

A Review on Source-Grid-Load-Storage Coordinated Planning and Operation Optimization for Carbon-Neutral Data Centers

Yiming Lu

*Hertfordshire College, Changzhou Institute of Technology, Changzhou, China
24260117@czust.edu.cn*

Abstract. The energy consumption of data centers continues to rise, currently accounting for around 1% of global electricity use, with the majority of carbon emissions originating from server operations and cooling systems. This paper reviews mechanisms and technologies for carbon-neutral data centers, with a particular focus on the integrated architecture of "Source-Grid-Load-Storage." Through a synthesis of relevant literature, it formulates carbon emission limits and evaluation metrics, including Scope 1-3 emissions and Carbon Usage Effectiveness (CUE). On this basis, a unified analysis framework is built that links low-carbon planning, such as site selection optimization, multi-site integration, efficient designs, liquid cooling technologies, and renewable-energy-adaptive capacity setups, with operational optimization, including mixed clean energy scheduling, electricity-carbon coordination, cascaded waste-heat recovery, carbon-aware workload shifting, AI-driven energy management, and multi-type energy storage coordination. The results indicate that synergy between geographically distributed computing flexibility and dynamic carbon signals can greatly reduce emissions and costs; however, full-life-cycle carbon management and privacy-preserving distributed optimization still lack sufficiently in-depth treatment.

Keywords: Carbon-Neutral Data Center, Source-Grid-Load-Storage, Low-Carbon Planning, Carbon-Aware Scheduling, Waste Heat Recovery

1. Introduction

With the advancement of the "dual carbon" goals and the rapid development of the digital economy, the electricity consumption and carbon emissions of data centers have become the most prominent issues among energy-intensive infrastructures. According to data from the International Energy Agency (IEA), global data center electricity consumption has been steadily increasing. Without effective mitigation measures, the resulting carbon footprint could significantly undermine efforts to meet the temperature control targets outlined in the Paris Agreement. However, advances in building efficiency, renewables, liquid cooling, and workload migration enable emission reduction, but remain siloed and lack system-level integration. In particular, issues like the overall framework for carbon emission accounting and "source-grid-load-storage" coordination, the coupling of cross-timescale planning and operational optimization, and the relationships among multi-agent market

mechanisms and cyber-physical integration control have rarely been explored in depth. Focusing on carbon-neutral data centers, this paper reviews recent literature, clarifies emission boundaries and key metrics (e.g., Scope 1-3, CUE), and proposes an integrated planning-operation framework. At the planning level, it includes site selection, multi-site integration, passive efficiency, liquid cooling, and renewable capacity planning; at the operational level, it covers multi-energy scheduling, electricity-carbon coordination mechanisms, waste heat recovery and utilization, carbon-aware spatiotemporal workload migration, AI-driven intelligent energy management, and coordinated multi-type energy storage systems. By summarizing the commonalities and limitations of existing research, this paper provides references for low-carbon design, operational strategies, and policy-making for data centers, and further presents prospects for future directions such as cross-domain collaboration, digital twins, privacy preservation, and full-life-cycle carbon management.

2. Foundations of carbon neutral data centers

2.1. Emission sources and performance metrics

Data centers require defined boundaries and metrics for emissions reporting and carbon neutrality. Accordingly, emissions are classified into three scopes under the Greenhouse Gas Protocol [1]. Scope 1 comprises direct emissions from fossil fuel combustion, corporate vehicle use, and backup diesel generators. Scope 2 covers indirect emissions from purchased energy (electricity, heat, steam), using location- or market-based accounting. Besides, Scope 3 comprises supply-chain emissions, including equipment and material production, transport, waste, and transmission and distribution losses [1].

From an energy consumption perspective, data centers are dominated by IT equipment (50%-60%) and cooling systems (30%-40%), with buildings and auxiliary facilities contributing about 10%-15% [2]. This distribution shows that emission reduction strategies should prioritize improving computing load efficiency and cooling system energy efficiency. In the meantime, concerning the evaluation indicators, PUE is utilized to evaluate the overall energy efficiency of the infrastructure, and CUE is used to measure the carbon emission per unit of IT energy consumption, thereby directly connecting energy efficiency and carbon emissions [1].

