Towards Joint Energy Efficiency and Performance Optimization: A Scheduling Strategy for Green-Aware Heterogeneous Computing

Ling Wu

China United Network Communications Co., Ltd. Nanjing Branch, Nanjing, Jiangsu 210000

Abstract: ***Against the backdrop of unprecedented growth in global computing demands and the rapid expansion of the artificial intelligence (AI) industry, the energy consumption and carbon emissions associated with computing infrastructure have emerged as pivotal challenges, significantly impeding the green transformation of the digital economy. The escalating reliance on high-performance computing (HPC) systems, data centers, and edge computing devices to support AI-driven applications, big data analytics, and cloud services has led to a substantial increase in energy use, exacerbating environmental concerns. This surge in energy demand not only strains power grids but also contributes to greenhouse gas emissions, undermining sustainability efforts in the digital sector. A critical challenge in this context is the spatio-temporal mismatch between resource scheduling and green energy supply within heterogeneous computing power networks. These networks, which integrate diverse computing resources such as CPUs, GPUs, and specialized accelerators, often operate in environments where renewable energy sources like solar and wind are intermittent. The variability in renewable energy generation, coupled with the dynamic nature of computing workloads, creates a complex scheduling problem. Traditional resource allocation strategies, which predominantly prioritize performance metrics such as latency and throughput, fail to account for the availability and utilization of green energy. Consequently, computing nodes frequently rely on non-renewable energy sources, leading to higher carbon footprints and inefficiencies in energy use. To address these challenges, this paper proposes a green energy-aware computing scheduling strategy designed to optimize the spatio-temporal coordination between heterogeneous computing resources and renewable energy supply. The strategy is grounded in a comprehensive sensing framework that integrates three key dimensions: task latency sensitivity, computational preference characteristics, and node-level green energy reserves. By analyzing task latency sensitivity, the strategy identifies workloads that can tolerate flexible execution times, enabling them to be scheduled during periods of abundant renewable energy. Computational preference characteristics, which include the type of processing required (e.g., CPU-intensive or GPU-intensive tasks), are used to match tasks with the most suitable computing nodes, thereby enhancing overall efficiency. Node-level green energy reserves are continuously monitored to ensure that tasks are allocated to nodes with sufficient renewable energy availability, minimizing reliance on non-renewable sources.***

Keywords: Green Energy, Computing Scheduling, Heterogeneous Computing, Renewable Energy, Digital Economy, Sustainability.

1. INTRODUCTION

Driven by exponential growth in global computing demand, the parameter scale and training compute requirements of large language models represented by the GPT series are expanding super-linearly. In 2022, global computing power reached 906 EFLOPS with an annual growth rate exceeding 47%, and it is projected to surpass 20 by 2030. Simultaneously, China's AI industry has entered a phase of rapid development; in 2023, the first batch of large-model products from eight institutions passed regulatory filing and entered commercial use. The digital infrastructure underpinning industrial upgrading—data centers—is expanding at an annual rate of 30%. However, the energy cost of computing infrastructure is becoming increasingly severe: by 2030, China's data-center energy consumption is expected to equal the total output of the Three Gorges Dam over five years, and carbon emissions will exceed 310 million tons, placing data centers at the forefront of the digital economy's carbon-neutral challenge.

Against this backdrop, raising the green level of computing power carries dual strategic value: on the one hand, as a new-quality productive force in the digital economy, every 1 yuan invested in computing power can leverage 3–4 yuan of economic output, yet the current data-center average utilization rate of only 55% in the extensive operating mode causes enormous resource waste; on the other hand, optimizing computing-power scheduling strategies across time and space to achieve spatiotemporal synergy between computing loads and renewable

energy has been proven to be a core pathway that delivers both economic and environmental benefits. This paper focuses on green scheduling mechanisms in the computing-power network, aiming to break through the global optimization problem of computing-network resources. By improving the utilization efficiency of green computing power, it pursues the dual goals of "improving quality and efficiency while reducing cost and carbon," providing both infrastructure support for overcoming AI computing bottlenecks and a theoretical paradigm and practical pathway for the sustainable development of the digital economy.

