload has the earliest start time point t_e and the latest start time point t_l, allowing the start node of the workload to shift within $[t_e, t_l]$. The constraint is expressed as:

$$t_e \le U_a^{index} \le t_l \tag{6}$$

The relation between the computing power demand time series of workload after shifting F_a^{mov} and the time-shift vector of the workload start node U_a is expressed as Eq. (7):

$$F_a^{mov} = U_a * M_a^{mat} \tag{7}$$

where the dimension of F_a^{mov} is 1^*T; M_a^{mat} is a constructed auxiliary constant matrix, with a dimension of T^*T, as shown in Eq. (8).

$$M_a^{mat} = \begin{bmatrix} m_a^{1,1} & m_a^{1,2} & m_a^{1,3} & \ldots & m_a^{1,t} & \ldots & m_a^{1,T-1} & m_a^{1,T} \\ m_a^{2,1} & m_a^{2,2} & m_a^{2,3} & \ldots & m_a^{2,t} & \ldots & m_a^{2,T-1} & m_a^{2,T} \\ m_a^{3,1} & m_a^{3,2} & m_a^{3,3} & \ldots & m_a^{3,t} & \ldots & m_a^{3,T-1} & m_a^{3,T} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ m_a^{t,1} & m_a^{t,2} & m_a^{t,3} & \ldots & m_a^{t,t} & \ldots & m^{t,T-1} & m_a^{t,T} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ m_a^{T-1,1} & m_a^{T-1,2} & m_a^{T-1,3} & \ldots & m_a^{T-1,t} & \ldots & m_a^{T-1,T-1} & m_a^{T-1,T} \\ m_a^{T,1} & m_a^{T,2} & m_a^{T,3} & \ldots & m_a^{T,t} & \ldots & m_a^{T,T-1} & m_a^{T,T} \end{bmatrix} \tag{8}$$

M_a^{mat} can be more specifically written as Eq. (9). In this matrix, elements from 'column k, row k' to 'row k, column $k+K-1$' ($k = 1, 2, \ldots, T+1-K$) are filled with the computing power demand values corresponding to the start node and subsequent nodes of the long-running uninterruptible workload, and the elements at the remaining positions are 0.

$$M_a^{mat} = \begin{bmatrix} f_a^1 & f_a^2 & \ldots & f_a^K & 0 & \ldots & 0 & 0 \\ 0 & f_a^1 & \ldots & f_a^{K-1} & f_a^K & \ldots & 0 & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & \ldots & 0 & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ldots & \ldots & \ldots & 0 & 0 \\ 0 & 0 & \ldots & f_a^1 & f_a^2 & \ldots & f_a^K & 0 \\ 0 & 0 & \ldots & 0 & f_a^1 & \ldots & f_a^{K-1} & f_a^K \end{bmatrix} \tag{9}$$

3.1.3 Modelling of Long-Running Interruptible Workload

Introduce F_b^{ori} to denote the original time series of computing power demand of long-running interruptible workload. F_b^{ori} is a constant vector with a dimension of 1^*T, represented as follows:

$$F_b^{ori} = \left[y_b^1, y_b^2, y_b^3, \ldots y_b^t, \ldots y_b^{T-1}, y_b^T\right] \tag{10}$$

where y_a^1, y_a^2, y_a^t are the computing power demand values of the long-running interruptible workload at the time point of 1, 2, and t.

According to the time-shift characteristics of long-running interruptible workload, this study considers the long-running interruptible workload as a holistic task composed of multiple subtasks in sequence. The duration of each subtask is consistent with the scheduling time step Δt. Record the start node of the long-running interruptible workload as the 1st subtask, and the subtasks in subsequent time points as the 2nd, …, k^{th} and K^{th} subtask in sequence, and the corresponding computing demand value as f_b^1, f_b^2, …, f_b^k, …, f_b^K. K represents the processing time of long-running interruptible workload execution, which is a constant. The constant vector F_b^{mat} can be constructed from f_b^1 to f_b^K, as illustrated below:

$$F_b^{mat} = \left[f_b^1, f_b^2, \ldots f_b^k, \ldots, f_b^K\right] \tag{11}$$

Hence, the original computing demand time series of the long time interruptible workload F_b^{ori} can be more specifically written as:

$$F_b^{ori} = \left[0, 0, \ldots, f_b^1, f_b^2, \ldots, f_b^k, \ldots, f_b^K, \ldots, 0, 0\right] \tag{12}$$

In this vector, there exists a period of time with computing power demand value, and the elements before and after this period are zero.

The mitigation pattern for the long-running interruptible workload is scheduling individual subtasks to new time points within a specified time frame while preserving their sequential dependencies. Once the execution time of each subtask is determined, the computing power demand for the workload can be uniquely ascertained. Thus, the key parameter for time-shift modelling of long-running interruptible workload is the time-shift matrix of its subtasks.

Use $u_b^{t,k}$ to indicate whether the k^{th} subtask is moved to time point t, with a value of 0 or 1. The time shift matrix variable U_b with a dimension of T^*K is formed, which is represented as:

$$U_b = \begin{bmatrix} u_b^{1,1} & u_b^{1,2} & \cdots & u_b^{1,k} & \cdots & u_b^{1,K} \\ u_b^{2,1} & u_b^{2,2} & \cdots & u_b^{2,k} & \cdots & u_b^{2,K} \\ u_b^{3,1} & u_b^{3,2} & \cdots & u_b^{3,k} & \cdots & u_b^{3,K} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ u_b^{t,1} & u_b^{t,2} & \cdots & u_b^{t,k} & \cdots & u_b^{t,K} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ u_b^{T-1,1} & u_b^{T-1,2} & \cdots & u_b^{T-1,k} & \cdots & u_b^{T-1,K} \\ u_b^{T,1} & u_b^{T,2} & \cdots & u_b^{T,k} & \cdots & u_b^{T,K} \end{bmatrix} \tag{13}$$

For each column in matrix U_b, the sum of the element values in each row is 1, represented by Eq. (14), implying that a subtask can only transfer from one time point to another, not to multiple time points.

$$\sum_{t=1}^{T} u_b^{t,k} = 1\,(k = 1, 2, \ldots, K) \tag{14}$$

The matrix of time points where subtasks locate after shifting is denoted as U_b^{index}, which is a variable with a dimension of 1*K, expressed as Eq. (15). U_b^{index} is calculated using Eq. (16). $u_b^{index,k}$ represents the time point where each subtask locates after shifting. Eq. (17) enforces that the time position at which the $(k+1)$-th subtask moves to should be greater than the time position at which the k^{th} subtask moves to.

$$U_b^{index} = \left[u_b^{index,1}, u_b^{index,2}, \ldots, u_b^{index,k}, \ldots, u_b^{index,K}\right] \tag{15}$$

$$U_b^{index} = [1, 2, 3, \ldots, T] * U_b \tag{16}$$

$$u_b^{index,k+1} - u_b^{index,k} \geq 1 \tag{17}$$

The computing power demand time series of each subtask after shifting is denoted as F_b^{mov}, which is a variable with a dimension of 1*T, expressed as below:

$$F_b^{mov} = \left[U_b * \left(F_b^{mat}\right)^{transpose}\right]^{transpose} \tag{18}$$

3.1.4 Modelling of Short-Running Workloads

Introduce F_c^{ori} to denote the original computing power demand time series of short-running workloads. F_c^{ori} is a constant vector with a dimensions of 1*T, represented as Eq. (19). Use F_c^{mov} to denote the computing power demand time sequence of the short-running workloads after shift. F_c^{ori} is a variable with a dimension of 1*T, represented Eq. (20):

$$F_c^{ori} = \left[f_c^1, f_c^2, \ldots f_c^k, \ldots, f_c^T\right] \tag{19}$$

$$F_c^{mov} = \left[f_{c*}^1, f_{c*}^2, \ldots f_{c*}^t, \ldots, f_{c*}^T\right] \tag{20}$$

where f_c^1, f_c^2, f_c^k are the computing power demand values of short-running workloads at original status at time point 1, 2, and k; f_{c*}^1, f_{c*}^2, f_{c*}^t are the computing power demand values of short-running workloads after shifting at time point 1, 2, and t.

The time shifting pattern for short-running workloads is shifting all or part of the tasks from the original time periods to new time periods. The key parameter for the time shift modelling of short-running workloads is the time shift matrix of the inflow and outflow amount of computing power demand at each time point.

Use $u_c^{t,k}$ to represent the amount of computing power demand of short-running workloads that is transferred from the original time point k to another time point of t. The time shift matrix variable of the inflow and outflow amount of computing power demand U_c with a dimension of T*T is formed, which is expressed as:

$$u_c = \begin{bmatrix} u_c^{1,1} & u_c^{1,2} & \ldots & u_c^{1,k} & \ldots & u_c^{1,T-1} & u_c^{1,T} \\ u_c^{2,1} & u_c^{2,2} & \ldots & u_c^{2,k} & \ldots & u_c^{2,T-1} & u_c^{2,T} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ u_c^{t,1} & u_c^{t,2} & \ldots & u_c^{t,k} & \ldots & u_c^{t,T-1} & u_c^{t,T} \\ u_c^{t+1,1} & u_c^{t+1,2} & \ldots & u_c^{t+1,k} & \ldots & u_c^{t+1,T-1} & u_c^{t+1,T} \\ u_c^{t+2,1} & u_c^{t+2,2} & \ldots & u_c^{t+2,k} & \ldots & u_c^{t+2,T-1} & u_c^{t+2,T} \\ u_c^{t+3,1} & u_c^{t+3,2} & \ldots & u_c^{t+3,k} & \ldots & u_c^{t+3,T-1} & u_c^{t+3,T} \\ u_c^{t+4,1} & u_c^{t+4,2} & \ldots & u_c^{t+4,k} & \ldots & u_c^{t+4,T-1} & u_c^{t+4,T} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ u_c^{T-1,1} & u_c^{T-1,2} & \ldots & u_c^{T-1,k} & \ldots & u_c^{T-1,T-1} & u_c^{T-1,T} \\ u_c^{T,1} & u_c^{T,2} & \ldots & u_c^{T,k} & \ldots & u_c^{T,T-1} & u_c^{T,T} \end{bmatrix} \tag{21}$$

U_c should meet the following constraint:

$$0 \le U_c \le C_c^{mat} * M \tag{22}$$

In this formula, C_c^{mat} is a constructed constant matrix with the same dimension as the variable U_c. C_c^{mat} is expressed by:

$$C_c^{\mathrm{mat}} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 1 & 1 \\ 1 & 1 & \cdots & \cdots & \cdots & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & \cdots & \cdots & 1 & 1 \\ 1 & 1 & \cdots & \cdots & \cdots & 0 & 1 \\ 1 & 1 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & 1 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & \cdots & 1 & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & 1 & 1 \end{bmatrix} \tag{23}$$

To illustrate the specific form of C_c^{mat}, we first introduce a constant value H to represent the maximum deferrable time periods of short-running workloads. The short-running workloads of the original time points are allowed to be fully or partially transferred to the *1*st to H^{th} time point thereafter. In C_c^{mat}, when k = 1, 2, …, $T - H$, the elements from 'column k, row k,' to 'column k, row $k + H$', and when $k = T - H + 1$, $T - H + 2$, …, T, the elements from 'column k, row k' to 'column k, row T' and from 'column k, row 1' to 'column k, row $H - T + k$', are filled with the value of 1, and the elements at the remaining positions are 0. Set the maximum deferrable time periods H to be 11, the C_c^{mat} can be specifically written as Eq. (23). The vector C_c^{mat} can be changed as the parameter H varies.

M is a constant, equal to the peak computing power demand f_c^{max} of the short-running workloads. Eq. (22) indicates that the transferred computing power demand from the original time point to the new time period should be less than f_c^{max} and greater than 0.

Another constraint for U_c is that, for the k^{th} column, the sum of all the elements equals to the original computing power demand f_c^k at time point k. For the t^{th} column, the accumulated elements of all the rows equals to the computing power demand at time point t of workloads after shifting f_{c*}^t.

$$f_c^k = \sum_{t=1}^{T} u_c^{t,k} (t = 1, 2, \ldots, T) \tag{24}$$

$$f_{c*}^t = \sum_{k=1}^{T} u_c^{t,k} (k = 1, 2, \ldots, T) \tag{25}$$

To prevent invalid shifting, a penalty cost coefficient matrix D_c^{mat} with dimension of T*T is established, as expressed in Eq. (26). Setting the unit penalty cost coefficient for the transfer amount of computing power demand at all possible time points to be 0.05 ¥/kWh, when k = 1, 2, …, $T - H$, the elements in positions of D_c^{mat} from the 'column k, row $k + 1$' to 'column k, row $k + H$', and when $k = T - H + 1, T - H + 2, \ldots, T - 1$, elements from the 'column k, row $k + 1$' to 'column k, row T' and from 'column k, row 1,' to 'column k, row $H - (T - k)$', and when $k = T$, elements from the 'column k, row 1' to 'column k, row H' are filled with 0.05, and

the other elements are set to be 0. One can adjust the matrix D_c^{mat} according to concrete parameter value of the unit penalty cost coefficient.

$$D_c^{mat} = \begin{bmatrix} 0 & 0 & \dots & \dots & \dots & 0.05 & 0.05 \\ 0.05 & 0 & \dots & \dots & \dots & 0.05 & 0.05 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.05 & 0.05 & \dots & \dots & \dots & 0.05 & 0.05 \\ 0.05 & 0.05 & \dots & \dots & \dots & 0 & 0.05 \\ 0.05 & 0.05 & \dots & \dots & \dots & 0 & 0 \\ 0 & 0.05 & \dots & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & 0.05 & 0 \end{bmatrix} \tag{26}$$

The penalty cost function C_{data} is expressed as:

$$C_{data} = \sum_{t=1}^{T} \sum_{k=1}^{T} \frac{U_c. * D_c^{mat}}{F_{cpu}^{max}} * E_{cpu}^{max} * \Delta t \tag{27}$$

where F_{cpu}^{max} represents the maximum computing power of the CPU chip, which is 1600 GFLOPS in this paper, as indicated in Table 1. E_{cpu}^{max} represents the maximum power consumption of the CPU used in this study, which is 150 W, as indicated in Table 1.