2.2. Integrated source-grid-load-storage framework

The achievement of carbon neutrality in data centers is a cross-timescale, multi-factor coupled system optimization problem, necessitating coordinated planning and operation within the "source-grid-load-storage" framework. The planning stage involves site selection, low-carbon building design, energy system capacity configuration, and multi-station convergence; these decisions collectively determine the energy structure and define the carbon emission boundaries [3-5]. Corresponding to the planning constraints, the operational stage realizes emission reduction and cost optimization through energy dispatch, electricity market participation, load regulation, and energy storage coordination [6-9]. At the system level, the "source" side represents low-emission energy provisioning, including distributed photovoltaics, wind power, fuel cells, hydrogen energy, and waste heat recovery, whose generation profiles shape overall carbon intensity [1, 10, 11]. The "grid" side forms a coupled relationship with the power system and the carbon market, incorporating carbon costs into operational decisions via tiered carbon trading, electricity-carbon coordinated demand response, and carbon allowance trading mechanisms, thereby altering the economic boundaries of energy dispatch [7, 9, 12]. The "load" side leverages the schedulable nature of

computing workloads to achieve carbon-aware allocation through task decomposition and cross-region migration, thus reshaping spatio-temporal computing demand [8, 13, 14]. The "storage" side operates through the combined effect of electrochemical energy storage, cold storage systems, and the virtual energy storage capability formed by server thermal inertia, enhancing system flexibility and buffering renewable energy fluctuations [4, 15]. Under the joint influence of planning constraints and operational scheduling, these elements form dynamic coupling relationships that support carbon neutrality in data centers.

3. Low-carbon planning and data center integration

3.1. Site selection and multi-station integration

Driven by user behavior and service workloads, data center computing demand exhibits significant spatio-temporal uncertainty, making traditional deterministic site-selection methods difficult to balance cost and performance. To address this challenge, existing approaches jointly model workload fluctuations and site-selection decisions using a two-stage robust optimization model that minimizes total cost under worst-case scenarios [3]. By analyzing the fluctuation range of computing demand, this model evaluates the robustness of different siting schemes under load variations and applies improved Benders decomposition and column-and-constraint generation (CCG) algorithms, thus achieving about a 4.73% cost reduction while ensuring availability during peak workloads. Once the site is determined, the data center further undergoes multi-station convergence planning with energy stations and substations, enabling the thermal storage capacity of the data center to serve as a flexible load for power regulation. The alternating direction method of multipliers (ADMM) is used to jointly optimize multiple energy vectors, including electricity and heat [4]. This multi-station coordination approach optimizes the distribution of electricity and heat, enhances energy utilization efficiency, incorporates carbon emissions into scheduling constraints, and supports the low-carbon deployment of cross-regional data centers while maintaining service quality and system stability.

3.2. Energy-saving architecture and liquid cooling

Passive energy-saving design at the building level is a key approach to lowering overall energy use due to the high energy density of data centers. Common strategies include hot-aisle containment to mitigate air mixing losses, vertical thermal stratification to improve inter-floor heat distribution, and building-integrated photovoltaics to reduce dependence on external energy supply. Building on these foundations, liquid cooling and free cooling technologies overcome the limitations of traditional air cooling for heat dissipation, with designs achieving a PUE as low as 1.07 [2]. As single-server power rises to tens of kilowatts, air-cooling systems, constrained by limited heat-transfer capacity, are increasingly unable to meet high-density heat dissipation requirements. Under low load conditions, cooling power consumption may exceed IT power consumption, resulting in energy consumption inversion, risk of localized overheating and increasing operational costs. To solve this problem, liquid cooling has gradually become the main heat dissipation method for high-power servers, which can improve the heat exchange efficiency through direct heat transfer and meet the needs of higher rack density. On this basis, dynamic cooling control adjusts coolant flow rate and temperature in real time according to thermal conditions and load, achieving approximately 10% energy savings compared with constant control [16]. Further optimization of piping layouts combined with air-liquid hybrid cooling structures enables on-demand refrigeration and reduces

pump and fan energy consumption. Concurrently, thermal inertia management is introduced in high-density computing scenarios to buffer short-term load fluctuations and ensure operational stability.

