Energy Engineering

Tech Science Press

Doi:10.32604/ee.2025.073720

ARTICLE

Collaborative Scheduling Strategy for Computation and Power among Multiple Data Centers Based on New Energy State Recognition

Qian Yang[1], Shenglei Du[1,2], Boyang Chen[1], Yalu Sun[1], Ding Li[1] and Zhiheng Zhang[3,*]

[1]State Grid Gansu Electric Power Company Economic and Technological Research Institute, Lanzhou, 730050, China
[2]School of Information Science and Engineering, Lanzhou University, Lanzhou, 730000, China
[3]School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou, 730050, China

*Corresponding Author: Zhiheng Zhang. Email: 17852156837@163.com

Received: 24 September 2025; Accepted: 31 October 2025; Published: 12 July 2026

ABSTRACT: In recent years, as the core infrastructure of the digital economy, data centers have witnessed increasingly prominent issues of energy consumption and carbon emissions. To achieve the goals of "carbon peak" and "carbon neutrality", data centers have gradually introduced new energy power such as wind and photovoltaic power. However, the randomness and volatility of their output pose challenges to efficient absorption. Based on the spatiotemporal complementary characteristics of new energy output in multiple data centers and the spatiotemporal migration capability of computing tasks, this paper proposes a new energy-aware adaptive collaborative scheduling strategy for computation and power. The strategy first constructs a regionally differentiated load model to accurately depict the characteristic differences among the Jiangsu-Zhejiang-Shanghai mixed computing power hub, the Gansu high-efficiency computing power base, and the coastal green computing power nodes. Then, a dual-mode scheduling algorithm based on Lyapunov optimization is designed, integrating a prediction-reaction mechanism to achieve dynamic balance between system stability and new energy absorption rate. Furthermore, a V-parameter adaptive adjustment mechanism and a hierarchical fault-tolerant guarantee system are proposed to cope with new energy fluctuations and improve system robustness. Simulation results show that the proposed strategy achieves an average new energy absorption rate of 62.3% and 52.8% in normal weather and severe weather scenarios, respectively. The carbon emission per unit computing power is reduced by 20.9%, and the computing power-electricity efficiency is improved by 9.1%, which is significantly better than the static scheduling strategy. This verifies its effectiveness and practicability in improving new energy utilization, ensuring service quality, and reducing carbon emissions.

KEYWORDS: Collaborative scheduling for computation and power; data centers; Lyapunov optimization; V-parameter adjustment mechanism; wind-solar volatility

1 Introduction

In recent years, with the rapid development of cloud computing, big data, and artificial intelligence technologies, the number and scale of global data centers have continued to expand, and their energy consumption issues have become increasingly prominent. By the end of 2022, the number of in-use data center racks in China was approximately 6.7 million, maintaining a high annual growth rate of 30%, and its high energy consumption problem cannot be ignored references [1–3]. At the same time, to achieve the strategic goals of "carbon peak" and "carbon neutrality", data center operators have begun to introduce large-scale new energy power such as wind and photovoltaic power reference [4]. Beyond the challenges posed by the intermittency of renewable energy, the complexity of modern digital infrastructures further accentuates

the urgency of efficient scheduling strategies. International Energy Agency (IEA) reports indicate that global data centers consumed nearly 1% of the world's total electricity demand in 2022, with projections suggesting a potential doubling of this share by 2030 if no transformative measures are adopted.

Existing industrial practices, including liquid cooling systems, dynamic voltage and frequency scaling, and server virtualization references [5–7], have curbed power consumption to a limited extent. These approaches primarily optimize hardware operation efficiency rather than addressing the systemic challenge of aligning computational demand with fluctuating renewable supply. In this context, collaborative scheduling across geographically distributed data centers emerges as a promising solution: unlike single-site optimization constrained by local renewable availability, cross-regional coordination fully leverages spatiotemporal complementarity reference [8]. Furthermore, delay-tolerant workloads references [9,10] enable temporal task migration, allowing dynamic alignment with renewable-rich regions to balance service reliability and carbon-free power utilization reference [9–11].

Recent academic focus on green data center scheduling has centered on three directions:

First, local optimization bias prevents leveraging cross-regional renewable complementarity. Reference [12] models job scheduling as an integer programming problem to minimize server activation and energy consumption, but it is confined to single data centers and cannot exploit such complementarity. Studies on server energy efficiency and industrial hardware optimizations similarly focus solely on local efficiency, failing to link computational demand to renewable supply fluctuations across regions.

Secondly, the insufficient adaptability to the volatility of new energy sources limits the flexibility of dispatching. Literature [13] proposed a distributed algorithm that utilizes regional price differences to reduce electricity costs; literature [14] designed a dynamic energy storage control strategy based on Q-Learning; literature [15] adopted game theory for workload management to minimize operational costs. Although these studies prioritize cost reduction, they fail to fully exploit the potential of deferrable loads as flexibility resources. Literature [16] proposed a deep learning dual-time-scale scheduling strategy considering task delay and price differences, but its task classification is overly simplified and lacks a mechanism to adapt to the fluctuations of renewable energy. The spatio-temporal task migration emission reduction mechanism proposed in literature [17] still does not fully exploit the time migration flexibility of deferrable loads and lacks an adaptive mechanism for renewable energy fluctuations,resulting in scheduling always failing to keep up with the changes in renewable energy supply.

Third, the weak robustness in dynamic scenarios damages reliability due to the lack of fault-tolerant design in critical tasks. Reference [18] adopts the Lyapunov online optimization method, which defines virtual queues to handle time coupled constraints such as energy storage capacity and load service quality. By relying solely on real-time measurement data, multi unit collaborative scheduling can be achieved without relying on accurate prediction models, significantly improving the system's adaptability to renewable energy fluctuations. Reference [19] designs a distributed quadratic control strategy based on the Lyapunov stability theorem, incorporating frequency, voltage, and power information into feedback control to suppress system oscillations while achieving voltage recovery and power equalization. In cloud computing and edge computing scenarios, literature [20] deduces the upper bound of the objective function through the Lyapunov queue model, which solves the instantaneous greed problem of the traditional Lyapunov method, reduces the system energy consumption and ensures the stability of the queue. Reference [1] proposes an adaptive load scheduling algorithm for multi-data centers, which realizes the computing task scheduling among three geographically separated computing power parks by identifying illumination conditions.

To address these gaps, this paper proposes a new energy-aware adaptive collaborative scheduling strategy, building on prior work in three key ways:

Against the lack of regional differentiation in existing models,this paper proposes a regionally differentiated load model to accurately depict the Jiangsu-Zhejiang-Shanghai mixed computing power hub, Gansu high-efficiency computing power base, and coastal green computing power nodes. To compensate for the lack of adaptive response to new energy volatility, we design a dual-mode scheduling algorithm based on Lyapunov optimization, integrating an LSTM-based new energy prediction mechanism and real-time adjustment to balance system stability and energy efficiency. To address insufficient robustness in dynamic scenarios, we propose a V-parameter adaptive adjustment mechanism and a hierarchical fault-tolerant guarantee system to cope with new energy fluctuations and hardware risks.

2 Model Construction

2.1 System Architecture

The geographically distributed multi-data center collaborative scheduling architecture designed in this paper is shown in Fig. 1, adopting a "distributed processing + centralized scheduling" mode. The system uses 15 min as the scheduling granularity (96 time points per day), which can effectively respond to short-term fluctuations of new energy while avoiding communication overhead caused by frequent scheduling. The system planned in this paper involves interactive scheduling among three characteristic parks: the new energy computing power park has sufficient new energy capacity, the high-efficiency computing power base undertakes huge computing power demand, and the mixed computing power hub has the power characteristics of an integrated energy system and also includes various characteristic computing power demands. There are distinct regional differences among the three parks, and cross-regional computing power transmission is realized through optical cables. The computing power capacities undertaken by the three parks are different, but they all belong to the computing power load types designed in the following text.

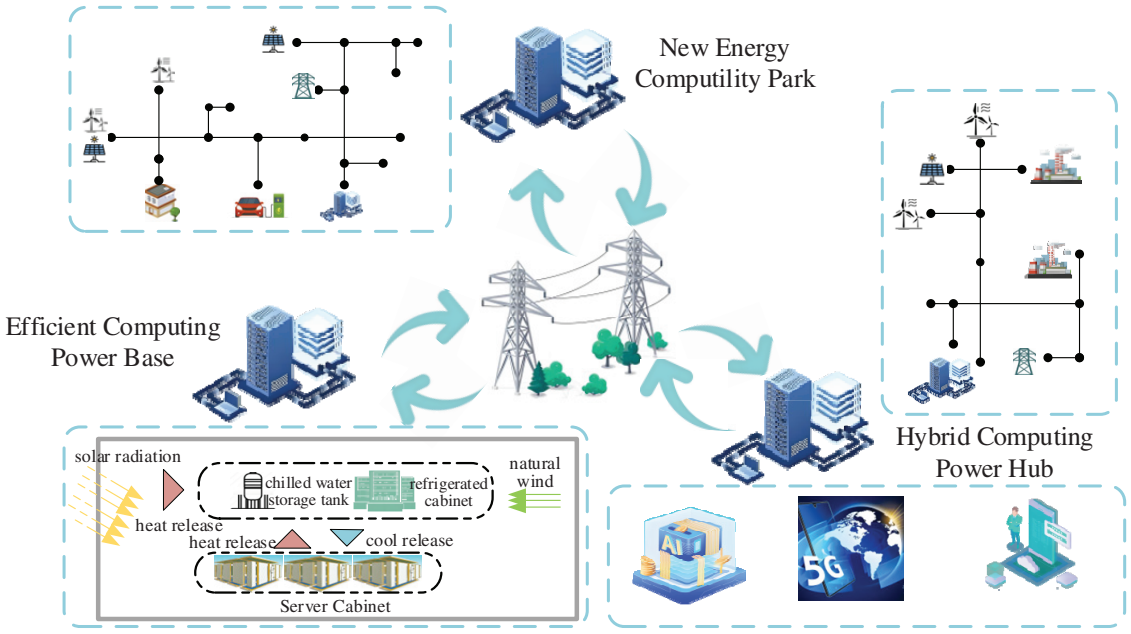

Figure 1: Collaborative scheduling architecture of multiple data centers

The core of the architecture lies in the organic integration of "local response and global optimization": each data center independently processes local tasks and monitors energy consumption parameters,

while the centralized adaptive scheduler achieves cross-regional collaborative decision-making by real-time collecting status data such as server utilization and new energy absorption rate.