3.1.5 Relation of Computing Power and Electricity Power

Computing power refers to a device's capability to perform data processing and generate specific outputs. The computing power is implemented through various computing chips such as CPU, GPU, FPGA, ASIC, which are carried by computers, servers, and other systems. The standard measurement for computing power is the number of FLoating-point Operations executed Per Second (FLOPS).

The computing power of a chip is determined by three key factors: the number of computing cores of the chip N_{core}, the core frequency f_{core}, and the double-precision floating-point operands per clock cycle Z_{core} of the core. The computing power F_{chip}^{max} of the chip follows the formula:

$$F_{chip}^{max} = N_{core} * f_{core} * Z_{core} \tag{28}$$

The computing power capacity of the data center F_{dc}^{max} using CPU chips is expresses as:

$$F_{dc}^{max} = N_{rack} * N_{server} * N_{CPU} * F_{cpu}^{max} \tag{29}$$

where N_{rack} is the number of racks in the datacenter. N_{server} is the number of servers per rack. N_{CPU} is the number of CPU chips per server. F_{cpu}^{max} is the maximum computing power of the CPU chip studied in this paper.

Based on the assumption that: (1) all servers in the data center are homogeneous; (2) all servers in the computer room are turned on. The power consumption of servers can be modelled as Eq. (30) [21,22]:

$$E_{data} = E_{idle} + \left(E_{peak} - E_{idle}\right) * u_{cpu} \tag{30}$$

where u_{cpu} is the processor utilization rate. E_{idle} represents the power of servers when u_{cpu} is zero, E_{peak} represents the power of servers when u_{cpu} reaches 100%.

The formula for calculating CPU utilization is u_{cpu}, expressed as:

$$u_{cpu} = \frac{F_{data}}{F_{dc}^{max}} \tag{31}$$

where F_{data} is the computing power demand of workloads.

This model provides a quantitative approach to estimate the electricity power consumption of workloads based on their computing power demand.

Table 1: Parameters for CPU, server, and rack

Item	Parameter
Number of racks (N_{rack})	60
Type of rack	42U
Type of server	DELL PowerEdge R940xa 4U
Number of servers per rack (N_{server})	10
Total number of servers in the data center	600
Number of CPU per server (N_{CPU})	4
Type of CPU	Intel Gold 6210U [23]
Number of cores per CPU (N_{core})	20
Core frequency (f_{core})	2.5 GHz
Double-precision floating-point operands per clock cycle of the core (Z_{core})	32
Rated power for each CPU (E_{cpu}^{max})	150 W
Rated computing power for each CPU (F_{cpu}^{max})	1600 GFLOPS

3.2 Time-Variant Model of Computer Room Temperature

Data center computer room temperature is influenced by multiple factors including server cluster heat generation, environmental maintenance structure, heat load, cooling power, and outdoor temperature. To simplify the problem-solving process when constructing a thermal model for the data center, the heat transfer through the walls is treated as a one-dimensional problem, and the physical properties of the walls is assumed to be remain constant over time. In addition, the indoor temperature is represented as a single node. The simplified heat transfer process for the data center room is depicted in Fig. 2.

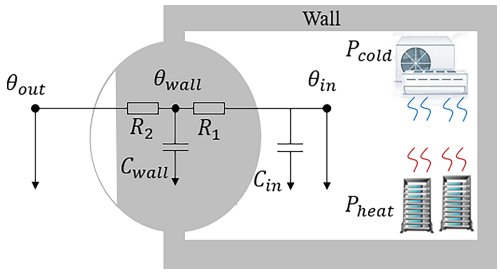

Figure 2: Schematic diagram of the equivalent model of heat transfer process in data center computer room

The energy conservation equation for the indoor air and the walls are constructed as Eqs. (32) and (33).

$$\frac{C_{in} * d\theta_{in}^{t}}{dt} = P_{heat}^{t} - P_{cold}^{t} - \frac{\theta_{in}^{t} - \theta_{wall}^{t}}{R1} \tag{32}$$

$$\frac{C_{wall} * d\theta_{wall}^{t}}{dt} = \frac{\theta_{in}^{t} - \theta_{wall}^{t}}{R1} - \frac{\theta_{wall}^{t} - \theta_{out}^{t}}{R2} \tag{33}$$

In the formula, t represents the current time point; C_{in} is the equivalent heat capacity of indoor air; C_{wall} is the equivalent heat capacity of the wall; $R1$ is the equivalent thermal resistance of indoor air and the inner side of the wall; $R2$ is the equivalent thermal resistance of the outer wall and outdoor air; θ_{in}^{t} is the indoor temperature at time t; θ_{wall}^{t} is the wall temperature at time t; θ_{out}^{t} is the outdoor temperature at time t; P_{heat}^{t} is the heat generation of the servers at time t; P_{cold}^{t} is the refrigeration power at time t.

In order to calculate the indoor temperature changes at different discrete time periods within a day, differential equations of Eqs. (32) and (33) are transformed into differential equations, as expressed as Eqs. (34) and (35), with a time step of $\Delta\tau$.

$$C_{in}\frac{\theta_{in}^{t+\Delta\tau} - \theta_{in}^{t}}{\Delta t} = \left(P_{heat}^{t} - P_{cold}^{t}\right) - \frac{\theta_{in}^{t} - \theta_{wall}^{t}}{R1} \tag{34}$$

$$C_{wall}\frac{\theta_{wall}^{t+\Delta\tau} - \theta_{wall}^{t}}{\Delta t} = \frac{\theta_{in}^{t} - \theta_{wall}^{t}}{R1} - \frac{\theta_{wall}^{t} - \theta_{out}^{t}}{R2} \tag{35}$$

Transforming Eqs. (34) and (35) yields Eqs. (36) and (37):

$$\theta_{in}^{t+\Delta\tau} = \left[\left(P_{heat}^{t} - P_{cold}^{t}\right) - \frac{\theta_{in}^{t} - \theta_{wall}^{t}}{R1}\right] * \frac{\Delta\tau}{C_{in}} + \theta_{in}^{t} \tag{36}$$

$$\theta_{wall}^{t+\Delta\tau} = \left(\frac{\theta_{in}^{t} - \theta_{wall}^{t}}{R1} - \frac{\theta_{wall}^{t} - \theta_{out}^{t}}{R2}\right) * \frac{\Delta\tau}{C_{wall}} + \theta_{wall}^{t} \tag{37}$$

Eqs. (36) and (37) can be used to update the indoor temperature and wall temperature at each time point. Furthermore, according to the practical engineering situation where the variation amplitude of θ_{wall}^{t} is very small compared to that of θ_{in}^{t}, θ_{wall}^{t} can be regarded as its mean value θ_{wall}^{stable}, and Eq. (38) can be derived from Eq. (37).

$$\theta_{wall}^{t} = \frac{R1 * \theta_{out}^{t} + R2 * \theta_{in}^{t}}{R1 + R2} \tag{38}$$

Regarding the determination of the time step $\Delta\tau$, it is crucial to balance accuracy with computational complexity. In this context, $\Delta\tau$ is set as 1 min. Combining Eqs. (36) and (38), the equilibrium constraints of the indoor temperatures at adjacent time points of θ_{in}^{t+1} and θ_{in}^{t} can be expressed as:

$$\theta_{in}^{t+1} = \theta_{in}^{t} + \frac{P_{heat}^{t} - P_{cold}^{t}}{C_{in}} - \frac{\theta_{in}^{t} - \theta_{out}^{t}}{(R1 + R2) * C_{in}} \tag{39}$$

Combined with the initial indoor temperature θ_{in}^{0}, the dynamic indoor temperature throughout the day can be calculated using the time-varying temperature model of computer room. Through the temperature time-variant model, dynamically adjusting the refrigeration power within the temperature range can achieve the regulation of electricity, thereby further promoting the consumption of renewable energy.

4 Two-Stage Multi-Time Scale Optimization Model for Data Center[1]

4.1 Two-Stage Multi-Time Scale Scheduling Framework

4.1.1 Day-Ahead and Intra-Day Optimization Scheduling Diagram

The multi-time scale optimization scheduling includes a two-stage day-ahead and intraday rolling optimization scheduling, as illustrated in Fig. 3. The day-ahead scheduling operates on a 24-h optimization scale with 15 min as the time step. Based on the day-ahead forecasts of renewable power and load, the scheduling plan for next 24 h can be formulated with the goal of minimizing the daily operating cost. The intraday rolling optimization scheduling takes 15 min as a rolling round, 60 min as the optimization scale, and 1 min as the time step for each round of optimization. Each round of optimization scheduling is based on 60-min ahead predictions of renewable power and load, aiming to minimize the operating costs of each round of optimization and minimize the fluctuations of exchange power between day-ahead and real-time plans.

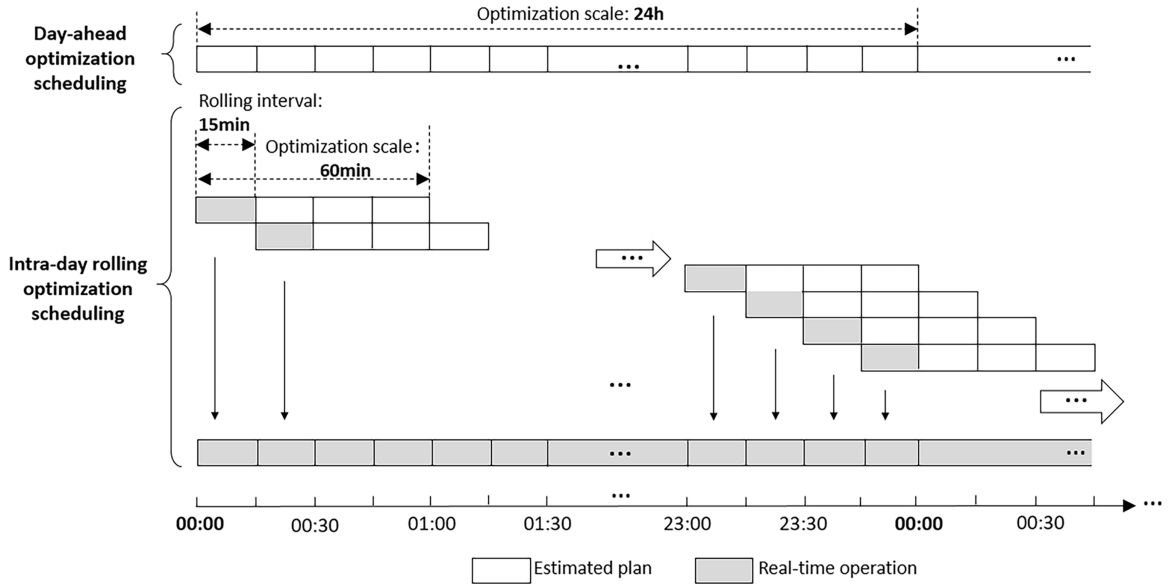

Figure 3: Diagram of the two-stage multi-time scale optimization scheduling framework of data center

4.1.2 Regulation Characteristics at Different Time Scales of Computing Workloads and Refrigeration System

The time step of day-ahead scheduling is generally set as 15 min, which is consistent with sampling interval of power supply and demand. On the one hand, the fluctuations trend of renewable power throughout a day typically exhibit periodic changes. The 15-min interval allows for capturing the primary fluctuation characteristics of wind and PV power, while mitigating the computational burden caused by overly frequent data points. On the other hand, the workloads power consumption is relatively steady throughout the day. Using 15 min as the interval for the measuring of power consumption and as the minimum unit for its movement offers sufficient accuracy to illustrate the load shift mechanism while simplifying data processing. Consequently, the time-shifting model of workloads is suitable to be incorporated into the day-ahead optimization scheduling model. By shifting these workloads along the time axis, the fluctuations of wind and PV power can be offset.

[1]The optimization in this article is without considering cross day load shifting; To isolate the time-shift effect of computing workloads, the energy storage in data center is not considered; The purchase and sale price of electricity in the power grid does not change over time.