2. OVERVIEW OF THE COMPUTING-POWER NETWORK

The computing-power network aims to achieve unified management and resource allocation across network, storage, and computing dimensions, thereby enabling global scheduling of computing resources. The overall architecture of the computing-power network is shown in Figure 1, with its core divided into four layers: the infrastructure layer, the resource pooling layer, the task scheduling layer, and the network operation layer.

In the infrastructure layer, computing resources consist of local computing centers, edge MEC servers, and remote cloud computing centers, while the communication system comprises optical transport networks and routing/forwarding nodes. Within the coverage of the computing-power network, computing resources at various points connected by the communication network are uniformly measured in the resource pooling layer, covering key metrics such as CPU resources, GPU resources, storage resources, and bandwidth resources. These measurements provide the basis for efficient resource allocation and utilization.

In the resource pooling layer, the status information of each computing node is reported in real time to the upper-layer SDN controller via a northbound interface. This status information includes electricity price information of the computing node, network availability, and the current load of the computing cluster. By integrating this information, the SDN controller provides data support for global resource management, enabling dynamic optimization of resources and efficient task scheduling.

In the domain of privacy protection, Li, Lin, and Zhang (2025) developed a framework incorporating federated learning and differential privacy for advertising personalization[1]. Intelligent system design includes Tu's (2025) platform-aware framework for 5G network test automation[2], Xie and Liu's (2025) multimodal sentiment analysis for recruitment processing[3], and Zhu's (2025) reliability engineering with automated causal tracking[4]. Industry-specific applications feature Zhuang's (2025) analysis of digital transformation in real estate marketing[5], Han and Dou's (2025) hierarchical graph attention network for recommendation systems[6], Zhang et al.'s (2025) AI-driven sales forecasting in gaming[7], and Cheng et al.'s (2025) investigation of executive human capital effects on stock volatility[8]. Computer vision research shows substantial progress with Chen et al.'s (2022) gaze-estimation based object referring[9] and Tian et al.'s (2025) cross-attention multi-task learning for digital advertising[10]. Medical imaging advances include Chen et al.'s (2023) vision-language pretraining for unified segmentation[11], while environmental finance applications feature Zhang et al.'s (2025) deep learning approach for carbon market forecasting[12]. The field of 3D vision and human pose estimation is significantly advanced by Peng et al.'s work on 3D vision-language Gaussian splatting[13] and their research on aggregation and segregation of representations for domain adaptive human pose estimation[14].

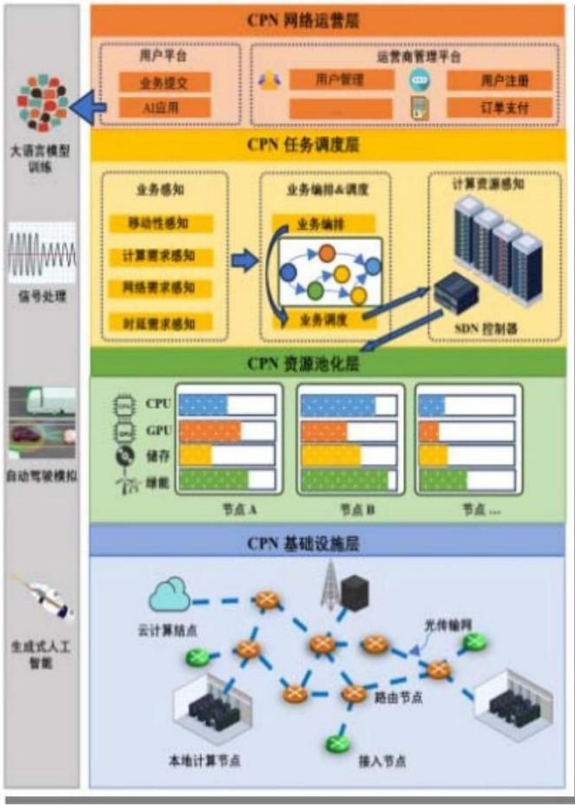

Figure 1: Overall architecture of the computing-power network

In the task-scheduling layer, when a user platform submits a service, the layer first performs multi-dimensional perception of that service, including mobility perception, computing-demand perception, network-demand perception, and latency-demand perception. Latency-demand perception specifically refers to the maximum tolerable response delay from the moment the user submits the service to the moment the task finishes computing and returns the result. For scenarios in which multiple tasks compete for resources in a distributed-computing workload, the service-orchestration unit intelligently orchestrates the current service queue, coordinates the execution order of different computing tasks and data flows, and then issues the orchestration result to the service-scheduling unit. The service-scheduling unit obtains global compute-network state information from the SDN controller, combines the orchestration result with real-time state information, generates a service-deployment plan through predefined policies, and, via southbound interfaces, interacts with the compute-network infrastructure to perform network-configuration activation, compute-node container configuration, and other operations, thereby achieving efficient utilization of compute-network resources and optimized task scheduling.