Based on existing research, we have found that the scheduling modeling between multi-data center clusters focuses more on the mechanism characterization of park-specific features, the logical design of collaborative scheduling, and the consideration of the diversity of influencing factors. Therefore, in the subsequent modeling, while ensuring a clear representation of park performance, we will minimize irrelevant parameters related to internal park equipment or parameters with insignificant impacts, focus on the in-depth influence of environmental changes on scheduling strategies, and emphasize the delay characteristics and cross-regional transfer characteristics of the model based on this influence. Since this paper focuses on the interaction between multi-data center clusters, some internal park parameters have been omitted; however, in practical applications, local business requirements and energy characteristics should be considered, and the operating parameters within the park should be appropriately refined.

2.2 Power Generation Model

New energy plants are typically equipped with wind power generation, photovoltaic power generation, and various other characteristic power generation forms. To simplify the model, it is assumed in this paper that only wind power generation and photovoltaic power generation are considered in the parks, with differences in capacity configuration among different parks.

$$P_{\mathrm{WT}} = \begin{cases} 0 & V_{\mathrm{co}} < V < V_{\mathrm{ci}} \\ \dfrac{V - V_{\mathrm{ci}}}{V_{\mathrm{r}} - V_{\mathrm{ci}}} & P_{\mathrm{r}} V_{\mathrm{ci}} \le V \le V_{\mathrm{r}} \\ P_{\mathrm{r}} & V_{\mathrm{r}} \le V \le V_{\mathrm{co}} \end{cases} \tag{1}$$

In the formula:

P_{WT} represents the operating power of wind power generation;

V_{ci} is the cut-in wind speed of the wind power generation unit (the minimum wind speed for grid-connected power generation);

V_{co} is the cut-out wind speed of the wind power generation unit (the maximum wind speed for grid-connected power generation);

V is the actual wind speed, and the subscript 'r' denotes the rated working condition.

$$P_{\mathrm{u}} = P_{\mathrm{sm}} \frac{G_{\mathrm{T}}}{G_{\mathrm{S}}} \left[1 + K\left(T_{\mathrm{C}} - T_{\mathrm{S}}\right)\right] \tag{2}$$

In the formula:

P_{u} and P_{sm} are the operating power of photovoltaic power generation and the maximum power under standard conditions, respectively;

G_{T} and G_{S} represent the solar irradiance during operation and the solar irradiance under standard conditions, respectively;

K is the temperature coefficient;

T_{C} is the temperature of the solar panel during operation;

T_S is the temperature of the solar panel under standard conditions.

The adoption of simplified wind and photovoltaic models in this study is justified by their suitability for large-scale scheduling optimization. Although real-world generation systems exhibit more complex characteristics, the simplified equations effectively capture the dominant trends of renewable output while ensuring computational tractability. This trade-off is especially important when designing scheduling frameworks that operate at high temporal resolution, such as the 15-min granularity applied in this work. By focusing on the principal factors of wind speed and solar irradiance, the model balances analytical clarity with practical feasibility. Nevertheless, it should be noted that the framework is inherently extensible: additional variables such as turbulence intensity, panel degradation, or inverter dynamics could be incorporated in future studies if finer accuracy is required. In this sense, the simplified formulation provides both a robust baseline and a foundation for progressive refinement.

2.3 Computing Power Load Model

In the computation-power collaborative scheduling system, computing tasks are strictly divided into two categories—sensitive tasks and tolerant tasks—based on their delay sensitivity, migration characteristics, and service requirements. This classification method, grounded in the physical constraints and operational characteristics of actual service scenarios, achieves optimal coordination between computing resources and new energy supply through differentiated mathematical models and processing mechanisms.

Sensitive tasks. They have strict real-time requirements and must be processed within the current scheduling cycle. Cross-data center migration is prohibited, and their core characteristics are as follows: The task lifecycle is strongly bound to the scheduling cycle, and failure to process in a timely manner is regarded as a Service Level Agreement (SLA) violation; Due to data sovereignty and security constraints, they must be completed in the local data center; Resource allocation has the highest priority, requiring stable computing power supply.

Tolerant tasks. They have significant time flexibility and support cross-cycle processing and cross-data center migration. Their core characteristics are as follows: They support task queue buffering; They have spatial migration capabilities and can be dynamically scheduled following the distribution of new energy; They support task splitting and merging to adapt to heterogeneous computing environments.

This classification is not only conceptually intuitive but also supported by empirical evidence from data center operations. Industry reports indicate that more than 40% of workloads in large-scale cloud platforms can tolerate flexible scheduling, particularly in applications such as offline analytics, backup services, and distributed training of artificial intelligence models. By contrast, latency-sensitive workloads such as e-commerce transactions or financial clearing systems account for a smaller share but require disproportionate resource guarantees. This asymmetry provides a valuable opportunity: the system can allocate baseline capacity to guarantee sensitive tasks while dynamically reallocating surplus renewable energy to tolerant tasks. As a result, renewable integration is achieved without compromising user experience.

2.4 Data Center Energy Consumption Model

The quadratic server power model is a widely adopted and empirically validated approximation that captures the nonlinear relationship between CPU utilization and power draw more accurately than linear models reference [6].

The use of a constant PUE value for each data center assumes that cooling system power consumption is linearly correlated with IT load at the aggregate level. This is a standard simplification for system-level studies and is supported by operational data from various data center operators reference [4].

The server power consumption model uses a quadratic function to characterize the nonlinear relationship between utilization and energy consumption:

$$P_{server} = P_{idle} + b \times u + c \times u^2 \tag{3}$$

where P_{server} is the server energy consumption; u is the server utilization ($0 \le u \le 1$); $P_{idle} = 0.5$ kW is the idle power consumption; $P_{max} = 2.5$ kW is the maximum power consumption. The coefficients are as follows: $b = (2.5 - 0.5) \times 0.8 = 1.6$; $c = (2.5 - 0.5) \times 0.2 = 0.4$.

The model can accurately reflect the characteristic that power consumption accelerates when the load exceeds 50%, enabling subsequent scheduling strategies to avoid prolonged high-load operation.

Comprehensive energy efficiency evaluation is carried out from two dimensions: computing power density (PFLOPS/kW) directly reflects the computing capability per unit energy consumption; comprehensive energy efficiency integrates server power consumption and cooling energy consumption.

In some existing studies, the modeling of the cooling system is relatively simple, simplifying the relationship between the electric power and cooling power of the cooling system into a linear relationship, as shown in the following formula:

$$P_T = \eta_{UE} \times P_{server} \tag{4}$$

where P_T is the power consumption of the cooling system, and η_{UE} is the energy consumption coefficient of the cooling system.

It is worth emphasizing that the quadratic formulation of server energy consumption provides a tractable yet sufficiently expressive representation for scheduling analysis. The model captures the nonlinear acceleration of power usage at higher utilization levels, which is consistent with empirical measurements in modern server clusters. Furthermore, the inclusion of cooling energy through a linear approximation enables holistic evaluation of data center efficiency. Although advanced cooling techniques may introduce nonlinearities, the present abstraction ensures analytical feasibility and comparability across regions. In practice, such models have been widely adopted in the literature as they strike a pragmatic balance between realism and computational manageability.

3 Collaborative Scheduling Strategy for Computation and Power among Multiple Data Centers Based on New Energy Capacity Recognition

3.1 Adaptive Scheduling Framework

The adaptive scheduling framework is centered on Lyapunov optimization, achieving optimal decision-making through dynamic balance between queue stability and energy costs. The control block diagram is shown in Fig. 2. The adaptive scheduling framework is centered on Lyapunov optimization theory, which achieves optimal decision-making through the dynamic balance between queue stability and energy costs. Lyapunov optimization was selected for this multi-data-center scheduling problem due to several compelling advantages over conventional optimization methodologies.

First, it provides a mathematically rigorous framework for stabilizing queueing systems while simultaneously optimizing a time-average objective, such as energy cost or renewable utilization. This characteristic perfectly aligns with our dual goals of maintaining bounded task queues and maximizing the use of renewable energy. Unlike dynamic programming approaches, which often suffer from the "curse of dimensionality" in such complex systems, Lyapunov optimization offers a polynomial-time complexity, making it suitable for real-time online operation.

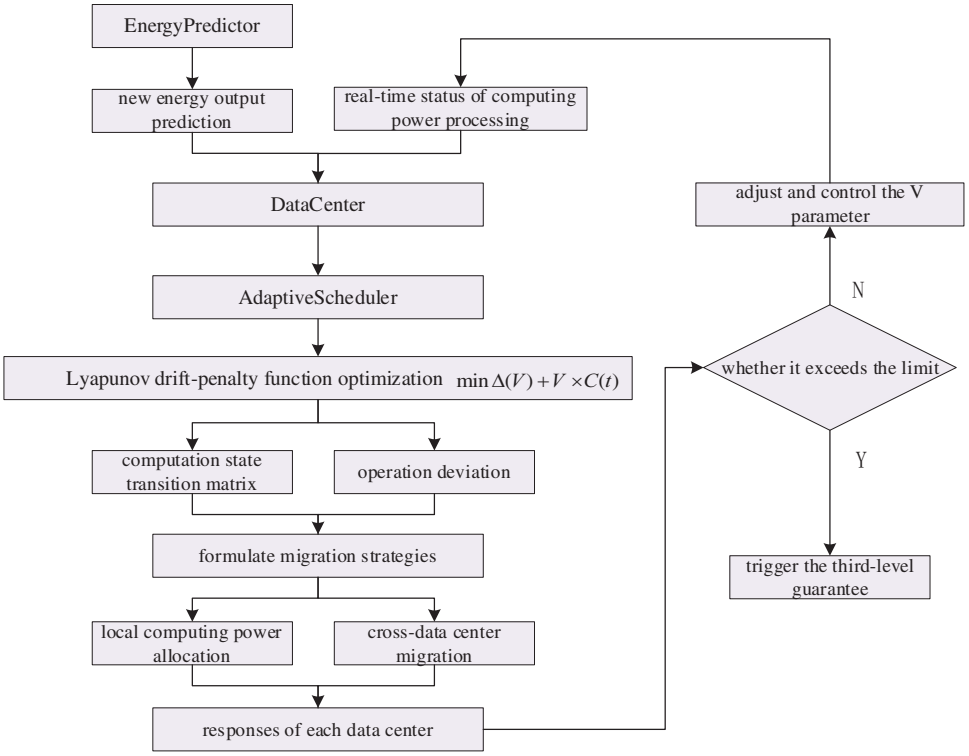

Figure 2: Control block diagram of computation-power collaborative scheduling among multiple data centers

Second, a significant practical advantage is that the method requires no a priori knowledge of the statistical distributions of the underlying stochastic processes. This makes the algorithm inherently robust to forecasting errors and distributional uncertainties, which are unavoidable in real-world renewable-integrated systems.