However, the volatilities of wind and PV power at the 1-min level are notably faster and more irregular compared to the daily fluctuations forecasted within a 15-min interval. To address this problem, intra-day real-time scheduling are needed. In this study, the time step in the real-time scheduling stage is 1min, which is consistent with the time interval of the time-variant model of the computer room temperature. Given that refrigeration power serves as the only controllable variable for regulating computer room temperature, by embedding the time-varying temperature model into the intra-day optimization scheduling model, adjusting the refrigeration power every 1 min can effectively cope with the instantaneous fluctuations of wind and PV power. Thus, the refrigeration power can be used as a control variable for intra-day optimization scheduling.

4.2 Day-Ahead Scheduling Optimization Objective

The day-ahead optimization scheduling is designed to minimize operational costs within the dispatching day, expressed as:

$$OBJ = \min C_{grid} + C_{cur} + C_{data} \tag{40}$$

The optimization function comprises the net cost of electricity transactions with the main-grid C_{grid} (¥), penalties for wind and PV power curtailment C_{cur}, and penalty costs for the computing workloads scheduling C_{data}(¥), which are expressed as Eqs. (41)–(43).

$$C_{grid} = \sum_{t=1}^{T} K_{buy} E_{buy}^{t} \Delta t - K_{sell} E_{sell}^{t} \Delta t \tag{41}$$

$$C_{cur} = \sum_{t=1}^{T} K_{cur} E_{cur}^{t} \Delta t \tag{42}$$

$$C_{data} = \sum_{t=1}^{T} \sum_{k=1}^{T} \frac{U_{c\cdot} * D_{c}^{mat}}{F_{cpu}^{max}} * E_{cpu}^{max} * \Delta t \tag{43}$$

where K_{buy}/K_{sell} is the unit electricity purchase/sale price from/to the main grid during time period t, which is 0.8 and 0.4 ¥/kWh, respectively[2]; E_{buy}^{t}/E_{sell}^{t} is the day-ahead planned purchased/sold electricity power during time period t; K_{cur} is the unit penalty price for wind and PV power curtailment, which is set as 0.6 ¥/kWh; E_{cur}^{t} is the day-ahead planned renewable power curtailment in time period t. Δt is the duration of time period t, set at 15 min in this paper.

By considering system operation constraints and combining day-ahead forecasts of renewable generation and electricity consumption of fundamental interactive computing workloads, this model can optimize the time shift plan of various delay-tolerant workloads, refrigeration output plan, exchange power plan, wind and PV power utilization plan under the goal of minimizing the operating costs of data center micro-grid during the scheduling day.

[2]The primary objective of this study is to demonstrate how workload shifting alleviates renewable energy curtailment, which is a core technical contribution. To isolate this mechanism, this paper initially adopt a simplified price assumption, i.e., the purchase and sell price from/to the main grid is a constant, to avoid conflating the curtailment-mitigation effect with price-driven load redistribution. Readers can modify the model's price settings for further analyses, and the framework remains applicable under these extended scenarios.

4.3 Constraints of Day-Ahead Optimization Scheduling

(1) Constraints of delay-tolerant batch processing workloads

The control variables of the long-running non-interruptible workloads, long-running interruptible workloads, and short-running workloads are U_a, U_a^{index}, F_a^{mov}; U_b, U_b^{index}, F_b^{mov} and U_c, F_c^{mov}. Constraints of the control variables are detailed in Sections 3.1.2–3.1.4.

In addition to the above constraints, the other constraints are listed as follows.

(2) Computing power demand balance constraint

This study considers two long-running non interruptible workloads A1 and A2, two long-running interruptible workloads B1 and B2, and short-running workloads C. The total computing power demand of all the workloads at the original state F_{data}^{ori} and after shifting F_{data}^{mov} should follow the below constraints.

$$F_{data}^{ori} = F_{a1}^{ori} + F_{a2}^{ori} + F_{b1}^{ori} + F_{b2}^{ori} + F_{c}^{ori} + F_{mg}^{ori} \tag{44}$$

$$F_{data}^{mov} = F_{a1}^{mov} + F_{a2}^{mov} + F_{b1}^{mov} + F_{b2}^{mov} + F_{c}^{mov} + F_{mg}^{ori} \tag{45}$$

where $F_{a1}^{ori}, \ldots, F_{C}^{ori}$ are the computing power demand of delay-tolerant workloads A1, . . . , C at the original state. $F_{a1}^{move}, \ldots, F_{c}^{move}$ are the computing power demand of delay-tolerant workloads A1, . . . , C after shift. Since the delay-sensitive workloads don't shift, their computing power demand are denoted as F_{mg}^{ori}.

(3) Relation of computing power demand and electricity consumption

$$E_{data}^{t} = E_{idle} + \left(E_{peak} - E_{idle}\right) * \frac{F_{data}^{t}}{F_{dc}^{max}} \tag{46}$$

where F_{data}^{t} is the computing power demand of all workloads during time period t.

(4) Indoor temperature constant constraint

$$\theta_{in}^{t} = 25°\mathrm{C} \tag{47}$$

If day-ahead forecasted indoor temperature are set to be variable, aligning with the 1-min time step of the time-variant computer room temperature model, day-ahead scheduling would require matching this granularity, which would result in each day-ahead decision variable increasing to 1440 and significantly increasing computational burden. Thus, indoor temperature is fixed at 25°C during day-ahead scheduling. Updates are applied via intraday optimization, which incorporates a refined 1-min temperature variation model.

(5) Refrigeration constraint

The refrigeration power at time point t should meet the upper and lower limits constraints:

$$P_{cold}^{min} \le P_{cold}^{t} \le P_{cold}^{max} \tag{48}$$

In this research, P_{cold}^{min} is 0, P_{cold}^{max} is 500 kW. Detailed analysis can be found in Section 5.1.3.

The relation of P_{cold}^{t} and the electricity power consumption of refrigeration system E_{cold}^{t} is:

$$E_{cold}^{t} = P_{cold}^{t}/3.5 \tag{49}$$

(6) Upper and lower limit of renewable power utilization

The utilized wind and PV power at time point t. E_{wind}^{t} and E_{pv}^{t} have the constraints as below.

$$0 \le E_{wind}^{t} \le E_{wind,max}^{t} \tag{50}$$

$$0 \le E_{pv}^{t} \le E_{pv,max}^{t} \tag{51}$$

where $E_{wind,max}^{t}$ and $E_{pv,max}^{t}$ are the forecasts of maximum wind and PV power generation at time t.

(7) Exchange power constraint

The day-ahead planned purchased power and sold power E_{buy}^{t} and E_{sell}^{t}, as well as the state of power purchase and sell B_{buy}^{t} and B_{sell}^{t} have the following constraint.

$$B_{buy}^{t} * E_{buy}^{min} \le E_{buy}^{t} \le B_{buy}^{t} * E_{buy}^{max} \tag{52}$$

$$B_{sell}^{t} * E_{sell}^{max} \le E_{sell}^{t} \le B_{sell}^{t} * E_{sell}^{max} \tag{53}$$

$$B_{buy}^{t} + B_{sell}^{t} = 1 \tag{54}$$

where E_{buy}^{min} is 0, E_{buy}^{max} is 360 kW, E_{sell}^{min} is 0, E_{sell}^{max} is 200 kW.

(8) Power supply and demand balance constraint

The total power at supply side of data center equals to the power on demand side:

$$E_{wind}^{t} + E_{pv}^{t} + E_{buy}^{t} = E_{data}^{t} + E_{cold}^{t} + E_{sell}^{t} \tag{55}$$

4.4 Intraday Rolling Optimization Objective

Due to the different time steps of intra-day and day-ahead scheduling, in order to distinguish, the time point in intra-day scheduling is recorded as τ; The time step is recorded as $\Delta\tau$; The number of time points of a scheduling day is denoted as Γ. For each round of real-time optimization, the objective is to minimize the fluctuation of power purchase and sale and to optimize the operating cost during the optimization time frame, expressed as:

$$obj = min\frac{V_{grid}}{V_{grid}^{max}} + \frac{C_{op}}{C_{op}^{max}} \tag{56}$$

In the formula, there is:

$$V_{grid} = \sum_{\tau=1}^{60} \left|E_{buy*}^{\tau} - E_{buy}^{\tau}\right| + \left|E_{sell*}^{\tau} - E_{sell}^{\tau}\right| \tag{57}$$

$$C_{op} = \sum_{\tau=1}^{60} E_{buy*}^{\tau} K_{buy} \Delta\tau - E_{sell*}^{\tau} K_{sell} \Delta\tau + \sum_{\tau=1}^{\Gamma} K_{cur} E_{cur*}^{\tau} \Delta\tau \tag{58}$$

where V_{grid} is the fluctuation of the purchased and sold power between day-ahead plan and real-time plan for each round. C_{op} is the optimized operation cost for each round. V_{grid}^{max} is the fluctuation of the purchased and sold power between day-ahead and real-time plan with the minimization of operation costs as the only goal for each round of optimization. C_{op}^{max} is the operating costs with the minimization of fluctuation of the purchased and sold power between day-ahead and real-time plan as the only goal for each round of optimization. $E_{buy}^{\tau}/E_{sell}^{\tau}$ is the day-ahead planned purchased/sold electricity power at time point τ. $E_{buy*}^{\tau}/E_{sell*}^{\tau}$ is the real-time purchased/sold electricity power at time period τ. E_{cur*}^{τ} is the real-time renewable power curtailment at time period τ.

For each round, the control variables are real-time refrigeration power P_{cold*}^{τ}, refrigeration electricity consumption E_{cold*}^{τ}, utilized wind power E_{wind*}^{τ}, utilized solar power E_{pv*}^{τ}, purchased power E_{buy*}^{τ}, sold power E_{sell*}^{τ}, and indoor temperature θ_{in*}^{τ}. The workloads aren't suitable to be the control variable in

the rolling optimization. The day-ahead time-shift plan of workloads E^{t}_{data} are only adopted as boundary conditions in the intra-day scheduling. By considering the system operation constraints and combining real-time forecasts of renewable power generation and server electricity consumption ($E^{\tau}_{wind*,mppt}$, $E^{\tau}_{pv*,mppt}$, E^{τ}_{data}), the control variables are solved using the rolling optimization model.

With reference to Fig. 3, for each round optimization, the first 15-min time period of E^{τ}_{cold*}, E^{τ}_{wind*}, E^{τ}_{pv*}, E^{τ}_{buy*}, E^{τ}_{sell*}, θ^{τ}_{in*} will be saved. After multiple rounds of rolling optimization scheduling, the results of the control variables in 1440 min are obtained.

4.5 Constraints of Intraday Rolling Optimization

Constraints regarding these variables are described in the following section.

(1) Indoor temperature constraint

For each round, the real-time indoor temperature θ^{τ}_{in*} need to satisfy the upper and lower limit:

$$5^{\circ}C \le \theta^{\tau}_{in*} \le 30^{\circ}C \tag{59}$$

The adjacent indoor temperature $\theta^{\tau+1}_{in*}$ and θ^{τ}_{in*} in the intra-day scheduling satisfies the equilibrium constraint:

$$\theta^{\tau+1}_{in*} = \theta^{\tau}_{in*} + \frac{P^{\tau}_{heat} - P^{\tau}_{cold*}}{C_{in}} - \frac{\theta^{\tau}_{in*} - \theta^{\tau}_{out*}}{R1 * C_{in} + R2 * C_{in}} \tag{60}$$

where the indoor temperature at the start time θ^{0}_{in*} is 25°C. For each round of optimization thereafter, θ^{0}_{in*} is equal to the optimized indoor temperature at the 15th time point of the last optimization round. P^{τ}_{heat} is equal to E^{τ}_{data}, which is a known boundary condition derived from day-ahead scheduling plan.

The refrigeration power should satisfy the ramp constraint.

$$\theta^{\tau+15}_{in*} - \theta^{\tau}_{in*} \le \Delta\theta^{max15}_{in*} \tag{61}$$

$$\theta^{\tau+1}_{in*} - \theta^{\tau}_{in*} \le \Delta\theta^{max1}_{in*} \tag{62}$$

where $\theta^{\tau+1}_{in*}$ and θ^{τ}_{in*} are the indoor temperature at time $\tau + 1$ and τ. In this paper, $\Delta\theta^{max15}_{in*}$ and $\Delta\theta^{max1}_{in*}$ is the max fluctuation of indoor temperature within 15 min and 1minute, which are set to be 15°C and 0.33°C [24].

(2) Refrigeration power

The refrigeration power should satisfy the ramp constraint.

$$P^{\tau+1}_{cold*} - P^{\tau}_{cold*} \le \Delta P^{max}_{cold*} \tag{63}$$

where P^{τ}_{cold*} and $P^{\tau+1}_{cold*}$ are the refrigeration power at time point $\tau + 1$ and τ. In this paper, ΔP^{max}_{cold*} is set to be 100 kW.

The relationship of P^{τ}_{cold*} with the refrigeration electricity consumption E^{τ}_{cold*} is:

$$E^{\tau}_{cold*} = \frac{P^{\tau}_{cold*}}{3.5} \tag{64}$$

(3) Upper and lower limit of renewable power utilization

Real-time utilized wind and PV power at time τ E^{τ}_{wind*} and E^{τ}_{pv*} have the constraints as below.

$$0 \le E^{\tau}_{wind*} \le E^{\tau}_{wind*,mppt} \tag{65}$$

$$0 \le E^{\tau}_{pv*} \le E^{\tau}_{pv*,mppt} \tag{66}$$

where $E^{\tau}_{wind*,mppt}$ and $E^{\tau}_{pv*,mppt}$ are the 60-min ahead forecasts of maximum wind and PV power generation at time τ.