3.3. Capacity provisioning under renewables

When renewable energy sources such as wind and photovoltaic power are introduced, the power supply becomes highly susceptible to weather and environmental conditions, exhibiting pronounced intermittency and stochasticity. As a result, traditional fixed capacity planning cannot reliably sustain operations under load fluctuations, necessitating capacity configuration methods with adaptive capabilities. To address the single-site capacity uncertainty problem, a planning framework based on a physics-informed digital twin constructs a model consistent with the actual physical and energy systems of a data center, enabling real-time mapping of energy supply and load demand. On this basis, the framework dynamically adjusts power supply and distribution capacity, cooling capacity, and energy storage dispatch, and optimizes capacity configuration by integrating load forecasting with renewable energy output. This approach can reduce carbon emissions by approximately 27% under varying load and weather conditions [5]. In multi-site scenarios, capacity planning is further coupled with task scheduling: computing tasks are dynamically allocated to sites with more abundant renewable energy, and capacity configurations are updated via iterative optimization methods, so that each site achieves a match between energy consumption and renewable energy supply on an annual timescale. This synergistic mechanism enhances efficiency and reduces dependence on the power grid., thereby supporting carbon neutrality. During optimization, energy storage states, task migration costs, and resource utilization must be jointly considered to ensure quality of service under high-load conditions [11].

4. Operational optimization at the "source" and "grid" sides

4.1. Renewable energy integration and clean energy supply

Once the clean energy capacity is determined, the operational phase is characterized by dynamic dispatch of power grid, distributed photovoltaics, fuel cells and energy storage. These energy sources have very different characteristics: grid power is stable but with load-dependent carbon emissions and electricity price; renewable energy is low-carbon but highly intermittent; fuel cells may respond smoothly but can not fully match load; energy storage may respond quickly but with limited capacity. The dispatch strategies optimize cost and carbon emissions based on load and source characteristics [10]. A cost minimization model addresses uncertainty by converting stochasticity into a deterministic mixed-integer program, prioritizing renewable energy use while fuel cells and the grid provide supplementary power and energy storage handles transient fluctuations. This approach is more cost-effective, load-balanced, and carbon-aware than grid-only schemes [10]. In multi-site environments, joint data and energy migration enables cross-site coordinated scheduling, where computation and energy flows are co-optimized over data and energy networks, thus achieving 85.8% direct energy migration efficiency and a 1.8% efficiency loss per additional intermediate node. The optimization jointly considers migration losses, VM migration energy consumption, electricity purchase and feed-in losses under constraints of power balance, computing capacity, and storage status. Sequential short-window optimization improves robustness against solar variability, equipment aging, and low storage conditions [11].

4.2. Electricity-carbon collaborative market mechanisms

Market-based carbon pricing exposes carbon costs and drives low-carbon scheduling in data centers. It uses tiered carbon trading to raise penalties for excess emissions, increasing marginal abatement incentives during high-emission periods. This mechanism is combined with electricity buying and selling, renewable energy output, wind and solar curtailment penalties, load adjustment compensation, and energy storage operation in microgrid scheduling. Loads are classified into rigid, shiftable, and transferable types for flexible dispatch, while the framework is formulated as a mixed-integer linear program, reducing operating cost and carbon emissions [7]. In multi-data-center scenarios, carbon trading is extended to an internal allowance market, where cross-site coordination is achieved via bilevel optimization: the upper level sets electricity and carbon prices, while the lower level optimizes load allocation, storage scheduling, and internal trading to minimize cost under emission constraints. To reduce communication and privacy overhead, the method adopts a distributed solution based on the ADMM and incorporates a threshold-triggered mechanism, cutting the communication volume by roughly one-third, achieving a cost reduction of about 16.2%, and reducing dependence on external carbon markets. At the electricity-carbon market coupling level, carbon-coordinated demand response integrates carbon costs into load scheduling via dynamic carbon intensity and carbon price signals, using a two-layer privacy architecture to separate regulation and execution. The upper layer broadcasts electricity price and carbon intensity information, while the lower layer optimizes task allocation accordingly and returns only the aggregated load. This approach achieves a marked reduction in carbon emissions even when the data center load share rises to 20% [12].