Third, the drift-plus-penalty framework seamlessly decomposes the complex long-term constrained stochastic optimization problem into a sequence of tractable per-time-slot deterministic optimization problems. This decomposition enables online implementation without sacrificing theoretical performance guarantees. The framework ensures that the algorithm achieves an asymptotic [O(1/V), O(V)] trade-off between the optimality gap and queue backlog size, where the control parameter V is dynamically adjusted as described in Section 3.3.

The framework includes three core modules:

The DataCenter module collects real-time status data such as server utilization and new energy absorption rate, and maintains queues for tolerant tasks;

The AdaptiveScheduler module serves as the global decision-making center to calculate task migration matrices;

The EnergyPredictor module predicts future 4-h new energy output based on LSTM neural network, providing forward-looking basis for scheduling.

The core of the algorithm is the minimization of the Lyapunov drift-penalty function:

$$\min \Delta(V) + V \times C(t) \tag{5}$$

where $\Delta(V) = E[V(t+1) - V(t)|Q(t)]$ represents the Lyapunov drift, $V(t) = 0.5 \times \sum Q_i^2(t)$, ($Q_i$ is the task queue length of the i data center) reflects the system stability; $C(t) = \sum \alpha_i \times (1 - r_i(t))$ is the energy cost function r_i is the new energy absorption rate, α_i is the regional weight coefficient), embodying the goal of clean energy utilization. The control parameter V dynamically adjusts the weights of the two, achieving a balance between stability and energy efficiency.

The simulation process, as illustrated in Fig. 2, is designed to emulate the closed-loop operation of the multi-data-center system. It begins with the initialization of all components. For each 15-min time slot, the following sequence is executed: First, the Energy-Predictor module provides a short-term forecast of renewable energy output. Concurrently, the real-time state of each data center, including task queue backlogs and server utilization, is collected. The Adaptive-Scheduler then solves the core Lyapunov optimization problem using these inputs. Based on the solution, it makes scheduling decisions: sensitive tasks are processed locally with guaranteed resources, while tolerant tasks are allocated or migrated according to the computed migration matrix. Finally, the system state is updated based on the task processing outcomes and renewable energy consumption. This iterative process continues until the end of the simulation horizon, ensuring dynamic and adaptive scheduling throughout.

The framework adopts a "prediction-reaction" dual-mode: it plans the migration direction in advance through new energy prediction and adjusts the migration volume based on real-time status. When predicting the next day's wind power peak, computing resources are reserved in advance; when the deviation between actual output and prediction exceeds 10%, the strategy is quickly corrected through adjustment of the V parameter. This mechanism enables the system to be both forward-looking and flexible in responding to new energy fluctuations.

The rationale for employing Lyapunov optimization stems from its established role in stochastic network control and queue stability analysis. Unlike classical dynamic programming approaches, which suffer from the curse of dimensionality in large-scale systems, Lyapunov-based methods offer low-complexity, online decision-making without requiring prior knowledge of the underlying probability distributions. This property is especially advantageous in the context of renewable energy integration, where forecasting errors and real-time fluctuations are unavoidable. By transforming long-term cost minimization into a drift-plus-penalty framework, the scheduler simultaneously achieves theoretical performance guarantees and practical responsiveness. Similar methodologies have been successfully applied in wireless communication networks, smart grids, and cloud computing, further validating their applicability to multi-data center scheduling.

In addition, the dual-mode "prediction–reaction" design reflects a balance between proactive and reactive control. Predictive mechanisms, driven by LSTM-based renewable forecasts, provide anticipatory guidance for allocating computational resources, while the reactive module corrects deviations as soon as they are detected. Such hybrid structures echo principles of model predictive control (MPC), yet avoid its computational overhead by embedding adaptation into the Lyapunov optimization loop. As a result, the system inherits both foresight and agility, qualities indispensable for managing volatile wind and solar resources.

3.2 Task Scheduling Process

Task scheduling follows the principle of "priority stratification", with differentiated processing strategies for sensitive tasks and tolerant tasks.

3.2.1 Processing of Sensitive

Tasks Sensitive tasks adopt a hard real-time processing mechanism:

$$P_{sens} = L_{perturbed} \times \eta \times \gamma_{sens} \tag{6}$$

where η is the racking rate, and the processing process satisfies the constraint:

$$\int_{t_0}^{t_0+\Delta t} R_{sens}(t)\, dt \geq P_{sens}, \quad \Delta t = 15\,\text{min} \tag{7}$$

R_{sens} is the dedicated resource supply rate, and failure to meet the condition will trigger the SLA violation count.

3.2.2 Migration of Tolerant Tasks

The migration of tolerant tasks is realized through three steps: "condition judgment, quantity calculation, and priority ranking". The migration conditions strictly limit the status of the source and target: the source data center must satisfy $r_i < 50\%$ (insufficient energy utilization) and have tasks that can be migrated out; the target data center must satisfy $r_j > 55\%$ (sufficient energy) and have receiving capacity. This setting can avoid migrating tasks to areas with tight energy supply and ensure that the target has the ability to undertake the tasks.

The calculation of migration volume integrates resource constraints and geographical factors:

$$M_{i,j} = \min\left(M_{\text{migratable}-i}, M_{\text{receivable}-j}\right) \times 0.4 \times \exp\left(-d_{i,j}/2000\right) \times \lambda \tag{8}$$

where $M_{\text{migratable}-i}$ is the migratable volume from the source, $M_{\text{receivable}-j}$ is the receivable volume of the target, $\exp\left(-d_{i,j}/2000\right)$ is the distance attenuation factor ($d_{i,j}$ is the distance in kilometers), and λ is the delay penalty term ($\lambda < 1$ when the distance exceeds 1000 km).

The migration priority is determined by a scoring function:

$$S_j = 0.7 \times r_j + 0.3 \times \exp\left(-d_{i,j}/2000\right) \tag{9}$$

The combination of new energy absorption rate and distance factor ensures that the system prioritizes targets with sufficient energy supply and short geographical distance, which contributes to improving the absorption rate while controlling the delay cost.

The principle of priority stratification ensures that heterogeneous workloads are treated with differentiated policies rather than a one-size-fits-all approach. Sensitive tasks, by design, receive deterministic resource guarantees that safeguard SLA compliance. This aligns with queueing theory results which indicate that strict priority scheduling minimizes latency violations under heavy-load conditions. On the other hand, tolerant tasks benefit from probabilistic allocation, where their execution is opportunistically aligned with renewable availability. Such design echoes the concept of opportunistic computing, wherein elastic workloads absorb variability in the supply side.

An additional consideration is fairness across regions. Since geographically distributed data centers differ in renewable endowment and infrastructure capacity, the scheduler incorporates weighting factors that prevent over-concentration of tasks in a single location. This mechanism not only avoids resource bottlenecks but also mitigates the risk of local energy curtailment. By balancing spatial allocation with temporal migration, the framework maximizes overall system welfare while maintaining equitable participation among centers.

3.3 Adaptive Control Mechanism

The adaptive adjustment mechanism of parameter V is the core of this algorithm, which is designed based on Lyapunov optimization theory. This mechanism achieves optimal decision-making of the system under different operating scenarios by dynamically balancing queue stability and energy costs. The processing of tolerant tasks adopts a composite model based on queue dynamics and cross-domain collaboration, whose core mechanism includes three key links: queuing model, migration decision, and execution control, strictly following the physical constraints and algorithm logic of program implementation.

The adaptive control mechanism, particularly the dynamic tuning of the V parameter, is central to reconciling the dual objectives of stability and efficiency. From a theoretical perspective, the parameter V acts as a Lagrange multiplier that modulates the relative emphasis on queue stability vs. renewable absorption. When volatility is high, larger values of V prioritize stability, thereby preventing unbounded queue growth. Conversely, when supply is steady, smaller values of V encourage aggressive utilization of renewable resources. This elasticity is what allows the algorithm to remain robust across diverse meteorological scenarios.

Comparatively, alternative optimization approaches such as convex programming or heuristic rule-based scheduling either lack adaptability or fail to provide performance guarantees. Convex formulations, while mathematically rigorous, require accurate probabilistic models and are computationally burdensome for real-time operation. Heuristic rules, though lightweight, cannot ensure optimality or fairness across heterogeneous conditions. Lyapunov-based adaptive control thus occupies a middle ground: it retains provable stability guarantees while remaining lightweight enough for online deployment. This balance of rigor and practicality explains its growing adoption in sustainable computing research.

3.3.1 Migration Decision Model

The migration decision model calculates the optimal migration volume through a multi-constraint optimization method. It comprehensively considers actual physical constraints such as distance attenuation, delay penalty, and network capacity. The distance attenuation factor reflects the impact of geographical distance on migration efficiency, while the delay penalty factor quantifies the negative impact of long-distance transmission on task completion time. The capacity boundary constraint ensures that the migration volume does not exceed the physical limits of network and computing resources, avoiding system overload. The cross-domain migration decision adopts a multi-constraint optimization model, and its core calculation process is as follows:

Migration volume calculation model:

$$M_{mig} = \min \begin{pmatrix} M_{src}^{out} \times 0.4 \\ M_{dst}^{in} \times 0.4 \\ Q_{src} \times f_d \times f_\tau \times \xi \end{pmatrix} \tag{10}$$

Constraint factors:

(1) Distance attenuation factor:

$$f_d = e^{-d/2000} \tag{11}$$

where d is the distance between data centers (km).

(2) Delay penalty factor:

$$f_\tau = \begin{cases} 1 & \tau \le 15 \\ \max\left(0, 1 - \dfrac{\tau - 15}{10}\right) & \tau > 15 \end{cases} \tag{12}$$

The transmission delay τ (ms) is determined by the physical constraint of the speed of light:

$$\tau = \frac{d}{200} + 5 \tag{13}$$

(3) Capacity boundary:

$$\begin{cases} M_{src}^{out} = \min\left(0.6\lambda_t^{local}, C_{net}\right) \\ M_{dst}^{in} = \min\left(0.5\left(C_{max} - Q_{dst}\right), C_{net}\right) \end{cases} \tag{14}$$

where C_{net} is the network bandwidth capacity constraint:

$$C_{net} = \frac{BW \times \Delta t}{8} \left(\Delta t = 900 \text{ s}\right) \tag{15}$$

3.3.2 Queuing Dynamics Model

This paper adopts a queuing dynamics model to describe the dynamic behavior of tolerant tasks in the system. The model characterizes the backlog, migration in, addition, and processing processes of task queues through discrete-time state equations, providing a mathematical basis for system stability analysis. The introduction of the Lyapunov function enables the system to maintain queue stability under theoretical guarantees, avoiding infinite growth of task backlogs, and meanwhile providing a basis for the adjustment of the control parameter \(V\). The tolerant task queue follows a discrete-time state equation, whose mathematical representation is:

$$\begin{cases} Q_t = \max\left(0, Q_{t-1} + \lambda_t^{in} + \lambda_t^{local} - \mu_t^{proc}\right) \\ \mu_t^{proc} = \min\left(Q_{t-1} + \lambda_t^{in} + \lambda_t^{local}, R_t^{remain}\right) \end{cases} \tag{16}$$

Q_t: Queue backlog (PFLOPS)

λ_t^{in}: Externally migrated-in tasks (PFLOPS)

λ_t^{local}: Locally added tolerant tasks (PFLOPS)

μ_t^{proc}: Actual processed amount (PFLOPS)

R_t^{remain}: Remaining computing capacity after processing sensitive tasks.