(4) Exchange power constraint

Real-time electricity purchase and sale E^{τ}_{buy*} and E^{τ}_{sell*}, as well as the state of electricity purchase and sale B^{τ}_{buy*} and B^{τ}_{sell*} have the following constraint.

$$B^{\tau}_{buy*} * E^{min}_{buy*} \le E^{\tau}_{buy*} \le B^{\tau}_{buy*} * E^{max}_{buy*} \tag{67}$$

$$B^{\tau}_{sell*} * E^{min}_{sell*} \le E^{\tau}_{sell*} \le B^{\tau}_{sell*} * E^{max}_{sell*} \tag{68}$$

$$B^{\tau}_{buy*} + B^{\tau}_{sell*} = 1 \tag{69}$$

where E^{min}_{buy*} is 0, E^{max}_{buy*} is 360 kW; E^{min}_{sell*} is 0, E^{max}_{sell*} is 200 kW.

(5) Power supply and demand balance constraint

For each optimization round, the power balance equation constraint at time τ is expressed as:

$$E^{\tau}_{wind*} + E^{\tau}_{pv*} + E^{\tau}_{buy*} = E^{\tau}_{data} + E^{\tau}_{cold*} + E^{\tau}_{sell*} \tag{70}$$

5 Boundary Conditions and Multi-Time Scale Optimization Scheduling Results

5.1 Boundary Conditions and Parameters of the Data Center

5.1.1 Parameters of the CPU, Server, Rack and Data Center

This study selected a medium-sized data center room in northern China as the example. The data center room occupies an area of 200 m^2, with a height of 4 m. The external wall area of the computer room is 240 m^2, and the roof area is 200 m^2. The data center room is equipped with 60 42U racks, each housing 10 4-way 4U servers. Detailed parameters for the CPU chip, server and rack are outlined in Table 1.

Based on Table 1, in this paper, F^{max}_{dc} is $3.84 * 10^6$ GFLOPS, E_{peak} is 360 kW. The E_{idle} is 216 kW, which is calculated by multiplying E_{peak} with a coefficient of 0.6 [25].

The thermal parameters of the computer room is adapted from literature [26] in accordance with the standards and specifications of data center room construction. The thermal parameters for the equivalent model of heat transfer process in the data center room are given in Table 2.

5.1.2 Parameters of Delay-Sensitive and Delay-Tolerant Workloads

The study considers two long-running non-interruptible workloads (A1 and A2), two long-running interruptible workloads (B1 and B2), and short-running workloads (C). The original computing power demand time series curves for the delay-sensitive interactive workload and delay-tolerant batch processing workloads are depicted in Fig. 4.

The shifting time range $[t_e, t_l]$ for the start node of the long-running non-interruptible workload A1 and A2 is [33, 73] and [1, 87]. The short-running workloads are allowed to be fully or partially transferred to the 1st–11th time points after the initial time point.

Table 2: Parameters for the equivalent model of heat transfer process in data center room

Parameters	Value
Equivalent thermal resistance of indoor air and the inside of wall R1/(°C/W)	4.8×10^{-4}
Equivalent thermal resistance of outside wall and outdoor air R2/(°C/W)	3.68×10^{-3}
Equivalent heat capacity of indoor air C_{in}/(J/°C)	4.60×10^{5}

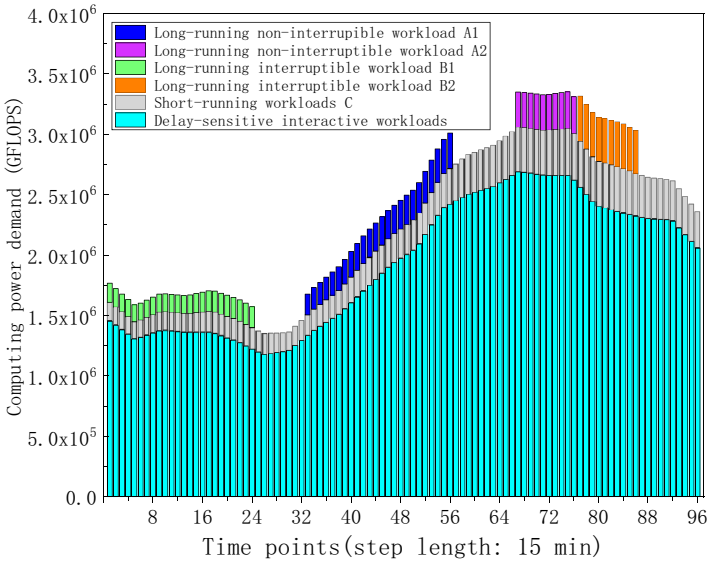

Figure 4: The original computing power demand time series curves of the delay-sensitive and various delay-tolerant workloads

5.1.3 Configuration of Refrigeration Power System

The heat load of data center mainly stems from the heat generated by servers and environmental maintenance structure. The rated refrigeration capacity of refrigeration system is determined based on the server equipment power and the area of data center room, as described by the formula:

$$P_{cold}^{rated} = P_{heat}^{max} + \beta S_a \tag{71}$$

where P_{heat}^{max} represents the maximum heat generation of the server cluster, which is equal to E_{peak} [17]. S_a denotes the area of the computer room, set as 200 m^2 for this study. The empirical coefficient β is assigned with a value of 0.7 [27]. Hence, the rated refrigeration capacity P_{cold}^{rated} is determined to be 500 kW for the computer room. With a comprehensive refrigeration performance coefficient of 3.5, the corresponding rated electricity consumption of the refrigeration system E_{cold}^{rated} is calculated to be 143 kW.

5.1.4 Configuration of Wind and PV Power System

The installed capacity of wind and PV power system, denoted as E_{vre}^{max}, is determined by formula (72) as follows:

$$E_{vre}^{max} * \eta_{vre} = E_{cold}^{rated} + E_{peak} \tag{72}$$

In this formula, the average power generation efficiency of renewable energy η_{vre} is assumed to be 0.5. With E_{cold}^{rated} and E_{peak} having the values of 143 and 360 kW, E_{vre}^{max} is calculated to be 1000 kW. This study stochastically set the proportion of wind and PV power capacity to be 3:7, resulting in capacities of 300 and 700 kW for wind and PV power system, respectively. This study focuses on developing a scheduling method considering the time-shiftable workloads and refrigeration system to facilitate the renewable energy utilization, while the economic analysis of allocating renewable energy is not the main concern. It is noted that varying the capacity ratios of wind and PV doesn't affect the effectiveness of the control strategy and conclusions drawn in this paper.

Wind and PV power generation data used in the study are sourced from a reliable database [28]. To align with the configured capacities of wind and PV power systems, the data is scaled to generate 24h-ahead forecasts of wind and PV power generation used in this study, as depicted in Fig. 5.

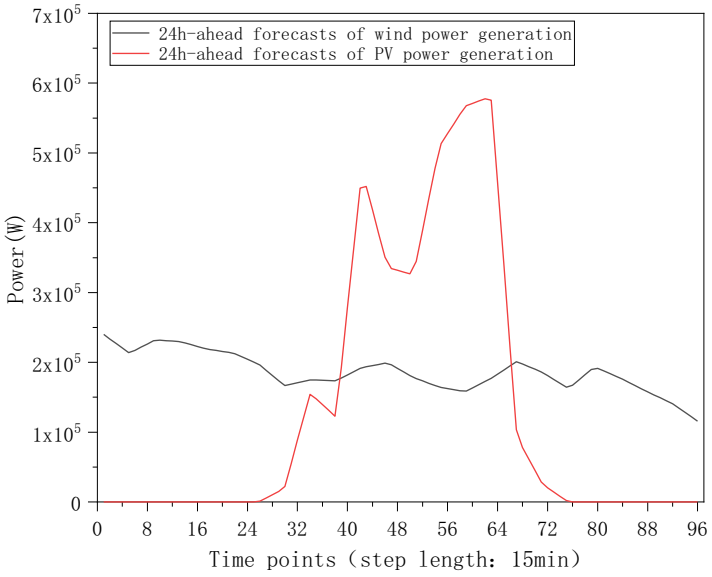

Figure 5: 24 h-ahead forecasts of wind and PV power generation for data center

5.2 Results of Multi-Time Scale Optimization Scheduling

This paper formulates both day-ahead scheduling and intra-day real-time scheduling optimization as Mixed Integer Programming (MIP) optimization problems. The optimization involves continuous variables, e.g., key time-shift parameter variables for various batch computing workloads, and binary variables, e.g., the status of power purchases and sales. Based on the decision variables, equality and inequality constraints, and objective function established in previous sections, the optimization model is solved using Gurobi solver in MATLAB R2024a. The Gurobi solver, widely recognized in academia and industry, employs techniques such as the interior point method, branch and bound method, and cut plane method to find optimal solutions.

This section presents the scheduling results after multi-time scale optimization. Section 5.2.1 details the operational costs and renewable power curtailment results after day-ahead scheduling and verifies its effectiveness by comparing with a simple scheduling scenario. Additionally, it analyzes the comparison of workloads at original status and after scheduling to illustrate the feasibility of the proposed time-shift modelling. Section 5.2.2 reports the operational costs and power exchange volatility results after real-time optimization scheduling. A comparison between the optimization scheduling and a plain real-time scheduling is also conducted to highlight the necessity of incorporating refrigeration regulation in real-time scheduling to cope with the instantaneous power volatility from renewable energy and load demand.

5.2.1 Results after Day-Ahead Optimization Scheduling

To quantitatively verify the effect of considering workloads time-shift, this paper sets up a simple scheduling scenario where the workloads don't shift over time. In the simple scheduling scenario, the power balance results of utilized wind power $E_{wind\#}$, utilized PV power $E_{PV\#}$, purchased power $E_{buy\#}$, server power consumption $E_{data\#}$, refrigeration electricity consumption $E_{cold\#}$ and sold power $E_{sell\#}$ are illustrated in Fig. 6. By considering the 24 h-ahead forecasts of wind and PV power generation and the system operation constraints, the daily operation plan in this scenario is obtained. Fig. 6 indicates that in this scenario, electricity sales mainly occur from the 39th–66th time periods, while electricity purchases predominantly occur from the 1st–38th and 67th–96th time periods. Considering the purchase and sale electricity prices, as well as the penalties for wind and PV power curtailment, the daily operating cost in this simple scheduling scenario amounts to ¥2980.6, and the planned curtailment of wind and solar power is 294.6 kWh (3.8% of the forecasted daily renewable generation).

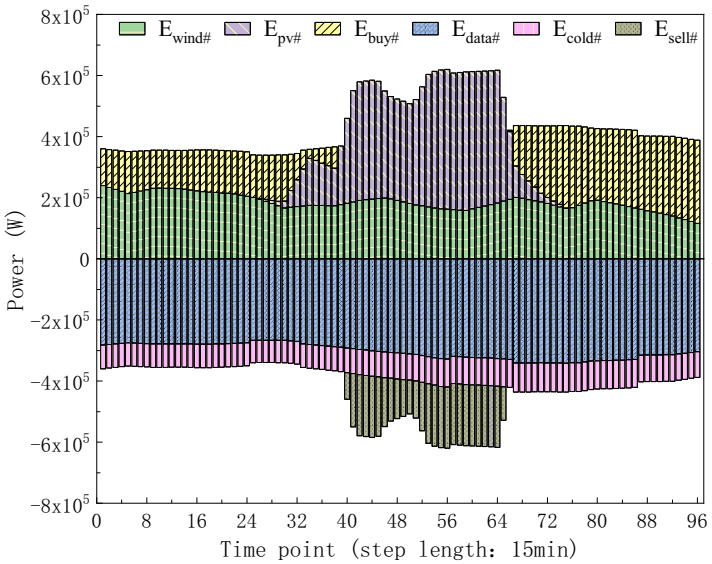

Figure 6: The power balance results in the day-ahead simple scheduling scenario

In the day-ahead optimization scheduling scenario where the time-shiftable workloads are involved, the power balance results of E_{wind}, E_{PV}, E_{buy}, E_{data}, E_{cold}, and E_{sell} for the scheduling day are presented in Fig. 7. A comparison of wind and PV power curtailment between the day-ahead simple scheduling scenario and the optimization scheduling scenario is depicted in Fig. 8. The computing power demand time series of workloads at the original state and after the temporal shift are shown in Figs. 9 and 10, respectively. Upon observing Figs. 9 and 10, it is evident that after optimization, long-running non-interruptible

workload A1 and A2, long-running interruptible workload B2 all moves to the 43rd–66th time periods, and long-running interruptible workload B1 and short-running workloads C partially shift to these time periods. Consequently, the renewable power curtailment during the 43rd–66th periods is reduced, while the electricity purchases during the 1st–38th periods and the 67th–96th periods decreases. The optimized operating cost plan is ¥1885.3 for the scheduling day, and the planned wind and PV power curtailment is 145.9 kWh (1.9% of the daily renewable generation), which respectively represents a 36.7% and 50.5% decrease compared to simple scheduling scenario. These comparison results prove that considering workload time-shift for the day-ahead optimization scheduling is advantageous in increasing the economic benefits of data center operation and promoting the consumption of renewable power.