4.3. Waste heat recovery and energy cascade utilization

Virtually all the electrical energy consumed by IT equipment is converted into heat; therefore, if the continuously generated waste heat can be recovered and utilized, data centers can be transformed into district heating nodes and achieve cascade utilization [10, 12]. Low-grade waste heat (typically 30–40 °C) can be upgraded by heat pumps for building heating or greenhouse use. For example, the Stockholm project connects to a district heating network via heat pumps, achieving a coefficient of performance of 8.2, while the Ulanqab project uses absorption heat pumps to raise waste heat to about 60 °C for greenhouse heating [12]. At higher temperature ranges, waste heat can be further used in absorption refrigeration systems for combined cooling and power supply, or in organic Rankine cycles for power generation at temperatures above 100 °C, and in low-temperature applications it can even be used for processes such as desalination [1]. The introduction of liquid cooling technology raises the temperature and stability of the waste heat, enhancing its exploitability and broadening its application scope while reducing the energy consumption of the cooling system itself. This process creates a synergistic effect at the system level: lower cooling energy consumption coincides with the delivery of higher-quality waste heat that provides additional energy value, thus improving life-cycle energy efficiency. Thus, the design of low-carbon data centers must consider the cooling system, IT layout, and external heat demand in a coordinated manner to form a synergistic thermal energy system that spans from chips and servers to computer rooms and district energy networks [2, 16].

5. Intelligent dispatch of "load" and "storage" resources

5.1. Spatio-temporal migration of computing workloads for carbon awareness

Computing tasks in data centers can be shifted across both time and space. In particular, temporal migration defers non-urgent workloads via queuing and batching to periods of lower grid carbon intensity, while spatial migration routes loads via optical fiber networks to regional data centers with higher renewable energy availability. This capability stems from the spatiotemporal complementarity of renewables, where PV peaks are staggered across longitudes and multi-regional wind aggregation reduces variability, enabling cross-domain scheduling [13]. On this basis, a spatio-temporal carbon-aware scheduling method jointly optimizes task timing and layout of tasks under SLA constraints, incorporating thermal power emissions, regional electricity use, and network communication energy into carbon accounting to enable low-carbon execution of delay-tolerant tasks. Task allocation is coupled with network routing and solved via Benders decomposition, where the master problem determines task allocation and the subproblem optimizes optical routing to minimize communication energy consumption and generate Benders cuts for iterative refinement. Simulations on the Google B4 topology, Parallel Workloads Archive traces, and real-world renewable energy data show that the proposed method reduces carbon emissions and improves renewable energy utilization compared with no-migration and time-only migration schemes, while remaining robust across seasons, load levels, and renewable penetration rates [13]. In large-scale carbon-aware scheduling, reinforcement learning agents are introduced to learn independently with different weights on carbon emissions and quality of service, approximate the Pareto front, and integrate decisions, reducing electricity-related carbon emissions while preserving service quality [8].

5.2. AI-driven energy management and regulating

In distributed data centers, deep reinforcement learning can optimize energy consumption, while model predictive control and online convex optimization remain effective in specific scenarios. For multi-agent collaboration, this problem is modeled as a Dec-POMDP and solved using the MADDPG method to address the issues of privacy constraints and the limited scalability of centralized methods. It utilizes online learning to adapt to workload and fluctuations in renewable energy, and achieves real-time interaction among agents through communication [6]. In multi-objective optimization, the integration of game theory and reinforcement learning enables efficient scheduling. AI inference tasks in geographically distributed data centers are modeled as independent "players" whose strategy is request-rate allocation to minimize carbon emissions or operational cost. The model considers multi-task interference, IT and cooling power coupling, renewable energy offset, and peak power charges. Each player learns a policy via PPO, guided by Nash equilibrium constraints, while reinforcement learning exploration helps avoid local optima. In experiments with 4–16 U.S. data centers, the method outperforms conventional scheduling, genetic algorithms, pure game-theoretic methods, and single-agent reinforcement learning in both carbon emissions and cost [14]. Online optimization provides theoretical performance guarantees. A Lyapunov optimization–based approach transforms long-term electricity cost minimization into per-time-slot mixed-integer linear programs, converts delay constraints into queue stability via virtual queues, and minimizes an upper bound on drift-plus-penalty in each slot. The method requires no future information, uses a single tunable parameter to balance cost and delay, and provides asymptotic performance guarantees [17].