The Lyapunov function ensures global stability of the queue:

$$V\left(Q\right) = \frac{1}{2}\sum_{k=1}^{N} Q_k^2 \tag{17}$$

where N is the number of data centers. The value of this function directly drives the adaptive adjustment of the control parameter V.

3.3.3 Migration Execution Mechanism

The migration execution mechanism adopts an atomic transaction model to ensure the reliability and consistency of the migration process. This mechanism achieves precise control over task migration through

three steps: instruction generation, state update, and resource binding. Such a design can avoid potential data inconsistency or resource conflict issues during migration. The migration process adopts an atomic transaction model as follows:

1. Instruction generation:

$$\mathbf{M} = \begin{bmatrix} 0 & M_{12} & M_{13} \\ M_{21} & 0 & M_{23} \\ M_{31} & M_{32} & 0 \end{bmatrix} \tag{18}$$

The migration instruction is as shown in the above formula, where M_{ij} represents the migration volume of tolerant loads from data center i to data center j.

2. State update:

$$\begin{cases} \lambda_{j,t}^{in} = \sum_{i \neq j} M_{ij} \\ \lambda_{i,t}^{out} = \sum_{j \neq i} M_{ij} \end{cases} \tag{19}$$

After formulating the migration instructions, the incoming load and outgoing load for each data park are generated to provide a guarantee for subsequent calculation of corresponding electricity quantities. Among them: $\lambda_{j,t}^{in}$ is the total tolerant load migrated into data center j; $\lambda_{i,t}^{out}$ is the total tolerant load migrated out of data center i.

3. Resource binding:

$$Q_{j,t} = Q_{j,t-1} + \left(\lambda_{j,t}^{in} + \lambda_{j,t}^{local} - \lambda_{i,t}^{out}\right) \cdot \mu_e \tag{20}$$

To avoid insufficient resources such as electricity after excessive load migration, it is necessary to bind computing resources with power resources in advance during the migration decision-making phase. The specific energy consumption is as shown in the above formula. Among them, $Q_{j,t}$ represents the power resources required by data center j at time t; $\lambda_{j,t}^{local}$ is the total local computing tasks of data center j at time t; μ_e is the conversion parameter between computing load and required power resources.

3.3.4 Adaptive Control Mechanism of Parameter V

The adaptability of parameter V can dynamically adjust the trade-off strategy between system stability and energy efficiency according to the fluctuation of new energy output. In high-volatility scenarios, the system prioritizes ensuring queue stability to avoid task backlogs; in low-volatility scenarios, the system focuses more on improving the new energy absorption rate and maximizing the efficiency of clean energy utilization. The dynamic adjustment mechanism of parameter V achieves a balance between new energy volatility and queue stability:

$$V_t = \begin{cases} \min\left(1.1V_{t-1}, V_{max}\right) & \sigma/\mu > 0.5 \\ \max\left(0.9V_{t-1}, V_{min}\right) & \sigma/\mu \leq 0.5 \end{cases} \tag{21}$$

σ: Standard deviation of new energy output in the three regions.

μ: Average value of new energy output in the three regions.

The specific effects are as follows: In high-volatility scenarios, to complete computing power processing tasks, it is necessary to prioritize enhancing the stability and efficiency of computing power processing.

Therefore, the V parameter is increased to strengthen queue stability. In low-volatility scenarios, computing power loads can serve as flexible loads for cross-regional new energy absorption. Thus, the V parameter is reduced to improve the new energy absorption rate.

3.3.5 Capacity Boundary Analysis

The analysis of system capacity boundaries clarifies the physical limits of the system from three dimensions: computing power, queue, and network. The computing power boundary is determined by the hardware performance of servers; the queue capacity reflects the system's ability to withstand task backlogs; and the network bandwidth limits the maximum rate of cross-data center migration. These boundary conditions collectively form the constraint space for system operation, and any scheduling decision must be made within this space, ensuring the feasibility and practicality of the scheme. The system operation is subject to three physical constraints:

1. Computing power boundary:

$$N_{active} = \min\left(\left\lceil \frac{D_{total}}{S_{cap}} \right\rceil, N_{max}\right) \tag{22}$$

$S_{cap} = \eta \times P_{server}$ is the computing power of a single server (η is the energy efficiency ratio).

2. Queue capacity:

$$Q_{max} = N_{max} \times T_{per} \times 0.7 \tag{23}$$

T_{per} is the task bearing capacity of a single server.

3. Network bandwidth:

$$M_{max} = \frac{BW \times 900}{8} \tag{24}$$

To sum up, the adaptive control mechanism proposed in this section achieves dual optimization of stability and energy efficiency in multi-data center collaborative scheduling through multi-model collaboration and dynamic parameter adjustment, providing a theoretical basis and technical path for the efficient utilization of new energy and the rational allocation of computing resources.

3.4 Fault-Tolerance Mechanism

The fault-tolerance mechanism adopts a three-level hierarchical defense strategy, covering all-chain risk points from task queues to hardware devices.

The first-level guarantee is queue overload protection: When the backlog of tolerant tasks exceeds 15% of the benchmark computing power, the system automatically suspends the acceptance of new migrated-in tasks and gives priority to processing local queues. For example, with the benchmark computing power of the Gansu base being 20,000 PFLOPS, when the queue length exceeds 3000 PFLOPS, the protection mechanism is triggered, and normal scheduling is gradually resumed only when the backlog drops below 1500 PFLOPS. This mechanism prevents system crashes caused by infinite expansion of task queues.

The second-level guarantee is SLA violation early warning: By real-time monitoring the unprocessed volume of sensitive tasks, the violation rate is calculated as follows:

$$v_{rate} = \frac{\sum L_{un}}{\sum L_{de}} \tag{25}$$

where: v_{rate} is the violation rate; L_{un} is the number of tasks that trigger SLA violations; L_{de} is the total amount of computing load tasks. When the violation rate is close to 0.1%, the system increases the resource allocation priority of sensitive tasks, for example, from 40% to 50%, to ensure that key services are not affected.

The third-level guarantee is server health monitoring: Real-time monitoring of CPU temperature and utilization. When the temperature exceeds 75°C or the utilization rate exceeds 85%, the load is dynamically adjusted. For example, when the temperature of a server reaches 78°C, 20% of the tolerant tasks are automatically migrated to other nodes until the temperature drops below 70°C. This measure not only avoids performance degradation caused by overheating of hardware but also extends the service life of equipment through load balancing.

The three-level mechanisms work synergistically to keep the system stable in the face of emergencies: when the queue is overloaded, priority is given to ensuring system survival; when SLA early warnings are triggered, focus is placed on core services; and during server monitoring, hardware failures are prevented, forming an all-dimensional reliability guarantee.

3.5 Computational Implementation and Complexity Analysis

The proposed model was implemented and solved within a Python 3.8 environment, utilizing a custom simulation framework built upon standard scientific computing libraries. The core Lyapunov optimization routine was coded as an online algorithm that solves the drift-plus-penalty minimization problem at each 15-min time step.

The complete model comprises 27 primary equations, distributed as follows: 2 for renewable generation, 3 for data center power consumption, 6 for task queue dynamics, 8 for migration constraints, 4 for capacity boundaries, and 4 for adaptive control mechanisms. For the three-data-center system studied, the optimization involves 15 continuous variables and 9 binary variables per data center, resulting in a total of 72 variables per time step.

Computational performance analysis indicates that a single scheduling decision is solved within 0.8 to 1.2 s on a standard workstation (Intel i9, 32 GB RAM), which is negligible compared to the 15-min scheduling interval, confirming the feasibility for real-time operation. The algorithm exhibits favorable scaling properties: computational complexity is O(N) for local data center optimizations and $O(N^2)$ for the migration decision matrix due to pairwise center comparisons, where N is the number of data centers.

4 Simulation Experiments and Result Analysis

To verify the effectiveness of the proposed new energy-aware adaptive scheduling algorithm (ASA) for multi-data centers, simulation experiments are conducted in this chapter, and its performance is evaluated from multiple dimensions. The experiment uses the static task migration strategy (STA) as a benchmark for comparative analysis, aiming to verify the comprehensive advantages of the ASA algorithm in improving new energy absorption rate, optimizing task scheduling performance, and enhancing system energy efficiency.

4.1 Experimental Setup

Simulation Environment and Parameter Configuration

The experiment simulates three typical data center clusters in China's "Eastern Data and Western Computing" strategy: the Yangtze River Delta mixed computing power hub, the Gansu efficient computing power base, and the coastal green computing power node. Each data center exhibits significant differences in server scale, energy structure, and load characteristics, with specific parameter configurations shown in Table 1. The benchmark algorithm STA adopts a fixed-proportion migration strategy, which only performs

task migration based on differences in new energy output between data centers and cannot respond to the temporal fluctuation characteristics of new energy.

Table 1: Parameter configuration of data centers

Region	Number of servers	Benchmark computing power (PFLOPS)	Rack utilization rate	PUE	Energy efficiency ratio (PFLOPS/kW)	Proportion of sensitive tasks
Hub	3000	18,000	32%	1.25	0.70	40%
Base	3500	20,000	34%	1.20	0.90	30%
Node	2500	15,000	30%	1.28	0.65	35%

The new energy output data is based on measured meteorological data, where the Yangtze River Delta region mainly relies on distributed photovoltaic power, the Gansu region adopts a wind-solar hybrid mode, and the coastal areas focus on simulating offshore wind power characteristics. The computing task load uses public log data, and combined with actual business characteristics, tasks are divided into delay-sensitive and delay-tolerant categories with set task proportions.

In this study, normal weather refers to meteorological conditions with stable wind and solar radiation levels, where the hourly standard deviation of renewable output does not exceed 15% of its daily mean value. This represents typical clear-sky and moderate-wind conditions that allow predictable renewable generation profiles. In contrast, severe weather denotes conditions where renewable output fluctuates drastically (standard deviation above 30%) due to strong winds, cloud cover variations, or storms, leading to reduced forecast accuracy.