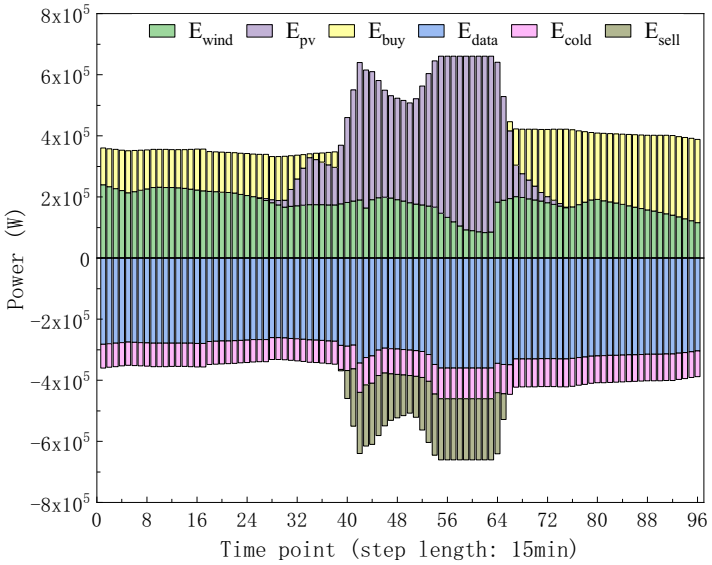

Figure 7: The power balance results in the day-ahead optimization scheduling scenario

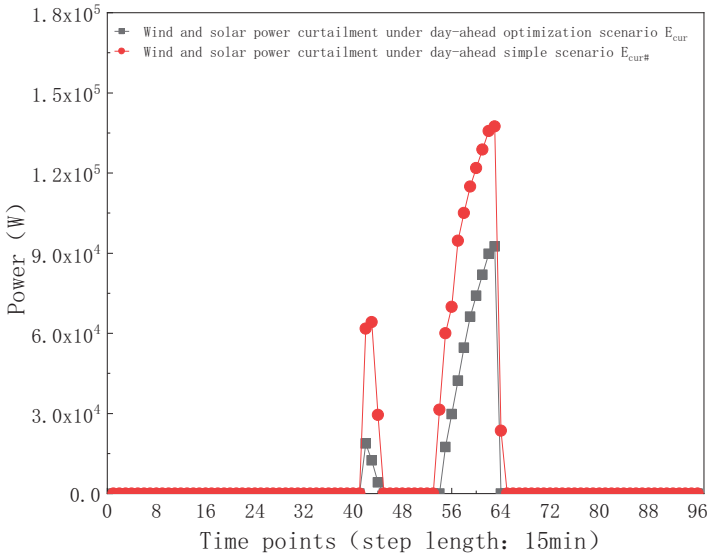

Figure 8: Comparison of wind and PV power curtailment in the day-ahead simple scheduling scenario and the optimization scheduling scenario

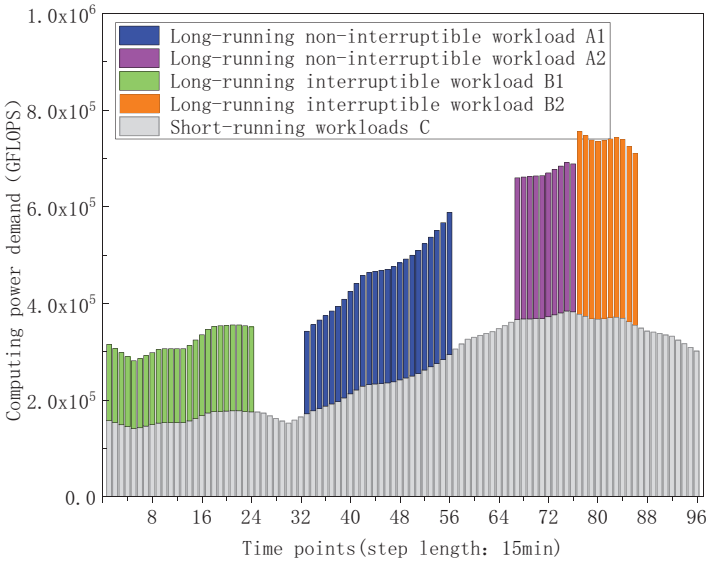

Figure 9: The computing power demand time series of various delay-tolerant workloads at the original state

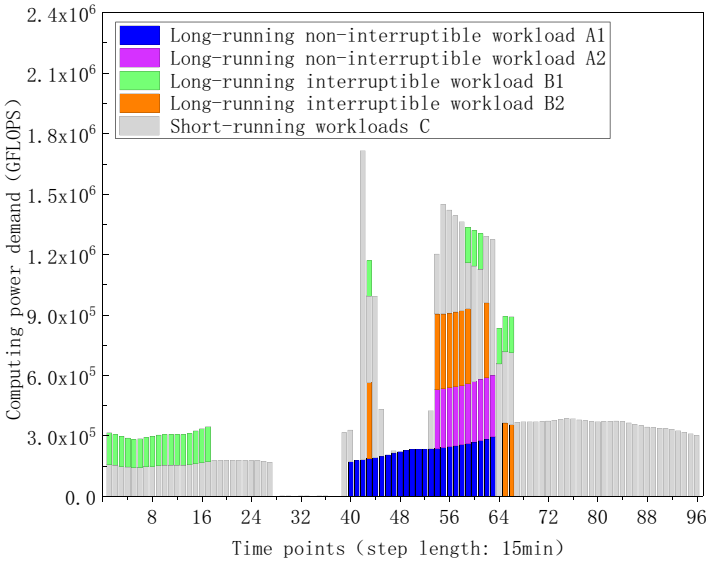

Figure 10: The computing power demand time series of various delay-tolerant workloads after temporal shift

The temporal shift details for the various delay-tolerant workloads are visualized in Figs. 11–15. Specifically, Figs. 11 and 12 display the computing power demand of long-running non-interrupted workloads A1 and A2 before and after shifting. Figs. 13 and 14 illustrate the computing power demand of long-running interruptible workloads B1 and B2, demonstrating that the sequential order of subtasks at different time points and the total computing power demand remain consistent before and after shifting. Additionally, Fig. 15 depicts the computing power demand of short-running workloads C before and after shifting. The results indicate a redistribution of computing power demand as short-running workloads transfer from initial time points to new time points, revealing a decrease in demand during the initial time periods and an increase at the new time periods. The total computing power demand stays unchanged before and after

the shifting process. These findings provide evidence supporting the feasibility of workloads' time shifting, demonstrating the adaptability and efficiency of the workloads scheduling strategy.

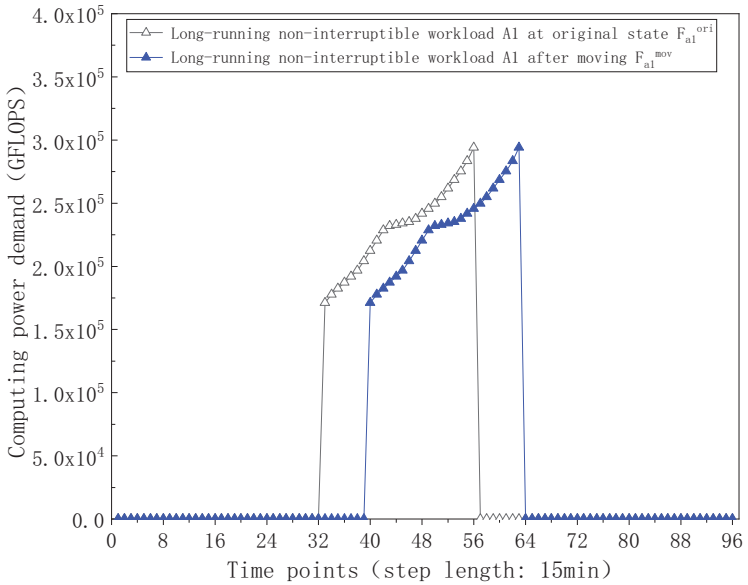

Figure 11: The computing power demand curve of long-running non-interruptible workload A1 at original state and after moving

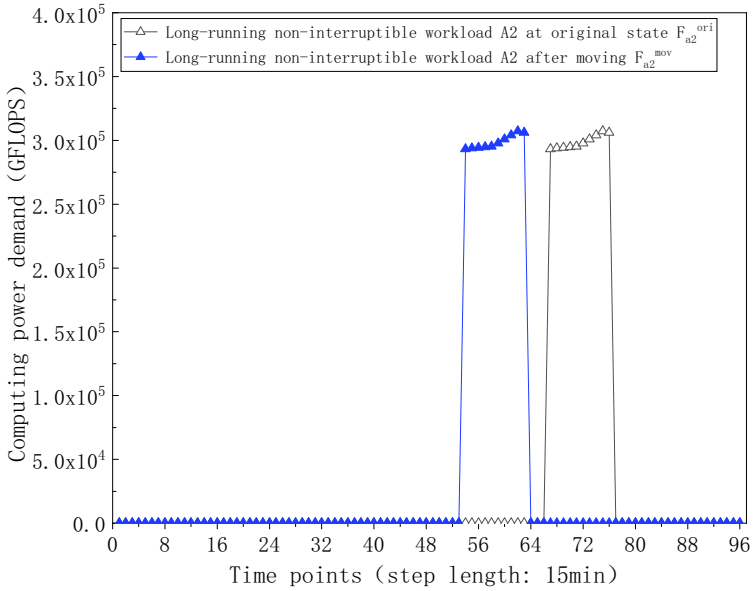

Figure 12: The computing power demand curve of long-running non-interruptible workload A2 at original state and after moving

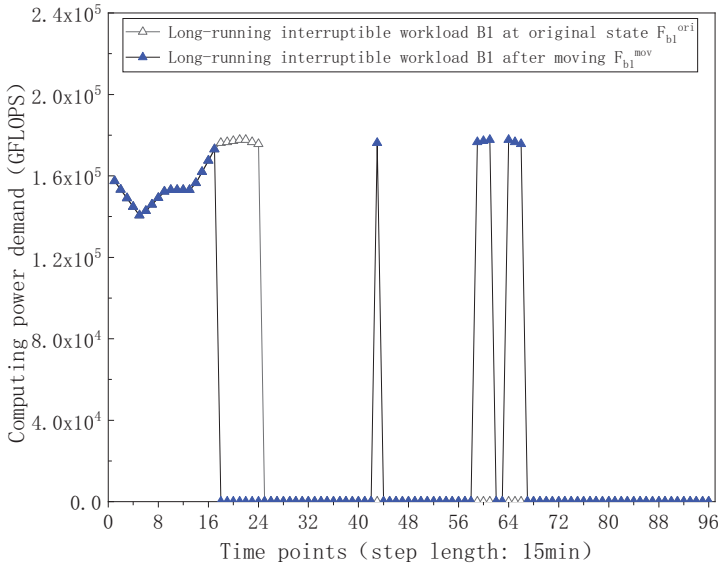

Figure 13: The computing power demand curve of long-running interruptible workload B1 at original state and after moving

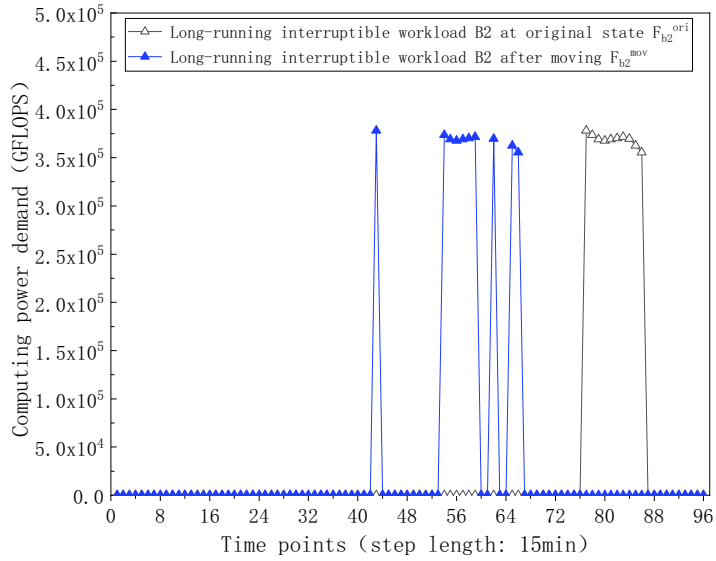

Figure 14: The computing power demand curve of long-running interruptible workload B2 at original state and after moving

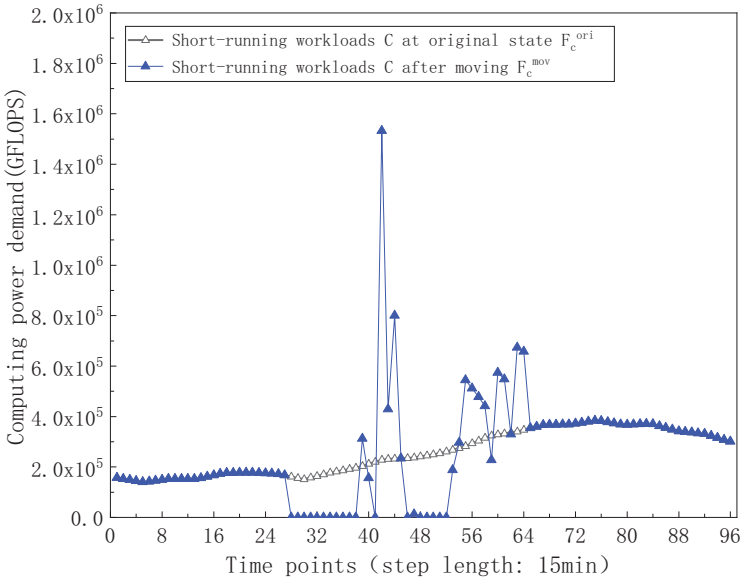

Figure 15: The computing power demand curve of short-running workloads C at original state and after moving

Among the three kinds of workloads, the short-running workloads have a special setting—a penalty coefficient in their time shift model. This paper discusses the rationale behind this setting by comparing the scenarios with and without the penalty coefficient. The time shift result of the various delay-tolerant workloads after the day-ahead optimization scheduling without the penalty coefficient is shown in Fig. 16.