The network operations layer delivers AI applications and the associated computing services to users via the user platform; based on users' quality requirements for these services, the computing provider submits their computational tasks to the task-scheduling layer through the southbound interface.

3. GREEN ENERGY-AWARE SCHEDULING STRATEGIES IN COMPUTING POWER NETWORKS

Computing-power scheduling is the core issue in optimizing computing networks; its main objective is to optimize overall system performance—including key metrics such as task completion time, computing resource utilization, and energy consumption—through rational resource allocation and task-scheduling strategies. The essence of computing-power scheduling lies in efficiently assigning user-submitted computing tasks to appropriate computing nodes to satisfy the tasks' latency requirements, computational demands, and energy constraints.

Compute scheduling in the compute power network faces multiple challenges. First, compute nodes in the network usually possess heterogeneous resources (e.g., CPU, GPU, TPU), so efficiently matching task requirements with node capabilities is a key issue. Second, green energy sources such as solar and wind are intermittent and uncertain;

leveraging them while minimizing their consumption is a major scheduling objective. Moreover, scheduling must balance latency and energy: latency-sensitive tasks prioritize completion time, whereas latency-tolerant tasks can focus more on energy optimization. The optimization goals of task scheduling are typically multi-objective, mainly including the following aspects:

(1) Latency guarantee: Ensure tasks are completed within the specified latency requirements to meet users' quality-of-service needs.

(2) Energy consumption optimization: Through rational scheduling, reduce the total energy consumption of the system, especially the consumption of brown energy, to achieve sustainable development.

(3) Improved resource utilization: By means of rational task allocation, computational resources are utilized more efficiently, waste is avoided, and overall system performance is maximized.

Figure 2 illustrates a typical computing-power scheduling process. Users submit their computing demands to a computing-power provider over the network; the scheduler selects an appropriate compute node and allocates computing resources according to task characteristics and network latency. The computing task is transmitted to the designated node following the schedule provided by the computing-power network scheduler, executed there, and the results are returned to the user upon completion.

During task scheduling, several key issues must be addressed:

(1) Task–node matching: the computational preference of a task (e.g., serial compute-intensive, parallel compute-intensive) must align with the node's computational capability to ensure efficient execution.

(2) Utilization of green energy: the scheduling algorithm assigns tasks preferentially to nodes rich in green energy according to the green-energy supply, reducing brown-energy consumption.

(3) Trade-off between latency and energy: the scheduling algorithm must balance latency and energy, ensuring latency-sensitive tasks finish on time while minimizing overall system energy consumption.