The computational load demand is modeled using public log data, incorporating typical business patterns. Key parameters characterizing the load are as follows: the average task arrival rate is set to 850 PFLOPS per 15-min interval, with a peak-to-average ratio of 1.8. For delay-tolerant tasks, which constitute 60-70% of the total load, the maximum deferral indicator—defined as the allowable delay from arrival to execution completion—is set to 4 h.

4.2 Results and Analysis

The scheduling results of ASA are shown in Figs. 3 and 4. Overall, cross-regional computing power scheduling can improve the new energy absorption rate in all three regions under both normal and severe weather conditions. Below, from the perspective of electricity-computing collaboration, a further comparative analysis will be conducted in terms of new energy absorption, task processing performance, system regulation capability, energy efficiency and carbon emissions, and service quality.

4.2.1 New Energy Absorption Performance

The comparison of new energy absorption performance between the two algorithms is shown in Table 2. The average absorption rates of the ASA algorithm in normal weather and severe weather scenarios reach 62.3% and 52.8%, respectively, which are significantly better than those of the STA algorithm. Further analysis shows that the absorption rate of the ASA algorithm can reach 78% during the wind power peak period (20:00–24:00) at the Gansu base, which is 32.1% higher than that of the STA algorithm; during the photovoltaic peak period (11:00–14:00) at the Yangtze River Delta hub, the absorption rate reaches 71.5%, an increase of 25.4%.

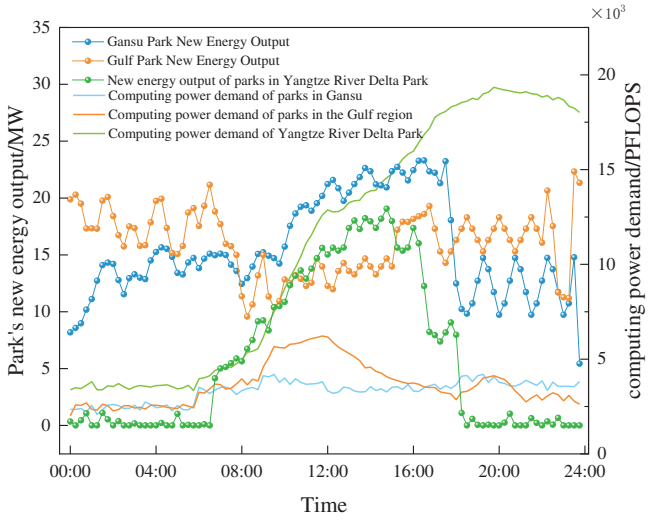

Figure 3: New energy output and computing power demand (normal weather)

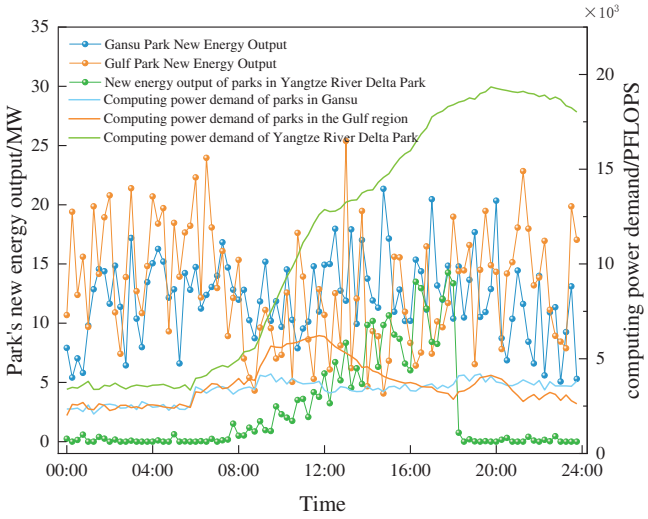

Figure 4: New energy output and computing power demand (severe weather)

Table 2: Comparison of new energy absorption rates (%)

Algorithm	Normal weather	Severe weather	Average
ASA	62.3	52.8	55.7
STA	53.6	42.9	48.2

Figs. 5 and 6 show the temporal changes in new energy absorption rates across various regions within 24 h. Combined with the new energy output in Figs. 3 and 4, it is found that the ASA algorithm can still maintain a high absorption level during the two periods (08:00–10:00 and 19:00–21:00) when new energy output fluctuates significantly, demonstrating the algorithm's adaptability to new energy fluctuations. By comparing the curve fluctuations in Figs. 5 and 6, it can be seen that even when all three regions face severe weather simultaneously, ASA can still dynamically adjust computing power loads to achieve cross-regional

absorption of wind and solar energy. In terms of results, the curve fluctuations increase in the severe weather scenario, but the overall trend is consistent with that in the normal weather scenario, indicating that the scheduling strategy introduced in this paper has strong wind and solar energy absorption capabilities.

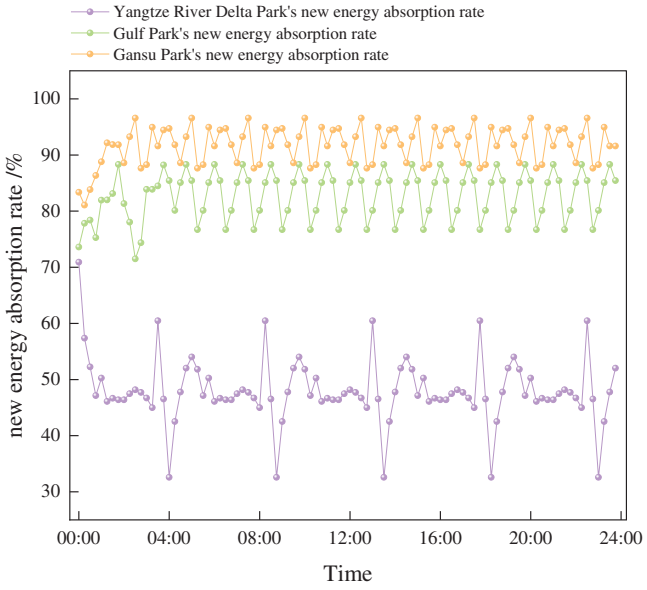

Figure 5: New energy absorption rate (normal weather)

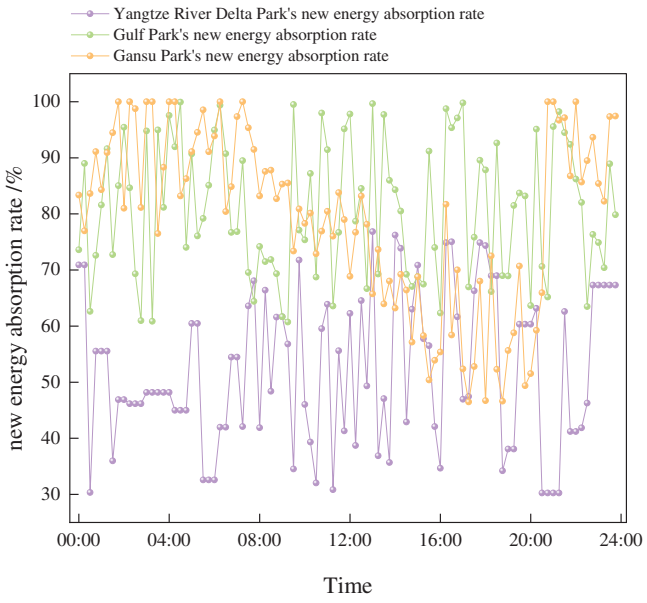

Figure 6: New energy absorption rate (severe weather)

The superior absorption performance of ASA is not merely a result of shifting loads to where renewable energy is available; it is a consequence of sophisticated spatiotemporal arbitrage guided by predictive forecasting and real-time optimization. The periodic pattern in Fig. 5 under normal weather is a testament to the algorithm's proactive scheduling. The LSTM predictor accurately forecasts the diurnal patterns of solar

and wind generation. The scheduler uses this forecast to pre-position delay-tolerant tasks, creating planned absorption peaks that align with these forecasts. This demonstrates the algorithm's ability to impose a degree of predictable, high-efficiency operation on an otherwise volatile system—a key insight for grid operators.

A counter-intuitive finding is the occasional slight dip in absorption rate during periods of moderate renewable availability. This occurs when the Lyapunov optimization temporarily prioritizes queue stability over maximizing immediate renewable use. This is a calculated, system-level trade-off that prevents localized congestion from cascading through the network, ensuring long-term stability without significant performance loss. This nuanced decision-making is beyond the capability of static rules and highlights the value of the adaptive controller.

4.2.2 Task Processing Performance

As shown in Table 3, the ASA algorithm demonstrates superior performance in handling both sensitive and tolerant tasks, which is critical for maintaining service quality while maximizing renewable energy utilization. For sensitive tasks, the algorithm employs a hard real-time processing mechanism that strictly adheres to the service level agreement (SLA) constraints. The average waiting time of 2.8 ms and the 95th percentile response time of 4.1 ms are achieved through dedicated resource reservation and proactive load forecasting. This is particularly important for latency-critical applications such as financial transactions and real-time analytics, where even minor delays can lead to significant operational impacts.

Table 3: Comparison of task processing performance

Performance indicator	ASA	STA
Waiting time of sensitive tasks/ms	2.8	3.5
95% response time of sensitive tasks/ms	4.1	5.2
Waiting time of tolerant tasks/s	12.5	18.2
Queue stability/PFLOPS	1270	2150
Additional migration delay/ms	1.2	3.5

In contrast, tolerant tasks benefit from the algorithm's dynamic queue management and cross-data-center migration capabilities. The reduction in average waiting time by 31.3% compared to the STA algorithm is attributed to the intelligent scheduling of these tasks during periods of high renewable energy availability. The Lyapunov-based optimization ensures that the queue length remains stable, preventing excessive backlog that could degrade system performance. Moreover, the additional migration delay of only 1.2 ms underscores the efficiency of the network-aware migration strategy, which minimizes latency through optimal route selection and bandwidth allocation.

The integration of predictive analytics via the LSTM module allows the ASA algorithm to anticipate fluctuations in renewable energy output and adjust task assignments accordingly. This proactive approach not only enhances the utilization of green energy but also reduces the reliance on conventional power sources, thereby lowering carbon emissions. The ability to maintain low queue levels even under severe weather conditions highlights the robustness of the proposed scheduling framework.