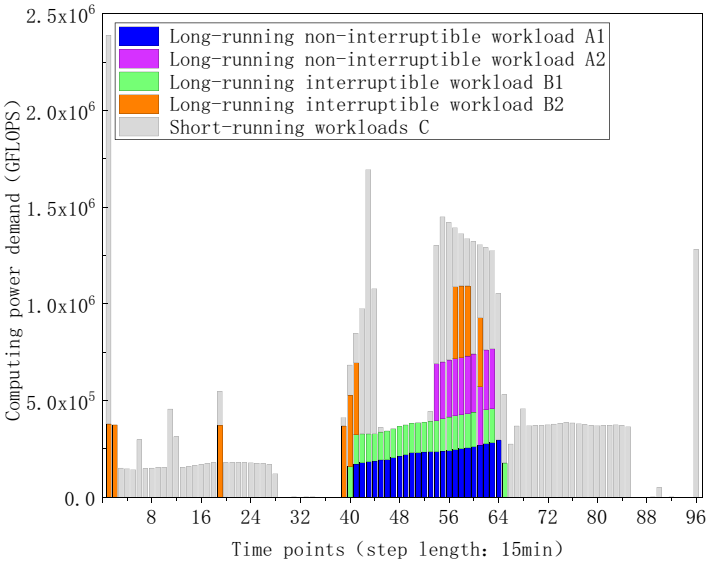

Figure 16: The computing power demand of delay-tolerant workloads after shifting in the scenario without the penalty coefficient setting for short-running workloads

(1) In the scenario with the penalty coefficient setting, most of the short-running workloads C are reallocated to the 43rd–66th time periods, where the surplus wind and solar power is utilized. The reason of this behavior is that, the cost of reducing the wind and PV power sales volume (0.4 + 0.05 =

0.45 ¥/kWh) is lower than the cost of purchasing electricity from other periods (1st–42nd and 67th–96th periods) (0.8 ¥/kWh). For another, this behavior may also lead to electricity revenue (0.6 − 0.05 = 0.55 ¥/kWh) due to reducing penalties for wind and PV power curtailment. Hence, operators are willing to pay lower costs or gain profits by using wind and PV power, rather than purchasing electricity from the main grid.

(2) In the scenario without the penalty coefficient setting, there is still an inflow and aggregation of short-running workloads in the 43rd–66th time periods. Meanwhile, there are also some workloads transferring from the 70th–96th period to the 1st–33rd time period. The mutual transfer of short-running workloads between the 1st–33rd and the 70th–96th time periods is an invalid time shift, since the unit price of power purchase in these two time frames is the same. Thus, the comparison indicates that setting the penalty coefficient for the time-shift model of short-running workloads can avoid invalid movement.

5.2.2 Results after Intra-Day Rolling Optimization

Considering the uncertainty of the day-ahead forecasts of power supply and demand, the day-ahead scheduling plans are updated and corrected by using the intra-day rolling optimization scheduling model. In order to carry out this model, the ultra-short-term 60 min-ahead forecasts of the wind, PV power generation, and server power consumption are needed. According to literature, the 7.5%, 5%, 2.5% mean absolute percentage error of day-ahead forecasts of wind, PV power and server electricity load compared to real-time forecasts are reasonable values [29,30]. Thus, the real-time forecasts of wind power, PV power generation, and server power consumption with a fluctuation ranges of 15%, 10%, and 5% compared to day-ahead forecasts are generated, as shown in Fig. 17a–c. The real-time outdoor temperature is given in Fig. 18.

Through the intra-day rolling optimization, the power balance results of utilized wind power E_{wind*}, utilized PV power E_{PV*}, purchased power E_{buy*}, server power consumption E_{data*}, refrigeration electricity consumption E_{cold*} and sold power E_{sell*} are as shown in Fig. 19. The fluctuation of real-time optimized exchange power relative to the day-ahead planed exchange power is shown in Fig. 20. Results show that the daily operating cost after real-time rolling optimization scheduling is ¥1895.9, and the deviation of real-time exchange power compared to day-ahead plan is 0.30 kW/min. Furthermore, optimized real-time indoor temperature, refrigeration power and heat generation, and net heat are shown in Fig. 21a–c. It can be observed that the indoor temperature fluctuates within the range of 15°C–30°C, and the temperature change rate is within the specified limits. The refrigeration power also meets the ramping constraint. Results indicate that the intra-day optimization scheduling can not only obtain the refined operation plan updates, but can also minimize the impact of uncertainty on the exchange power volatility, proving that the intra-day optimization scheduling plays an indispensable role in data center power scheduling.

For comparison, this paper sets up a simplified intra-day scheduling scenario that excludes refrigeration modulation flexibility. The calculation determines intra-day power purchases and sales by subtracting 60-min-ahead wind and PV generation forecasts from the combined day-ahead planned server and refrigeration electricity consumption. Fig. 22 depicts the fluctuation of real-time optimized exchange power to day-ahead planned exchange power in this simplified scenario.

After calculation, the exchange power deviation between real-time value and day-ahead plan is 14.92 kW/min, and the daily dispatching cost is ¥2141.4. In comparison to the simple intra-day scheduling scenario, the intra-day optimization scheduling model significantly reduces exchange power fluctuation (by 50 times) and lowers daily operation costs (by 12.9%). This demonstrates the model's effectiveness in minimizing fluctuations and optimizing system operation economy by leveraging the thermal inertia of the data center computer room and the flexibility of the refrigeration system.

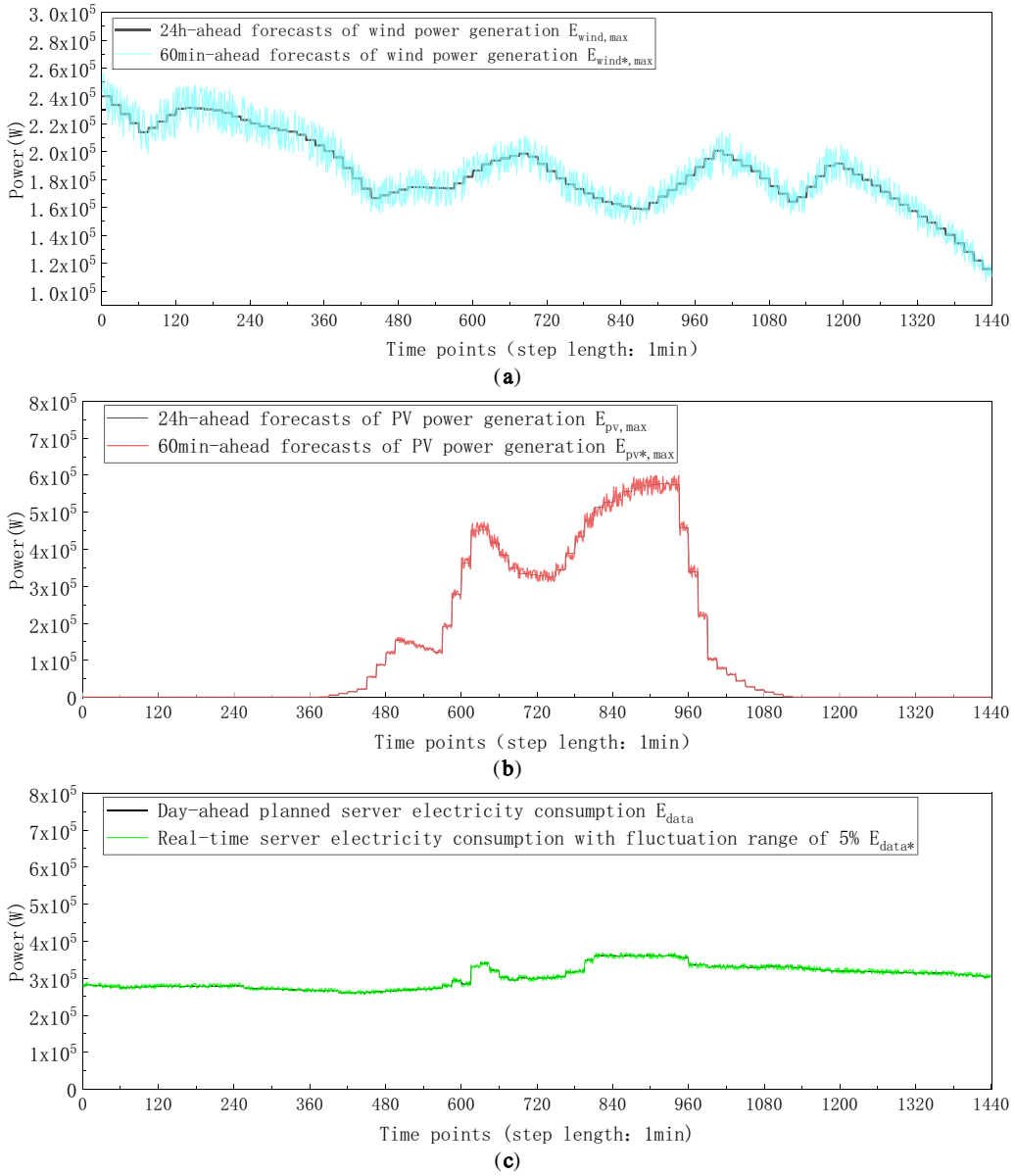

Figure 17: (**a**–**c**) Real-time forecasts (plans) and day-ahead forecasts (plans) of wind and PV power generation, and server power consumption. (**a**) The 60 min-ahead and 24 h-ahead forecasts of wind power generation (step length: 1 min). (**b**) The 60 min-ahead and 24 h-ahead forecasts of PV power generation (step length: 1 min). (**c**) Real-time server electricity consumption and day-ahead planned server electricity consumption (step length: 1 min)

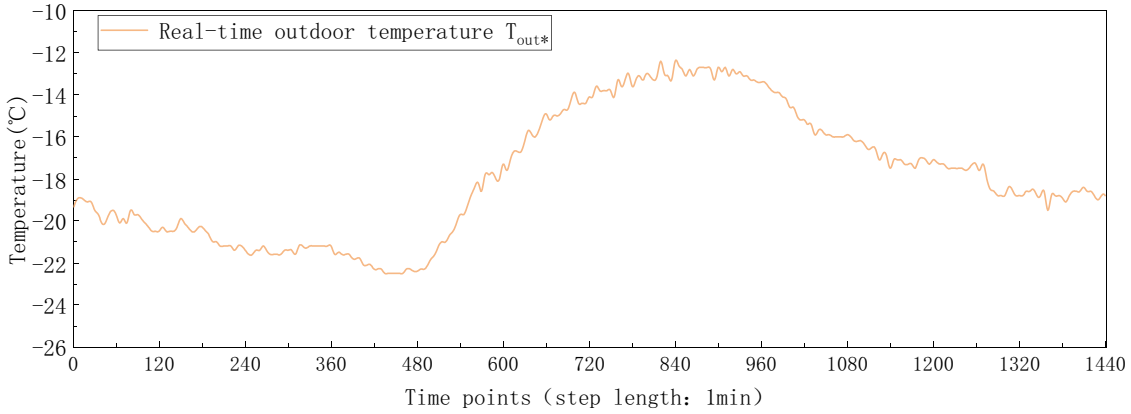

Figure 18: Real-time outdoor temperature (step length: 1 min)

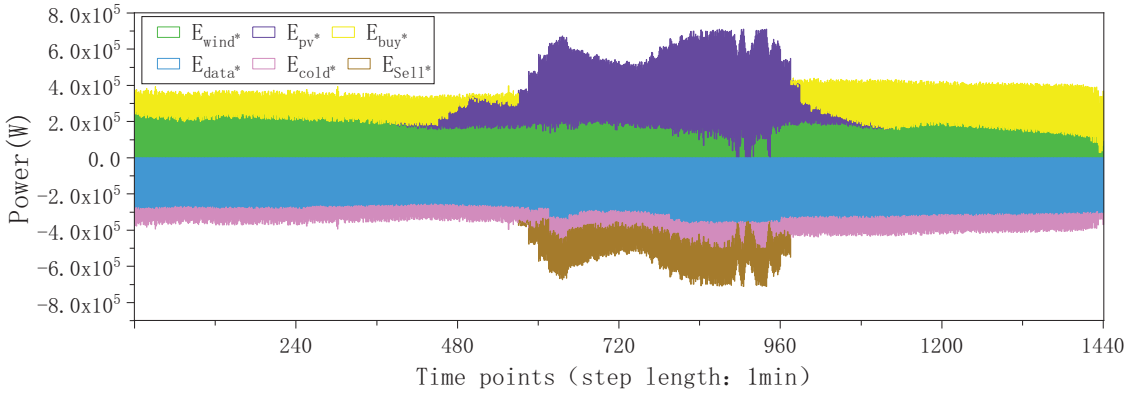

Figure 19: Power supply and demand results after intra-day rolling optimization (step length: 1 min)