5.3. Multi-type energy storage and virtual energy storage buffering

At the energy storage level, energy storage smooths renewable fluctuations and maintains power balance in data centers, including electrochemical storage, cold/thermal storage, and virtual storage via thermal inertia and task buffering [14]. Electrochemical storage provides millisecond-level response and enables bidirectional UPS–grid interaction; deployed systems connect UPS batteries to frequency regulation, achieving ~80 ms response while preserving backup reliability [15]. In terms of cold energy storage, chilled water storage stores cooling capacity in stratified layers, while ice slurry storage leverages the latent heat of phase change to increase energy density, and both are used in conjunction with time-of-use electricity pricing to charge cooling at night and discharge it during the day, lowering operating costs and shaving peak loads [15]. Virtual storage includes thermal inertia and computing-task caching. Thermal inertia uses the thermal capacity of building envelopes, racks, servers, and cooling water to stabilize temperature during short-term cooling reductions, providing regulation during price peaks or grid constraints; models quantify demand response capacity through discretized temperature dynamics with constraints on temperature and its rate of change [3]. By deferring batch tasks to later time periods, computing-task caching shifts instantaneous electricity demand and achieves load shifting; although it lacks bidirectional energy flow, it can be regarded as computational-side storage [15]. The coordination of multiple types of energy storage requires joint optimization of battery charging/discharging, cooling storage operation, thermal inertia regulation, and task scheduling. Due to differences in capacity, power, efficiency, and response time across storage types, and the discrete nature of task scheduling, the problem is a mixed-integer programming problem spanning multiple time scales and physical processes [11, 18].

6. Conclusion

Based on a carbon accounting framework and a "source-grid-load-storage" architecture, this paper reviews low-carbon planning and operational optimization in data centers. At the planning level, two-stage robust siting, multi-station integration, liquid and free cooling, and adaptive capacity configuration can effectively reduce carbon emissions, with the lowest current PUE approaching 1.07. At the operation level, priority dispatch of renewable energy, tiered carbon trading, and electricity–carbon co-optimized demand response make carbon costs explicit while enabling joint optimization of emission reduction and economic returns. Enabled by the spatiotemporal migratability of data center tasks, carbon-aware cross-regional scheduling is achieved. Meanwhile, reinforcement learning and game-theoretic methods are combined to improve low-carbon operation while ensuring quality of service. Besides, the co-optimization of thermal inertia, electrochemical, and chilled water storage turns data centers into flexible energy nodes with peak shaving and frequency regulation capabilities, while waste heat recovery enables energy supply to district systems. Existing studies lack quantitative comparison of economic viability and applicability across low-carbon technologies and insufficiently consider life-cycle embodied emissions, decommissioning, and grid carbon responsibility allocation. Future work should focus on multiscale electro-thermal-carbon digital twins, electricity–carbon market distributed decision-making, unified computing–thermal virtual storage modeling, and life-cycle carbon neutrality evaluation and cost allocation for low-carbon data centers.