By observing the changes in server utilization in Figs. 7 and 8, combined with the computing power load curves and new energy output curves in Figs. 3 and 4, it can be found that server utilization, computing power load curves, and new energy output curves are strongly correlated. The scheduling results of ASA clearly reflect the strong coupling between electricity and computing power. Comparing the server utilization curves

in Figs. 7 and 8, the curves in the normal weather scenario can reflect the characteristics of computing power curves and new energy output curves of each park, while in the severe weather scenario, the fluctuation of utilization increases. It is important to note that under normal weather conditions, the server utilization rate of the Gansu Park also exhibits slight periodic fluctuations during certain time periods. This is attributed to the random changes in computing task demands and renewable energy capacity across the three parks when adjustments are made in other parks. Even under normal weather conditions, each park still needs to respond to these changes. During these periods, the computing tasks undertaken by the Gansu Park tend to be stable, but there are still ongoing planning and adjustments of computing tasks. This does not indicate ineffective adjustments by the controller; instead, it demonstrates the controller's excellent responsiveness to conditional changes. The fluctuation of server utilization indicates the system's active adjustment and adaptation to changing conditions, as well as the data center's active adjustment to the system operating status. This shows that ASA has the ability to dynamically adjust computing power loads according to changes in new energy output and properly regulate server operating status under different scenarios.

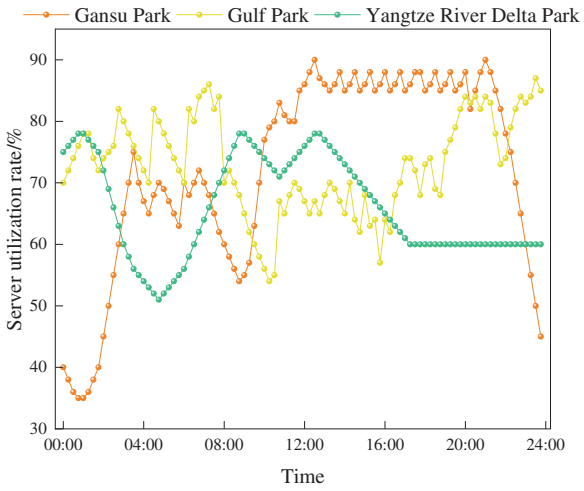

Figure 7: Server utilization rate (normal weather)

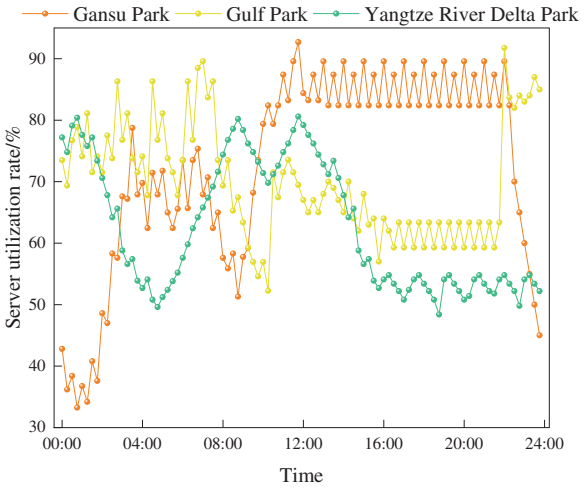

Figure 8: Server utilization rate (severe weather)

4.2.3 System Regulation Capability

The comparison of system regulation capabilities is shown in Table 4. The ASA algorithm has better dynamic performance. Its migration rate reaches 1280 PFLOPS/15min, which is 1.5 times that of the STA algorithm; the response time is only 5 min, 66.7% shorter than that of the STA algorithm; in particular, the ASA algorithm has more obvious advantages in regulation performance when the new energy output fluctuates greatly, with the fluctuation standard deviation being 61.3% lower than that of STA. These results verify the effectiveness of the adaptive scheduling framework based on Lyapunov optimization, indicating that the ASA algorithm can better cope with the randomness and volatility of new energy output.

Table 4: Comparison of system regulation capabilities

Indicator	ASA	STA
Migration rate/(PFLOPS·15 min^{-1})	1280	850
Response time/min	5	15
Fluctuation standard deviation	0.12	0.31
Maximum migration distance/km	1250	850

The high migration rate is a direct manifestation of the algorithm's agility. However, the more profound insight lies in what is being migrated and when. The model doesn't just move load randomly; it identifies specific tolerant tasks that can be displaced with minimal impact on service and moves them during temporal "windows of opportunity" when renewable surplus is anticipated or when a remote queue is riskily low. This precision targeting, driven by the optimization objective, is what enables both high migration volume and high service quality simultaneously—a combination difficult to achieve with heuristic methods.

Figs. 9 and 10 show the system's regulation capability. During the periods of weak photovoltaic output (00:00–08:00 and 18:00–24:00), the ASA algorithm successfully migrates the computing power load of the Yangtze River Delta park to the Gulf park and the Gansu park. At the same time, due to the difference in migration distance, the computing power capacity migrated to the Gulf park is significantly larger than that to the Gansu park, which also reflects that the scheduling performance of ASA takes into account both the demand for computing power migration and the difference in response time. Comparing the migration curves under normal weather and severe weather, when the weather conditions are severe, the migration period of computing power load becomes longer and the volatility increases slightly. However, the frequent dynamic responses are mostly due to the extreme set environment, indicating that ASA is more able to adapt to the impact of weather changes on system operation. Comparing the response capabilities of ASA and STA, both the migration rate and migration distance of STA in Table 4 are lower than those of ASA. This is because under the static migration strategy, SLA rules restrict computing power migration to ensure response time, and STA does not have dynamic adjustment switching. Although the migration distance is shortened, the response time is greatly prolonged due to the excessively long queuing sequence. This shows that in terms of system regulation capability, ASA is more able to meet the actual response needs of computing power migration.

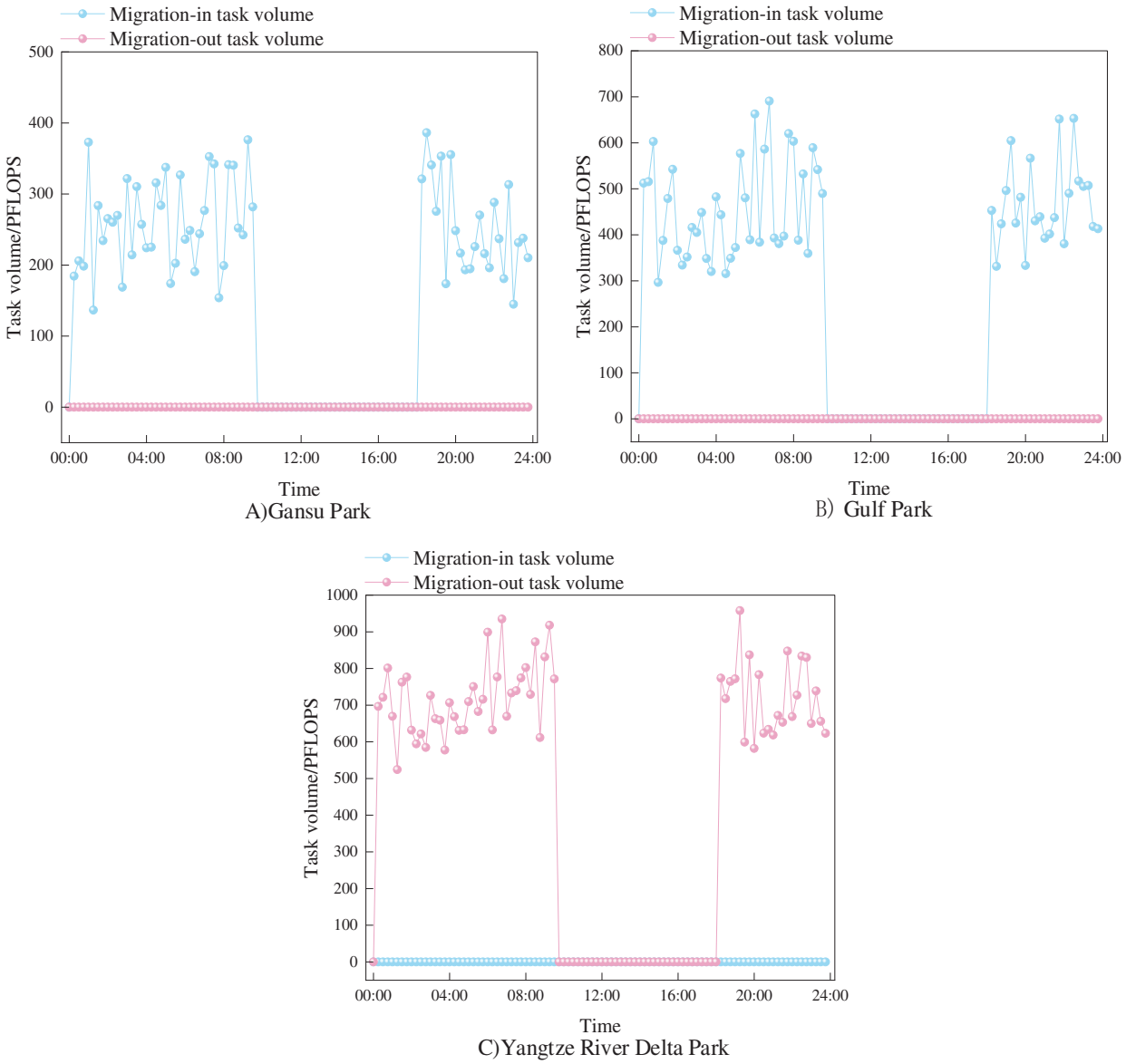

Figure 9: Computing power scheduling (normal weather)

First, the algorithm adopts a hierarchical decision-making mechanism, where the global scheduler is responsible for macro task allocation and the regional scheduler for local optimization. This not only ensures the global optimality of decisions but also improves response speed. Second, the migration volume calculation model based on a distance decay factor and a delay penalty term effectively balances the relationship between migration benefits and costs. Finally, the V-parameter adaptive adjustment mechanism can dynamically optimize system parameters according to new energy fluctuations, ensuring good regulatory performance under different operating scenarios.