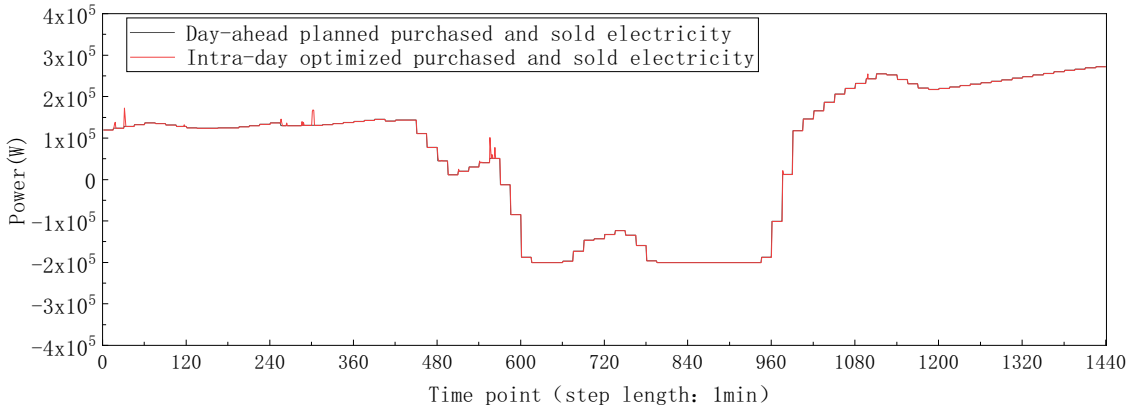

Figure 20: Fluctuation of real-time optimized exchange power relative to the day-ahead planed exchange power (step length: 1 min)

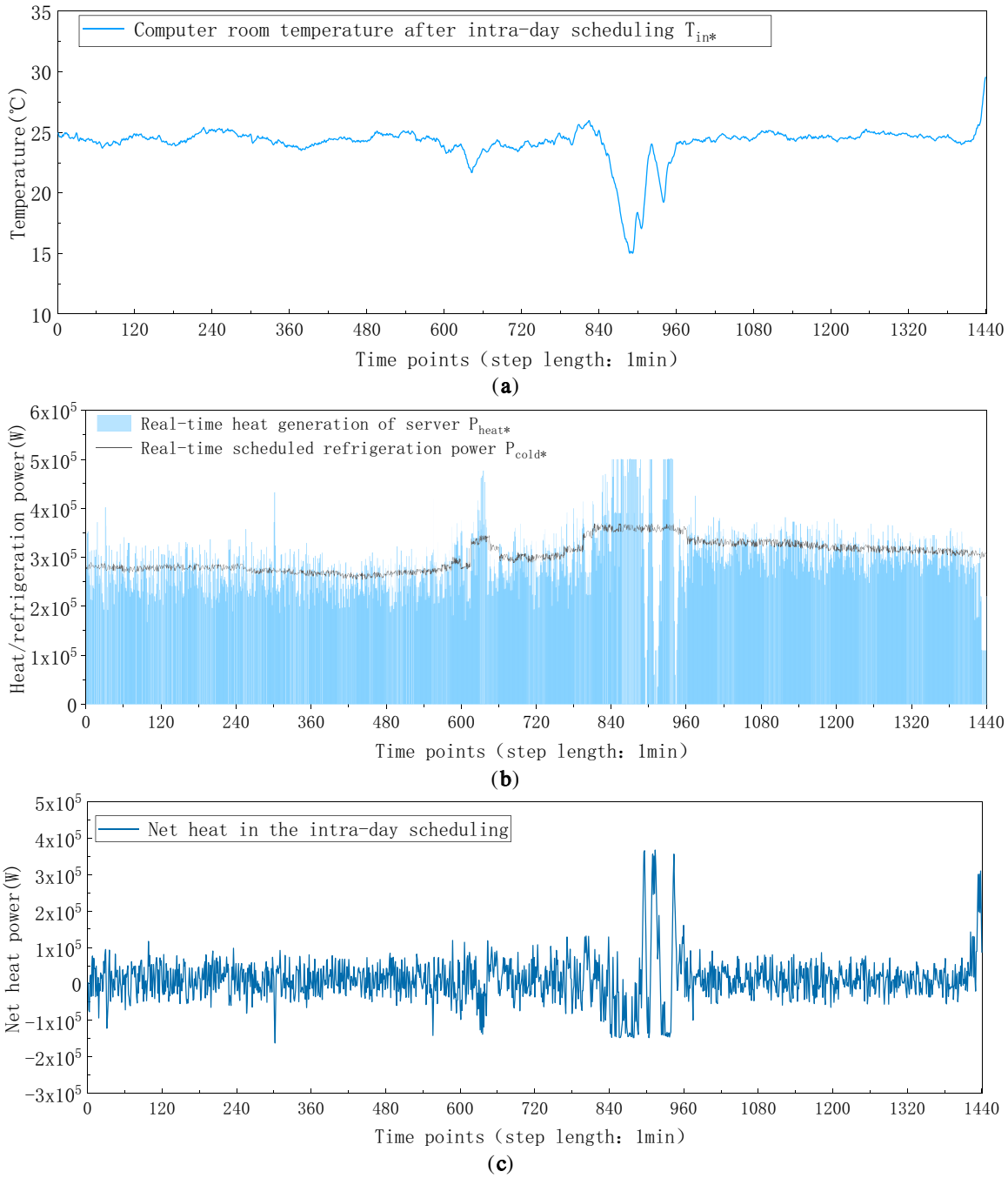

Figure 21: (**a**–**c**) Real-time indoor temperature, refrigeration power, heat generation, and net heat after the intra-day rolling optimization (step length: 1 min). (**a**) The indoor temperature after rolling optimization (step length: 1 min). (**b**) Real-time beat generation and seheduled refrigeration power (step length: 1 min). (**c**) Real-time net heat in the intra-day scheduling (step length: l min)

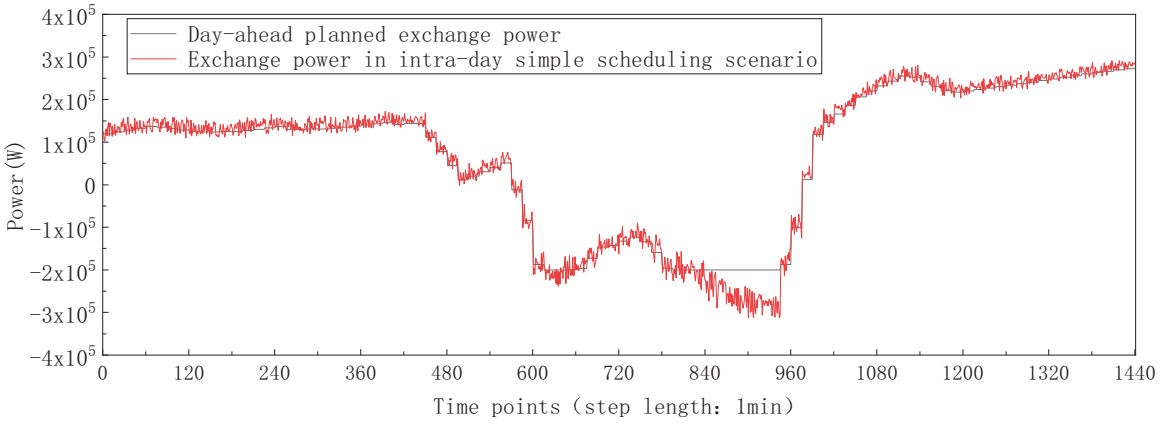

Figure 22: Fluctuation of real-time exchange power relative to day-ahead planned exchange power in the simple intra-day scheduling scenario

6 Conclusion

This study investigates the use of delay-tolerant computing workloads and refrigeration systems within data center micro-grid as flexible resources to accommodate volatile renewable generation. A two-stage multi-time scale optimization framework that integrates workload time-shift models into day-ahead scheduling and incorporates refrigeration power as a controllable variable in intra-day dispatch is developed to support green data-center operation. The proposed model is validated on a representative data-center case. Results show that accounting for workload shifting in day-ahead power scheduling reduces daily operating costs by 36.7% and decreases renewable curtailment by 50.5%. Including refrigeration control in intra-day scheduling further improves economic performance and mitigates the impact of renewable uncertainty on exchange-power volatility.

The contribution of this paper includes:

- This work identifies and categorizes three types of delay-tolerant workloads with distinct characteristics. Mathematical models are built to describe their time-shifting mechanisms. These models enable the workload time-shift behavior to be incorporated into renewable-aware power scheduling, allowing computing tasks to actively contribute to low-carbon data-center operation.
- The respective power regulation characteristics of delay-tolerant workloads and data center refrigeration system are analyzed and modelled. Based upon this foundation, a novel multi-time scale power-management framework is established, which achieves the coordination of these two kinds of flexible resources across different time scales to handle renewable power uncertainty.

In future, the deployment of renewable energy in data centers as well as the accelerated expansion of intelligent computation have become an inevitable trend. Further research into scheduling strategies and quantification of power regulation potential of the flexible computing workloads will therefore become essential. The time shift modelling methodology of workloads herein provides a solid foundation and could be applied in these application scenarios. Moreover, the renewable energy uncertainty continues to pose a significant challenge. The two-stage multi-time scale model developed in this study proves to be effective at managing power uncertainty and can be adapted to other green data centers. Overall, this paper contributes a practical management approach that leverages workload shifting and refrigeration flexibility to increase renewable-energy utilization and improve the operational economy of data centers, which provides support to the sector's green transition.

Acknowledgement: None.

Funding Statement: This work is supported by Science and Technology Standard Project of Guangdong Electric Power Design Institute (ER11301W; ER11811W).

Author Contributions: Luyao Liu: Conceptualization, formal analysis, investigation, methodology, writing—original draft; Xiao Liao: Conceptualization, fund acquisition, project administration, resources; Yiqian Li: Conceptualization, methodology, validation; Shaofeng Zhang: Validation, writing—review & editing. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be made available on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Nomenclature

ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
A	Long-running non-interruptible workload
F_a^{ori}	Original computing power demand time series of the long-running non-interruptible workload
f_a^1	Computing power demand value at the starting node of the long-running continuous workload
f_a^k	Computing power demand value of the long-running continuous workload at the k^{th} time points following the starting node
K	Processing time of the workload execution
U_a	Time-shift vector of the starting node of long-running continuous workload
u_a^t	Whether the starting node of the long-running non-interruptible workload is located at a certain time point, with a value of 0 or 1
U_a^{index}	The time at which the starting node is positioned
t_e	The earliest start time point of the long-running non-interruptible workload
t_l	The latest start time point of the long-running non-interruptible workload
F_a^{mov}	Computing power demand time series of the long-running non-interruptible workload after shifting
M_a^{mat}	A constructed auxiliary constant matrix, with a dimension of T^*T
B	Long-running interruptible workload
F_b^{ori}	Original computing power demand time series of long-running interruptible workload
f_b^1	Computing demand value of the *1st* subtask of long-running interruptible workload
f_b^k	Computing demand value of the k^{th} subtask of long-running interruptible workload
F_b^{mat}	A constant vector constructed based on values of $f_b^1,\ldots, f_b^K$
U_b	Time shift matrix variable of long-running interruptible workload
$u_b^{t,k}$	Whether the k^{th} subtask is moved to time point t, with a value of 0 or 1
U_b^{index}	Matrix of time points where subtasks locate after shifting
$u_b^{index,k}$	Time point where each subtask locates after shifting
F_b^{mov}	Computing power demand time series of each subtask after shifting
C	Short-running workloads
F_c^{ori}	Original computing power demand time series of short-running workloads
F_c^{mov}	Computing power demand time sequence of the short-running workloads after shift
f_c^k	Computing power demand value of short-running workloads at original status at time point k
f_{c*}^t	Computing power demand value of short-running workloads after shifting at time point t
$u_c^{t,k}$	The amount of computing power demand of short-running workloads that is transferred from the original time point k to another time point of t