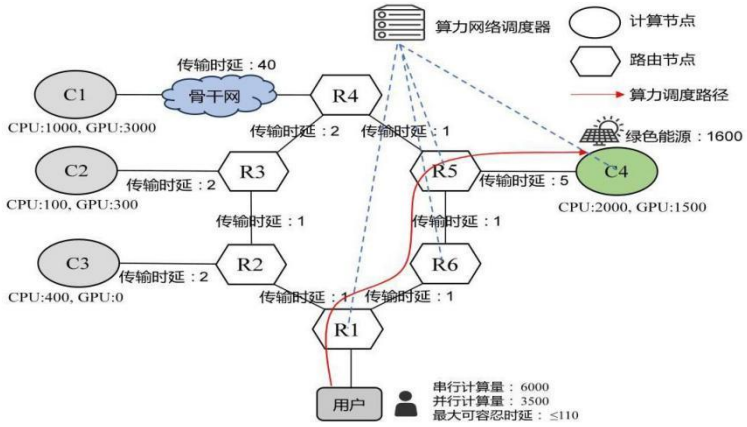

Figure 2: Green-energy-aware computing resource scheduling

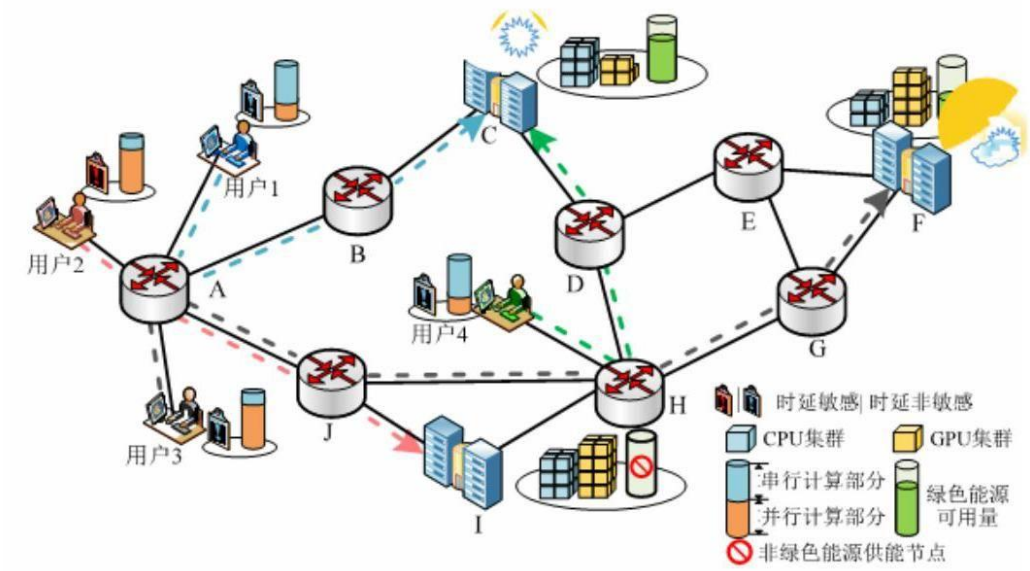

Figure 3: Green-energy-aware computing resource scheduling strategy

Computing resource scheduling is of great research value in computing networks, especially under the context of green-energy awareness. How to optimize the energy consumption and latency of computing tasks through reasonable scheduling strategies is a key direction for future research. To address the three key issues above, this paper proposes a Green Energy-Aware Computing Resource Scheduling (GECRS) strategy. Based on the latency requirements and computational preferences of tasks, and combined with the green-energy reserves of computing nodes in the network, GECRS balances energy consumption and latency during task execution.

As shown in Figure 3, a brief example is provided to illustrate how GECRS works. For ease of comparison, assume identical bandwidth and transmission loss between nodes. The computing network contains three nodes (C, F, H). Due to geographical and climatic differences, the renewable-energy capacity available to each node varies significantly. Moreover, the computing resource capacities of the nodes differ. In the example, node I has more CPU/GPU units than C and F, but, constrained by the environment, its power mainly comes from brown energy. Region C enjoys abundant solar energy; node C deploys twice as many CPU units as GPU units. Figure 3 also lists four example tasks, roughly classified along two dimensions—latency sensitivity and computational preference:

(1) latency-sensitive & parallel compute-intensive (user 2);

(2) latency-insensitive & serial compute-intensive (user 1, user 4);

(3) latency-insensitive & parallel compute-intensive (user 3).

When user 1 submits a computing task, the traditional Shortest Path First (SPF) method provides two paths: path 1: A − B − C, path 2: A-J-I. However, the SPF strategy cannot select an appropriate computing node according to the task's computing preference, which may lead to a decline in the overall resource utilization of the computing network. Meanwhile, because SPF lacks green-energy awareness, it cannot effectively utilize renewable energy within the computing network. The GECRS strategy first considers the latency sensitivity of the computing task and, while guaranteeing task quality, schedules tasks according to the principle of prioritizing green energy. For user 1, the task is latency-insensitive and serial-compute-intensive, so it can be preferentially scheduled to a node with green-energy supply and abundant CPU resources. In the example, node I has no green energy, and node F is a GPU-rich node that does not match user 1's computing preference; therefore, the optimal scheduling path chosen by GECRS is: A − B − C.