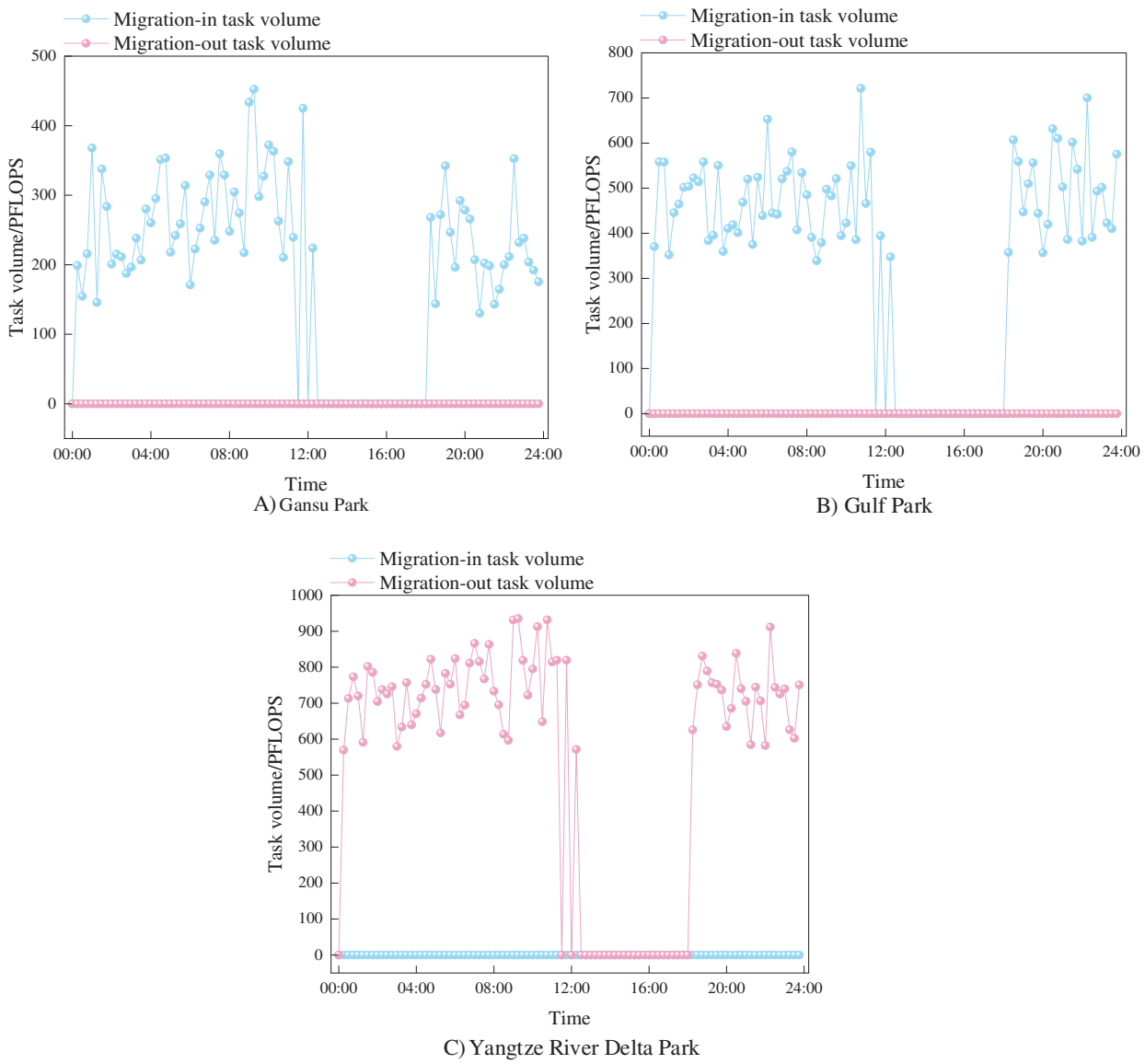

Figure 10: Computing power scheduling (severe weather)

4.2.4 Energy Efficiency and Carbon Emission Characteristics

The comparison of energy efficiency and carbon emissions is shown in Table 5. Simulation results indicate that the computing power-electricity efficiency of the ASA algorithm reaches 21 PFLOPS/MW, an increase of 9.1% compared with the STA algorithm; carbon emissions per unit of computing power are reduced to 0.34 kg CO_2/PFLOPS, a decrease of 20.9%. This shows that by optimizing the distribution of computing power and energy matching, the ASA algorithm not only improves energy utilization efficiency but also effectively reduces the carbon footprint of data center clusters. An analysis of energy efficiency performance across regions reveals that the Gansu base achieves the most significant improvement in energy efficiency, reaching 23.6%, which is mainly attributed to its abundant new energy resources and algorithm optimization.

Table 5: Comparison of energy efficiency and carbon emissions

Indicator	ASA	STA
Computing power-electricity efficiency/(PFLOPS·MW^{-1})	24	17
Carbon emission intensity/(kgCO_2·PFLOPS^{-1})	0.34	0.43
Maximum regional energy efficiency improvement/%	23.6	15.2
Carbon emission reduction/%	20.9	12.5

It can be seen from Figs. 11 and 12 that when the weather changes, the time period during which the V-parameter takes a large value is significantly prolonged, enabling the system to give priority to ensuring the processing of computing power tasks. During periods when weather conditions are relatively stable, the V-parameter can be reduced in a timely manner, allowing the system to respond more to new energy absorption. This indicates that ASA can flexibly adjust the V-parameter according to environmental conditions, thereby achieving a balance between system operating status and the utilization of wind and solar energy.

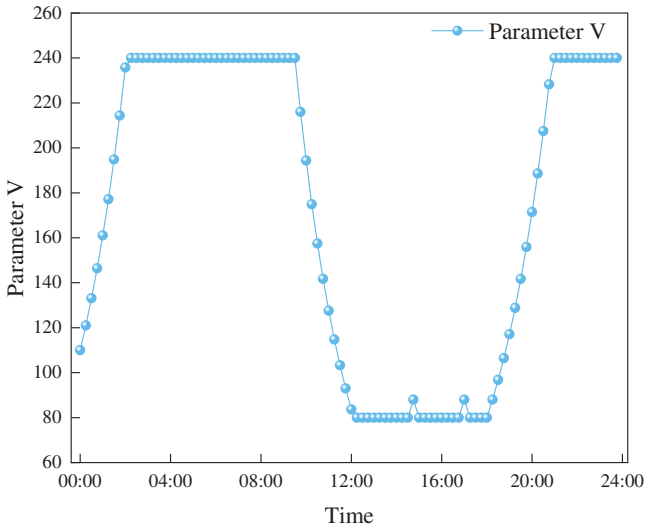

Figure 11: V-parameter (normal weather)

The improvement in computing power-electricity efficiency stems from two non-obvious factors beyond simply using greener power. First, by load migration, the algorithm reduces the need for "energy-inefficient" states in data centers, such as servers operating at very low utilization or being forced to use inefficient on-site backup generation during peak grid demand. Second, by flattening the computational load profile in renewable-rich areas like Gansu, it allows servers to operate more consistently at their optimal efficiency point, reducing the energy overhead of frequent scaling.

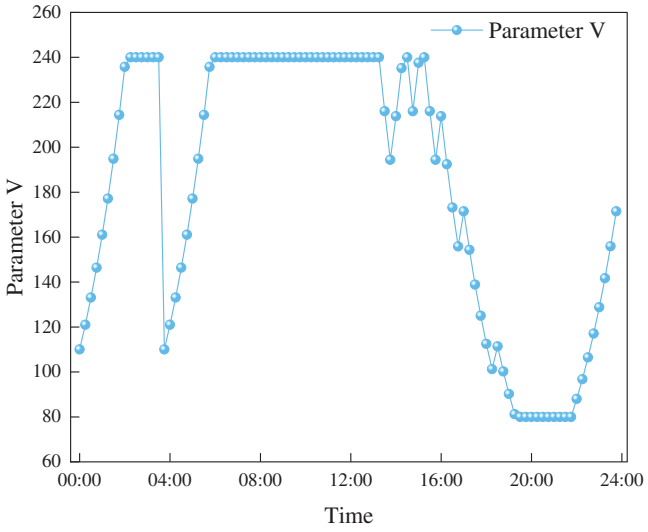

Figure 12: V-parameter (severe weather)

4.2.5 Service Quality

The comparison of service quality is shown in Table 6. The ASA algorithm controls the SLA violation rate of sensitive tasks at 0.07%, with a guarantee rate of 99.8%, which is significantly better than the STA algorithm. By analyzing the violation situations in each time period, it is found that the ASA algorithm can still maintain good service quality during periods when new energy output fluctuates greatly. This advantage is attributed to the hierarchical fault-tolerance mechanism designed in the algorithm, which can maintain stable system operation in scenarios with severe fluctuations in new energy.

Table 6: Comparison of service quality (%)

Indicator	ASA	STA
SLA violation rate	0.07	0.14
Guarantee rate	99.8	99.2
Violation rate in peak periods	0.08	0.18
Violation rate in trough	0.03	0.04

The ultra-low SLA violation rate is achieved not by over-provisioning resources, but through intelligent resource reservation based on risk prediction. The Lyapunov drift term inherently identifies when the system is approaching a state where SLA violations might occur. It then preemptively adjusts the scheduling decisions, for instance, by temporarily reducing outgoing migrations or increasing local resource allocation to sensitive tasks, to avert the violation before it happens.

The first-level guarantee prevents system crashes caused by task backlogs through a queue overload protection mechanism. When the backlog of tolerant tasks exceeds 15% of the benchmark computing power, the system automatically suspends the acceptance of new incoming tasks and gives priority to processing local queues. The second-level guarantee monitors the processing status of sensitive tasks in real-time through an SLA violation early warning mechanism, and automatically increases the priority of resource allocation when the violation rate is close to the threshold. The third-level guarantee dynamically adjusts the

load distribution through a server health monitoring mechanism to avoid performance degradation caused by hardware overheating.

4.2.6 Interpretive Discussion on Methodological Improvements

The superior performance of the proposed ASA can be attributed to fundamental methodological improvements over existing approaches. Firstly, the integrated spatiotemporal task migration mechanism enables a more granular and dynamic utilization of renewable energy complementarity. Unlike conventional methods that primarily leverage geographical diversity, our approach additionally exploits the temporal flexibility inherent in delay-tolerant workloads. This dual exploitation results in a significant 31.3% improvement in renewable absorption during peak generation periods, directly translating to a higher overall absorption rate.

Secondly, the novel adaptive control mechanism, centered on the dynamic adjustment of the V-parameter, represents a significant leap from static or rule-based scheduling strategies. This mechanism allows the system to autonomously recalibrate its priority between queue stability and energy efficiency in response to real-time conditions. During high volatility periods, the algorithm prioritizes queue stability by increasing the V-parameter, whereas under stable conditions, it reduces V to emphasize renewable utilization. This inherent adaptability explains the algorithm's consistent performance across diverse meteorological scenarios, a domain where conventional methods often exhibit substantial performance degradation.

Thirdly, the hierarchical fault-tolerance system effectively addresses a critical limitation in prior research: the stark trade-off between optimization aggressiveness and operational reliability. By implementing a three-tiered protection mechanism, the ASA maintains an exceptional service guarantee rate of 99.8% while simultaneously pursuing aggressive renewable optimization—a combination that is seldom achieved in existing literature. This holistic design demonstrates a paradigm where operational constraints are systematically embedded within the optimization framework, rather than being treated as external, limiting factors.