References

- [1] Cao, Z., Zhou, X., Hu, H., Wang, Z., & Wen, Y. (2022). Toward a systematic survey for carbon neutral data centers. *IEEE Communications Surveys & Tutorials*, 24(2), 895–936. <https://doi.org/10.1109/COMST.2022.3161275>
- [2] Chen, S. G. (2025). Low-carbon design of data center building structures and full life-cycle carbon management. *China Building Decoration and Renovation*, 00(18), 93. <https://doi.org/10.3969/j.issn.1672-2167.2025.18.014>
- [3] Deng, R., Song, M., Qiao, L., Xia, Y., Bai, W., & Gan, W. (2022). Interconnected data center siting planning method considering uncertainty of decision dependence. *Automation of Electric Power Systems*.
- [4] Li, L. H., He, S. J., Gao, H. J., & Liu, J. Y. (2026). Low-carbon distribution network planning method considering energy station–substation–data center multi-station integration. *Power System Technology*, 50(2), 494–504.
- [5] Cao, Z., et al. (2025). Adaptive capacity provisioning for carbon-aware data centers: A digital twin-based approach. *IEEE Transactions on Sustainable Computing*, 10(4), 730–743. <https://doi.org/10.1109/TSUSC.2025.3526192>
- [6] Liu, C., Wan, J., Li, L., Ren, G., & Wang, X. (2024). Distributed energy management for carbon neutral data centers. In *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications*. 1433–1438.
- [7] Zhao, D., Su, X., Huang, S., Chen, J., & Gao, L. (2024). Economic and low-carbon dispatch of a data center microgrid considering ladder carbon trading. In *2024 4th Power System and Green Energy Conference*. 1232–1237.
- [8] Wang, Y., Sun, W., Ren, P., & Harrison, G. (2026). Multi-objective low-carbon scheduling method for data centers based on ensemble reinforcement learning. *IEEE Transactions on Smart Grid*, 17(1), 297–308.
- [9] Yan, D., Chow, M.-Y., & Chen, Y. (2024). Low-carbon operation of data centers with joint workload sharing and carbon allowance trading. *IEEE Transactions on Cloud Computing*, 12(2), 750–761.
- [10] Hu, X., Li, P., & Sun, Y. (2021). Minimizing energy cost for green data center by exploring heterogeneous energy resource. *Journal of Modern Power Systems and Clean Energy*, 9(1), 148–159.
- [11] Da Silva, M. D. M., Gamatié, A., Sassatelli, G., Poss, M., & Robert, M. (2023). Optimization of data and energy migrations in mini data centers for carbon-neutral computing. *IEEE Transactions on Sustainable Computing*, 8(1), 68–81. <https://doi.org/10.1109/TSUSC.2022.3197090>
- [12] Ruan, J., et al. (2024). Privacy-preserving bi-level optimization of Internet data centers for electricity-carbon collaborative demand response. *IEEE Internet of Things Journal*, 11(14), 24948–24959.
- [13] Yang, T., Jiang, H., Hou, Y., & Geng, Y. (2022). Research on spatiotemporal dual-dimensional migration-based carbon-neutral control method for interconnected multi-data centers. *Proceedings of the CSEE*, 42(1), 164–177.
- [14] Hogade, N., & Pasricha, S. (2025). Game-theoretic deep reinforcement learning to minimize carbon emissions and energy costs for AI inference workloads in geo-distributed data centers. *IEEE Transactions on Sustainable Computing*, 10(4), 628–641. <https://doi.org/10.1109/TSUSC.2024.3520969>
- [15] Wu, Z., Chen, H., Jia, X., Wang, S., Pang, J., Jiang, G., et al. (2025). Grid-friendly low-carbon regulation methods and practices for data centers. *New Type Power Systems*, 3(3), 263–284.
- [16] Ni, L., Shao, L., & Yang, W. (2025). Energy-saving pathways of liquid cooling technology in data centers under the dual-carbon background. *Low Carbon World*, 15(12), 75–78.
- [17] Yu, L., Jiang, T., Cao, Y., & Qi, Q. (2014). Carbon-aware energy cost minimization for distributed Internet data centers in smart microgrids. *IEEE Internet of Things Journal*, 1(3), 255–264.
- [18] Du, B., Jia, H., Li, Y., Du, E., Zhang, N., & Liang, D. (2025). Estimating the carbon emission reduction potential of using carbon-oriented demand response for data centers: A case study in China. *iEnergy*, 4(1), 54–64.