U_c	The time shift matrix variable of the inflow and outflow computing power demand amount of short-running workloads
C_c^{mat}	A constructed constant matrix with the same dimension as the variable U_c
M	A constant, equal to the peak computing power demand of the short-running workloads.
D_c^{mat}	A penalty cost coefficient matrix to prevent invalid shifting
F_{cpu}^{max}	The maximum computing power of the CPU chip
E_{cpu}^{max}	The maximum power consumption of the CPU used in this study
FLOPS	FLoating-point Operations executed Per Second
N_{core}	The number of computing cores of the chip
f_{core}	The core frequency
Z_{core}	The double-precision floating-point operands per clock cycle of the core
F_{chip}^{max}	The computing power of the chip
F_{dc}^{max}	The computing power capacity of the data center using CPU chips
N_{rack}	The number of racks in the data center
N_{server}	The number of servers per rack
N_{CPU}	The number of CPU chips per server
u_{cpu}	The processor utilization rate
E_{idle}	The power of servers when u_{cpu} is zero
E_{peak}	The power of servers when u_{cpu} reaches 100%
F_{data}	The computing power demand of workloads
C_{in}	Equivalent heat capacity of indoor air
C_{wall}	Equivalent heat capacity of the wall
$R1$	Equivalent thermal resistance of indoor air and the inner side of the wall
$R2$	Equivalent thermal resistance of the outer wall and outdoor air
θ_{in}^t	Indoor temperature at time t
θ_{wall}^t	Wall temperature at time t
θ_{out}^t	Outdoor temperature at time t
P_{heat}^t	Heat generation of the servers at time t
P_{cold}^t	Refrigeration power at time t
θ_{wall}^{stable}	Mean value of wall temperature
θ_{in}^0	Initial indoor temperature
C_{grid}	Net cost of electricity transactions with the main-grid
C_{cur}	Penalties for wind and PV power curtailment
K_{buy}	Unit electricity purchase price from the main grid
K_{sell}	Unit electricity sale price to the main grid
E_{buy}^t	Day-ahead planned purchased electricity power during time period t
E_{sell}^t	Day-ahead planned sold electricity power during time period t
K_{cur}	Unit penalty price for wind and PV power curtailment
E_{cur}^t	Day-ahead planned renewable power curtailment in time period t
Δt	Duration of time period t
F_{a1}^{ori}	Computing power demand of delay-tolerant workloads A1
F_{a2}^{ori}	Computing power demand of delay-tolerant workloads A2
F_{b1}^{ori}	Computing power demand of delay-tolerant workloads B1
F_{b2}^{ori}	Computing power demand of delay-tolerant workloads B2
F_c^{ori}	Computing power demand of delay-tolerant workloads C
F_{a1}^{move}	Computing power demand of delay-tolerant workloads A1 after shift
F_{a2}^{move}	Computing power demand of delay-tolerant workloads A2 after shift
F_{b1}^{move}	Computing power demand of delay-tolerant workloads B1 after shift
F_{b2}^{move}	Computing power demand of delay-tolerant workloads B2 after shift

F_c^{move}	Computing power demand of delay-tolerant workloads C after shift
F_{mg}^{ori}	Computing power demand of the delay-sensitive workloads
θ_{in}^t	Day-ahead forecasted indoor temperature
P_{cold}^{min}	Lower limit of refrigeration power
P_{cold}^{max}	Upper limit of refrigeration power
E_{cold}^t	Day-ahead planned refrigeration electricity power consumption
E_{wind}^t	Day-ahead planned utilized wind power at time point t
E_{pv}^t	Day-ahead planned utilized PV power at time point t
$E_{wind,max}^t$	Day-ahead forecasts of maximum wind power generation
$E_{pv,max}^t$	Day-ahead forecasts of maximum PV power generation
B_{buy}^t	State of day-ahead power purchase
B_{sell}^t	State of day-ahead power sell
E_{buy}^{min}	Lower limit of power purchase
E_{buy}^{max}	Upper limit of power purchase
E_{sell}^{min}	Lower limit of power sell
E_{sell}^{max}	Upper limit of power sell
V_{grid}	Fluctuation of the purchased and sold power between day-ahead plan and real-time plan for each round
C_{op}	The optimized operation cost for each round
V_{grid}^{max}	Fluctuation of the purchased and sold power between day-ahead and real-time plan with the minimization of operation costs as the only goal for each round of optimization
C_{op}^{max}	The operating cost with the minimization of fluctuation of the purchased and sold power between day-ahead and real-time plan as the only goal for each round of optimization.
E_{buy}^{τ}	Day-ahead planned purchased electricity power at time point τ.
E_{sell}^{τ}	Day-ahead planned sold electricity power at time point τ.
E_{buy*}^{τ}	Real-time purchased electricity power at time period τ
E_{sell*}^{τ}	Real-time sold electricity power at time period τ
E_{cur*}^{τ}	Real-time renewable power curtailment at time period τ
θ_{in*}^{τ}	Real-time indoor temperature
θ_{in*}^{0}	Indoor temperature at the start time of intra-day rolling scheduling
P_{heat}^{τ}	Day-ahead planned heat generation of the servers at time τ
$\Delta\theta_{in*}^{max15}$	The max fluctuation of indoor temperature within 15 min
$\Delta\theta_{in*}^{max1}$	The max fluctuation of indoor temperature within 1minute
P_{cold*}^{τ}	Real-time refrigeration power at time period τ
ΔP_{cold*}^{max}	Real-time ramp constraint of refrigeration power
E_{cold*}^{τ}	Real-time refrigeration electricity consumption at time period τ
E_{wind*}^{τ}	Real-time planned utilized wind power at time point t
E_{pv*}^{τ}	Real-time planned utilized PV power at time point t
$E_{wind*,mppt}^{\tau}$	60-min ahead forecasts of maximum wind power generation
$E_{pv*,mppt}^{\tau}$	60-min ahead forecasts of maximum PV power generation
B_{buy*}^t	State of real-time power purchase
B_{sell*}^t	State of real-time power sell
P_{cold}^{rated}	Rated refrigeration capacity of refrigeration system
E_{cold}^{rated}	Corresponding rated electricity consumption of the refrigeration system
P_{heat}^{max}	The maximum heat generation of the server cluster
S_a	The area of the computer room
β	The empirical coefficient
E_{vre}^{max}	The installed capacity of wind and PV power system
η_{vre}	Average power generation efficiency of renewable energy

$E_{wind\#}$	Day-ahead planned utilized wind power in the simple scheduling scenario
$E_{PV\#}$	Day-ahead planned utilized PV power in the simple scheduling scenario
$E_{buy\#}$	Day-ahead planned purchased power in the simple scheduling scenario
$E_{data\#}$	Day-ahead planned server power consumption in the simple scheduling scenario
$E_{cold\#}$	Day-ahead planned refrigeration electricity consumption in the simple scheduling scenario
$E_{sell\#}$	Day-ahead planned sold power in the simple scheduling scenario

References

1. Masanet E, Shehabi A, Lei N, Smith S, Koomey J. Recalibrating global data center energy-use estimates. Science. 2020;367(6481):984–6. doi:10.1126/science.aba3758.
2. IEA (2024), Electricity 2024, IEA, Paris. [cited 2025 Aug 06]. Available from: https://www.iea.org/reports/electricity-2024.
3. Ni W, Hu X, Du H, Kang Y, Ju Y, Wang Q. CO2 emission-mitigation pathways for China's data centers. Resourc Conservat Recycl. 2024;202:107383. doi:10.1016/j.resconrec.2023.107383.
4. Patterson D, Gonzalez J, Le Q, Liang C, Munguia L, Rothchild D, et al. Carbon emissions and large neural network training. arXiv:2104.10350. 2021. doi:10.48550/arXiv.2104.10350.
5. Zhang Y, Tang H, Li H, Wang S. Integration and interaction of next-generation AI-focused data centers with smart grids and district energy systems: the state-of-the-art, opportunities and challenges. Renew Sustain Energ Rev. 2025;224:116097. doi:10.1016/j.rser.2025.116097.
6. Cao Y, Zhang S. Facilitating the provision of load flexibility to the power system by data centers: a hybrid research method applied to China. Utilities Policy. 2023;84:101636. doi:10.1016/j.jup.2023.101636.
7. Yang Z, Trivedi A, Liu H, Ni M, Srinivasan D. Two-stage robust optimization strategy for spatially-temporally correlated data centers with data-driven uncertainty sets. Elect Power Syst Res. 2023;221:109443. doi:10.1016/j.epsr.2023.109443.
8. Wang G, Pan C, Wu W, Fang J, Hou X, Liu W. Multi-time scale optimization study of integrated energy system considering dynamic energy hub and dual demand response. Sustain Energy, Grids Netw. 2024;38:101286. doi:10.1016/j.segan.2024.101286.
9. Zhou F, Gu W, Ma G. Advancements in data center cooling systems: from refrigeration to high performance cooling. Energy Build. 2024;320:114634. doi:10.1016/j.enbuild.2024.114634.
10. Chen M, Gao C, Shahidehpour M, Li Z. Incentive-compatible demand response for spatially coupled internet data centers in electricity markets. IEEE Trans Smart Grid. 2021;12(4):3056–69. doi:10.1109/TSG.2021.3053433.
11. Yang T, Jiang H, Hou Y, Geng Y. Carbon management of multi-datacenter based on spatio-temporal task migration. IEEE Trans Cloud Comput. 2021;11(1):1078–90. doi:10.1109/TCC.2021.3130644.
12. Wiesner P, Behnke I, Scheinert D,Gontarska K, Thamsen L. Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud. In: Proceedings of the 22nd International Middleware Conference; 2021 Dec 6–10; Online. p. 260–72. doi:10.48550/arXiv.2110.13234.
13. Niu T, Hu B, Xie K, Pan C, Jin H, Li C. Spacial coordination between data centers and power system considering uncertainties of both source and load sides. Int J Electr Power Energy Syst. 2021;124(2):106358. doi:10.1016/j.ijepes.2020.106358.
14. Liu Z, Chen Y, Bash C, Wierman A, Gmach D, Wang Z, et al. Renewable and cooling aware workload management for sustainable data centers. ACM SIGMETRICS Perform Eval Rev. 2012;40(1):175–86. doi:10.1145/2318857.2254779.
15. Cupelli L, Schutz T, Jahangiri P, Fuchs M, Monti A, Muller D. Data center control strategy for participation in demand response programs. IEEE Trans Ind Inform. 2018;14(11):5087–99. doi:10.1109/TII.2018.2806889.
16. Kwon S. Ensuring renewable energy utilization with quality of service guarantee for energy-efficient data center operations. Appl Energy. 2020;276:115424. doi:10.1016/j.apenergy.2020.115424.
17. Tang J, Su Z, Wang Y, Zhao M, Liu C, Xu J. Potential analysis of considering the participation of 5G Hongji station air conditioning load in demand response. Power Demand Side Manag. 2022;24(6):77–83. (In Chinese). doi:10.3969/j.issn.1009-1831.2022.06.013.

18. Zhu J, Yang S, Yu L, Jia H. Modeling and temperature control of real-time consumption of renewable energy in data center refrigeration systems. Pow Syst Automat. 2022;46(20):13–22. doi:10.7500/AEPS20220212003.
19. Reiss C, Tumanov A, Ganger GR, Katz RH, Kozuch MA. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In: Proceedings of the Third ACM Symposium on Cloud Computing; 2012 Oct 14–17; San Jose, CA, USA. p. 1–13. doi:10.1145/2391229.2391236.
20. Lu C, Ye K, Xu G, Xu C-Z, Bai T. Imbalance in the cloud: an analysis on Alibaba cluster trace. In: 2017 IEEE International Conference on Big Data (Big Data). Piscataway, NJ, USA: IEEE; 2018. p. 2884–92. doi:10.1109/BigData.2017.8258257.
21. Beloglazov A, Abawajy J, Buyya R. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener Comput Syst. 2012;28(5):755–68. doi:10.1016/j.future.2011.04.017.
22. Doyle J, Shorten R, O'mahony D. Stratus: load balancing the cloud for carbon emissions control. IEEE Trans Cloud Comput. 2013;1(1):1. doi:10.1109/TCC.2013.4.
23. Second-generation intelligent Intel ® xeon ® Scalable processor. [cited 2025 Aug 06]. Available from: https://ark.intel.com/content/www/cn/zh/ark/products/series/192283/2nd-generation-intel-xeon-scalable-processors.html.
24. ASHRAE 41.I-2020 standard method for temperature measurement. [cited 2025 Aug 06]. Available from: https://www.ashrae.org/.
25. Jin C, Bai X, Yang C, Mao W, Xu X. A review of power consumption models of servers in data centers. Appl Energy. 2020;265:114806. doi:10.1016/j.apenergy.2020.114806.
26. Chinese Society of Electrical Engineering. Topic A of the 15th CSEE national college students electrical mathematics modelling competition. [cited 2025 Aug 06]. Available from: http://shumo.neepu.edu.cn. (In Chinese).
27. Wang S, Tu R, Chen X, Yang X, Jia K. Thermal performance analyses and optimization of data center centralized-cooling system. Appl Therm Eng. 2023;222:119817. doi:10.1016/j.applthermaleng.2022.119817.
28. European commission Science Hub. [cited 2025 Aug 06]. Available from: https://joint-research-centre.ec.europa.eu/index_en.
29. Giebel G, Kariniotakis G. Wind power forecasting—a review of the state of the art. In: Woodhead publishing series in energy, renewable energy forecasting. Oxford, UK: Woodhead Publishing; 2017. p. 59–109. doi:10.1016/B978-0-08-100504-0.00003-2.
30. Liu L, Sun Q, Wennersten R, Chen Z. Day-ahead forecast of photovoltaic power based on a novel stacking ensemble method. IEEE Access. 2023;11:113593–604. doi:10.1109/ACCESS.2023.3323526.