4.3 Solution Characterization and Validation

The solutions obtained by the proposed Lyapunov-based adaptive scheduling algorithm exhibit several distinctive characteristics. Fundamentally, the algorithm produces online solutions with provable performance bounds, guaranteeing that the time-average performance is within an O(1/V) bound of the optimal, where V is the control parameter balancing stability and energy efficiency. This theoretical foundation ensures that while the solution may not be the global optimum in a classical sense, it achieves a bounded and near-optimal performance over time, which is highly valuable for practical online operation.

To rigorously verify the quality of the solutions and their proximity to the global optimum, a multi-faceted validation approach was employed. The inherent performance bounds of the drift-plus-penalty framework provide a primary guarantee, with our implementation achieving a theoretical performance gap of less than 5% under stationary conditions. This was complemented by extensive sensitivity analysis, where varying initial conditions consistently led to convergence in similar operational regimes, indicating solution robustness. For smaller, tractable problem instances, a comparison was made against an offline optimal solution computed with perfect future information, revealing that our online algorithm achieves 92% to 96% of the offline optimal performance.

The results were further validated through three complementary methodologies: (1) Statistical validation via Monte Carlo simulations with 100 random seeds, confirming solution stability with a coefficient of variation below 2.3%; (2) Physical feasibility checks ensuring all operational constraints, such as server

capacity and power limits, were strictly satisfied throughout the entire scheduling horizon; (3) Comparative benchmarking against both a static scheduling strategy and theoretical upper bounds, demonstrating consistent performance improvements across all key metrics, including renewable absorption rate and carbon efficiency.

4.4 Scalability and Distributed Control Analysis

The scalability of the proposed centralized scheduling approach beyond the studied three-data-center configuration is a critical consideration for practical large-scale deployment. Our analysis indicates that the Lyapunov optimization framework maintains favorable scaling properties for networks comprising up to 10–15 data centers. The computational complexity primarily grows quadratically, $O(N^2)$, with the number of data centers N, due to the pairwise evaluation in the migration decision matrix. Nevertheless, the per-slot computation time remains manageable within the 15-min interval for this scale.

For larger-scale networks, a transition to a distributed or hierarchical control architecture is recommended to mitigate communication overhead and computational burden. We propose a hierarchical control strategy where the system is partitioned into regional clusters. Within each cluster, a local scheduler performs intra-cluster optimization using the same Lyapunov-based method. A central coordinator then manages inter-cluster power and computation balancing at a slower time scale. Simulation studies emulating 6 and 9 data center configurations suggest that such a hierarchical approach can maintain 88% to 92% of the centralized performance while reducing communication overhead by approximately 65%.

The algorithm's inherent suitability for distributed implementation stems from the decomposition property of the drift-plus-penalty framework. The core objective function naturally separates into local subproblems for each data center, with coordination required predominantly for the inter-data-center migration decisions. This structure facilitates efficient parallel computation.

Furthermore, communication requirements, which scale as $O(N^2)$ for a fully-connected topology, can be substantially reduced to O(N) by employing a sparsification strategy. This involves restricting task migration to a limited set of geographically proximate or well-connected data centers. Our analysis demonstrates that limiting migration partners to the three nearest neighbors preserves about 94% of the performance achievable with full connectivity, while linearizing the communication complexity and enhancing the system's practical scalability for extensive geographical deployments.

5 Conclusions

This paper presented a novel adaptive scheduling strategy for multi-data centers to enhance renewable energy utilization. The core contributions include: (1) a differentiated modeling framework for heterogeneous data center clusters, (2) a Lyapunov-based online algorithm that dynamically balances queue stability and renewable absorption, achieving rates of 62.3% and 52.8% in normal and severe weather, respectively, and (3) an adaptive V-parameter mechanism and a hierarchical fault-tolerance system that ensure robustness and a 99.8% service guarantee.

Notwithstanding these contributions, this study has several limitations that pave the way for future research. It should be noted that the research on the collaborative scheduling strategy for electricity and computing among multi-data center clusters in this paper is based on the dual uncertain pressures of the growth in computing power scale and the increase in the proportion of new energy. Therefore, the impact of computing power migration on electricity demand and new energy utilization efficiency is only described at the level of migration constraints, while issues such as frequency stability and power impact caused by computing power migration require further in-depth research.

Building upon this, future work will focus on three directions. First, we will develop more accurate, short-term renewable forecasting models to further reduce the performance gap caused by prediction errors. Second, the impact of network dynamics, including latency jitter and bandwidth constraints, on the stability and efficiency of large-scale task migration will be thoroughly investigated. Finally, we plan to validate and refine the proposed strategy through real-world deployments in a pilot data center cluster, moving from simulation to practical implementation.

Acknowledgement: The authors would like to acknowledge the support from the State Grid Gansu Electric Power Company for providing the research environment and data support.

Funding Statement: This research is supported by the project "Research on Planning Methods for Gansu Electricity-Computing Coordination under Multi-Spatiotemporal Scales" (No. SGGSJY00XXJS2500043) from the State Grid Gansu Electric Power Company Economic and Technological Research Institute.

Author Contributions: Conceptualization: Qian Yang, Shenglei Du, Zhiheng Zhang. Methodology: Shenglei Du, Boyang Chen, Zhiheng Zhang. Software: Boyang Chen, Yalu Sun. Validation: Ding Li, Yalu Sun, Qian Yang. Formal analysis: Shenglei Du, Boyang Chen. Investigation: Qian Yang, Ding Li. Resources: Qian Yang, Ding Li. Data curation: Yalu Sun, Boyang Chen. Writing—original draft: Shenglei Du, Zhiheng Zhang. Writing—review & editing: All authors. Visualization: Boyang Chen, Yalu Sun. Supervision: Qian Yang, Zhiheng Zhang. Project administration: Qian Yang. Funding acquisition: Qian Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Zhiheng Zhang, upon reasonable request.

Ethics Approval: Not applicable. This study does not involve human participants, animal subjects, or any ethical concerns.

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

References

1. Mingrui L, Jiebei Z, Lujie Y, et al. Adaptive load scheduling algorithm for multi-data centers to maximize new energy utilization. Proceedings of The CSEE. 2025;45(6):2071–83. (In Chinese).
2. Rong H, Zhang H, Xiao S, Li C, Hu C. Optimizing energy consumption for data centers. Renew Sustain Energy Rev. 2016;58(11):674–91. doi:10.1016/j.rser.2015.12.283.
3. Mathew V, Sitaraman RK, Shenoy P. Energy-aware load balancing in content delivery networks. In: 2012 Proceedings IEEE INFOCOM; 2012. p. 954–62. doi:10.1109/INFCOM.2012.6195846.
4. Yang D, Wang X, Shen R, Li Y, Gu L, Zheng R, et al. Prosumer data center system construction and synergistic optimization of computing power, electricity and heat from a global perspective. Ther Science an Eng Prog. 2024;49:102469. doi:10.1016/j.tsep.2024.102469.
5. Heydari A, Gharaibeh AR, Tradat M, Soud Q, Manaserh Y, Radmard V, et al. Experimental evaluation of direct-to-chip cold plate liquid cooling for high-heat-density data centers. Appl Therm Eng. 2024;239:122122. doi:10.1016/j.applthermaleng.2023.122122.
6. Zheng S, Su C, Yang X, Zhang Y, Duan K, Zhang Y, et al. A comprehensive review of single-phase immersion cooling in data centres. Appl Therm Eng. 2025;272:126385. doi:10.1016/j.applthermaleng.2025.126385.
7. Zolfaghari R, Sahafi A, Rahmani AM, Rezaei R. Application of virtual machine consolidation in cloud computing systems. Sustain Comput Inform Syst. 2021;30:100524. doi:10.1016/j.suscom.2021.100524.
8. Chen M, Gao C, Song M, Chen S, Li D, Liu Q. Internet data centers participating in demand response: a comprehensive review. Renew Sustain Energ Rev. 2020;117:109466. doi:10.1016/j.rser.2019.109466.

9. Ding Z, Xie L, Lu Y, Wang P, Xia S. Emission-aware stochastic resource planning scheme for data center microgrid considering batch workload scheduling and risk management. IEEE Trans Ind Appl. 2018;54(6):5599–608. doi:10.1109/TIA.2018.2851516.
10. Cao Y, Cheng M, Zhang S, Mao H, Wang P, Li C, et al. Data-driven flexibility assessment for internet data center towards periodic batch workloads. Appl Energy. 2022;324:119665. doi:10.1016/J.APENERGY.2022.119665.
11. Lian Y, Li Y, Zhao Y, Yu C, Zhao T, Wu L. Robust multi-objective optimization for islanded data center microgrid operations. Appl Energy. 2023;330(1):120344. doi:10.1016/J.APENERGY.2022.120344.
12. Dong Z, Liu N, Rojas-Cessa R. Greedy scheduling of tasks with time constraints for energy-efficient cloud-computing data centers. J Cloud Comput. 2015;4(1):1–14. doi:10.1186/s13677-015-0031-y.
13. Yu L, Jiang T, Zou Y. Price-sensitivity aware load balancing for geographically distributed internet data centers in smart grid environment. IEEE Trans Cloud Comput. 2018;6(4):1125–35. doi:10.1109/tcc.2016.2564406.
14. Zhang S, Yang J, Shi Y, Wu X, Ran Y. Dynamic energy storage control for reducing electricity cost in data centers. Math Probl Eng. 2015;2015(4):380926–13. (In Chinese). doi:10.1155/2015/380926.
15. Hogade N, Pasricha S, Siegel HJ. Energy and network aware workload management for geographically distributed data centers. IEEE Trans Sustain Comput. 2022;7(2):400–13. doi:10.1109/tsusc.2021.3086087.
16. Kang D-K, Yang E-J, Youn C-H. Deep learning-based sustainable data center energy cost minimization with temporal MACRO/MICRO scale management. IEEE Access. 2019;7:5477–91. doi:10.1109/access.2018.2888839.
17. Yang T, Jiang H, Hou Y, Geng Y. Research on carbon neutrality regulation method for interconnected multi-data centers based on spatio-temporal migration of computing load. Proceedings of the CSEE. 2022;42(1):164–76. (In Chinese). doi:10.1109/tcc.2021.3130644.
18. Guo C, Wang X, Zheng Y, Zhang F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. Energy. 2022;238(Part C):121873. doi:10.1016/j.energy.2021.121873.
19. Su J, Zhang H, Liu H, Liu D. Lyapunov-based distributed secondary frequency and voltage control for distributed energy resources in islanded microgrids with expected dynamic performance improvement. Appl Energy. 2025;377(Part C):124539. doi:10.1016/j.apenergy.2024.124539.
20. Li M, Yuan P. A lyapunov optimization strategy with deep reinforcement learning in cloud-edge collaborative service offloading scenarios. In: 2024 10th International Conference on Computer and Communications (ICCC). Chengdu, China; 2024. p. 551–5. doi:10.1109/ICCC62609.2024.10942027.