User 4, whose computing type is similar to that of user 1, is located closest to node I in the computing network, and the transmission latency and energy consumption to reach node C and node F are identical. According to the GECRS principle, the node that best meets user 4's needs is still node C, and the optimal path is: H −D − C.

User 2's task is latency-sensitive and parallel-compute-intensive; GECRS prioritizes a low-latency scheduling strategy. Although the transmission latency to nodes C and I is equal and both are lower than that to node F , node I has the most computing resources and can provide the lowest latency; therefore, the optimal scheduling plan for user 2 is: A – J – I.

User 3 shares the same starting position as user 1 in the computing network and also has a latency-insensitive task. However, user 3's task is parallel-compute-intensive, so node C better matches user 3's computing preference than node F. Based on this, the optimal scheduling strategy for user 3 is: A-J-H-G-F. It is worth emphasizing that, compared with path A-B-C, this choice incurs higher transmission latency and energy consumption, yet it avoids the overall resource-utilization drop caused by a mismatch in computing type. The GECRS strategy can reduce service costs for computing-network operators while ensuring service quality.

4. CONCLUSION

In summary, as future demand for computing power accelerates, the energy consumption of computing infrastructure is becoming increasingly prominent. Effectively leveraging green energy to reduce computational energy use in heterogeneous computing environments has become one of the core objectives of green development. This paper introduces a computing-network architecture and proposes a green-energy-aware computing scheduling strategy. The strategy simultaneously considers task computing preferences, heterogeneous resources in the computing network, and green energy reserves, formulating scheduling plans based on the varying delay sensitivity of tasks, thereby achieving spatiotemporal matching between computing demand and green computing supply.

REFERENCES

[1] Li, X., Lin, Y., & Zhang, Y. (2025). A Privacy-Preserving Framework for Advertising Personalization Incorporating Federated Learning and Differential Privacy. arXiv preprint arXiv:2507.12098.
[2] Tu, Tongwei. "AutoNetTest: A Platform-Aware Framework for Intelligent 5G Network Test Automation and Issue Diagnosis." (2025).
[3] Xie, Minhui, and Boyan Liu. "EvalNet: Sentiment Analysis and Multimodal Data Fusion for Recruitment Interview Processing." (2025).
[4] Zhu, Bingxin. "REACTOR: Reliability Engineering with Automated Causal Tracking and Observability Reasoning." (2025).
[5] Zhuang, R. (2025). Evolutionary Logic and Theoretical Construction of Real Estate Marketing Strategies under Digital Transformation. Economics and Management Innovation, 2(2), 117-124.
[6] Han, X., & Dou, X. (2025). User recommendation method integrating hierarchical graph attention network with multimodal knowledge graph. Frontiers in Neurorobotics, 19, 1587973.
[7] Zhang, Jingbo, et al. "AI-Driven Sales Forecasting in the Gaming Industry: Machine Learning-Based Advertising Market Trend Analysis and Key Feature Mining." (2025).
[8] Cheng, Ying, et al. "Executive Human Capital Premium and Corporate Stock Price Volatility." Finance Research Letters (2025): 108278.
[9] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).
[10] Q. Tian, D. Zou, Y. Han and X. Li, "A Business Intelligence Innovative Approach to Ad Recall: Cross-Attention Multi-Task Learning for Digital Advertising," 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Shenzhen, China, 2025, pp. 1249-1253, doi: 10.1109/AINIT65432.2025.11035473.
[11] Chen, Yinda, et al. "Generative text-guided 3d vision-language pretraining for unified medical image segmentation." arXiv preprint arXiv:2306.04811 (2023).
[12] Zhang, Zongzhen, Qianwei Li, and Runlong Li. "Leveraging Deep Learning for Carbon Market Price Forecasting and Risk Evaluation in Green Finance Under Climate Change." Journal of Organizational and End User Computing (JOEUC) 37.1 (2025): 1-27.
[13] Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., Chen, C. and Wu, Z., 3D Vision-Language Gaussian Splatting. In The Thirteenth International Conference on Learning Representations.

[14] Peng, Qucheng, Ce Zheng, Zhengming Ding, Pu Wang, and Chen Chen. "Exploiting Aggregation and Segregation of Representations for Domain Adaptive Human Pose Estimation." In 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1-10. IEEE, 2025.