OPEN ACCESS

LETTER

Journal homepage: www.the-innovation.org/energy-use

Energy Use

INNOVATION PRESS

Energy Use 2 (2026):100046

Contents lists available at INNOVATION PRESS

Energy-Aware Computing Power Sharing For Sustainable Artificial Intelligence

Xiaochen Liu,[1,*] Bowen Zheng,[1] Zhi Fu,[1,4] Li-ang Huang,[2,4] Zhongziheng Xu,[3,4] Tao Zhang,[1] and Xiaohua Liu[1]

[1]Department of Building Science, School of Architecture, Tsinghua University, Beijing 100084, China

[2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China

[3]Department of Engineering Mechanics, Tsinghua University, Beijing 100084, China

[4]ComNergy Co. Ltd., Beijing 100080, China

*Correspondence: lxc4@tsinghua.edu.cn

Received: ; Accepted: June 8, 2026; https://doi.org/10.59717/ipj.energy-use.2026.100046

Citation: Liu X., Zheng B., Fu Z., et al. (2026). Energy-Aware Computing Power Sharing For Sustainable Artificial Intelligence. *Energy Use* **2**:100046.

The surging demand for artificial intelligence (AI) computing power has led to escalating energy consumption and carbon emissions. Meanwhile, low utilization of computing infrastructure and power grid constraints pose critical barriers to sustainable AI development. This paper proposes energy-aware computing power sharing to address these challenges. By leveraging the spatiotemporal flexibility of computing workloads, the sharing system enables load migration across regions and time to leverage available computing resources and renewable energy, thereby reducing energy costs and carbon emissions. Key enabling technologies include heterogeneous computing resource integration, privacy-preserving security isolation, rapid migration and elastic scaling, high-precision supply-demand forecasting, dynamic spatiotemporal scheduling, and market mechanism design. Stakeholder benefits encompass affordable computing access for AI developers, revenue generation from underutilized computing hardware, and reduced investment in power infrastructure. A pilot implementation in China aggregated over 700,000 computing nodes (> 10.6 EFLOPS) and served over 18,000 users. It has maintained 99.99% service availability since its operation. For two typical workload types, it reduced computing costs by 50% and 68%, respectively, while offering substantial potential for improving renewable energy utilization.

CHALLENGES IN SUSTAINABLE ARTIFICIAL INTELLIGENCE

The rapid advancement of AI technologies has led to surging demand for computing power (also termed "computility"). As of June 2025, global computing power has reached 4,495 EFLOPS; it is estimated that, by 2030, it will exceed 50 ZFLOPS, with AI computing power accounting for more than 95%.[1] The energy consumption of AI clusters poses a growing strain on power systems. According to the International Energy Agency, the electricity use from data centers worldwide was 485 TWh in 2025, and is projected to roughly double to 950 TWh by 2030, leading to a substantial increase in carbon emissions.[2] China and the United States are expected to drive this expansion, jointly responsible for nearly 80% of the global increase in data centers' electricity use by 2030.

Ensuring a reliable, affordable, and low-carbon energy supply remains the key challenge in supporting the sustainable development of AI computing power.[3] Due to the contradiction between the surging energy demand for computing power and the relatively slow upgrade of the utility grids, increasing enterprises worldwide plan to construct dedicated power plants (like solar, wind, or even nuclear power) and transmission lines for directly supplying green power to their AI data centers. Although such energy supply solutions are both reliable and low-carbon, the large-scale investment in power infrastructure significantly raises the cost of computing power, thereby hindering the inclusive and equitable development of AI.

Understanding the current characteristics of computing power facilitates the identification of feasible pathways to promote sustainable AI development. The utilization rate of AI computing infrastructure remains low. For example, the average utilization rate of AI computing centers in China was less than 30% in 2025, and in some regions even below 18%.[1] Fully utilizing these idle computing resources will be crucial for reducing the overall cost of AI computing power.[4] It will also help avoid the additional embodied carbon that would arise from constructing new dedicated computing infrastructure. Furthermore, although computing workloads are energy-intensive, they can be transferred across time and region to respond to dynamic electricity prices and renewable generation.[5] Based on this spatiotemporal flexibility, AI data centers can operate as grid-interactive resources, contributing to maintaining AI service quality within existing power infrastructure constraints.[6]

PATHWAY FORWARD: ENERGY-AWARE COMPUTING POWER SHARING

Computing power sharing is the "silver bullet" for unlocking the flexibility of computing workloads, thus minimizing energy costs and carbon emissions.[7] The concept of computing power sharing originates from the sharing economy, a socio-economic system in which participants share resources through renting, lending, and trading to reduce costs and mitigate underutilization. The sharing economy has been adopted in various fields, including ride-hailing (e.g., Uber) and homestay services (e.g., Airbnb). Similarly, the rapid increase in both the supply of and demand for AI computing power offers significant opportunities for sharing computing power across interconnected networks. In March 2026, China's National Development and Reform Commission officially incorporated the computing power network into the national modern infrastructure framework for the first time, designating it a national strategic priority.

Figure 1 illustrates the energy-aware computing power sharing for sustainable AI. The shared resources include CPUs, GPUs, memory, and other related computing infrastructure. Unlike current most cloud service providers that offer dedicated computing resources, the proposed resource providers offer various computing services via full-time or idle-time leasing, while ensuring high reliability. The system's supply side comprises geo-distributed heterogeneous computing devices, ranging from personal computers and local data centers to cloud computing infrastructure. On the demand side, current computing tasks are primarily driven by large language models (LLMs), including training, fine-tuning, and inference. As LLMs become widely deployed, inference tasks could account for up to 90% of LLM

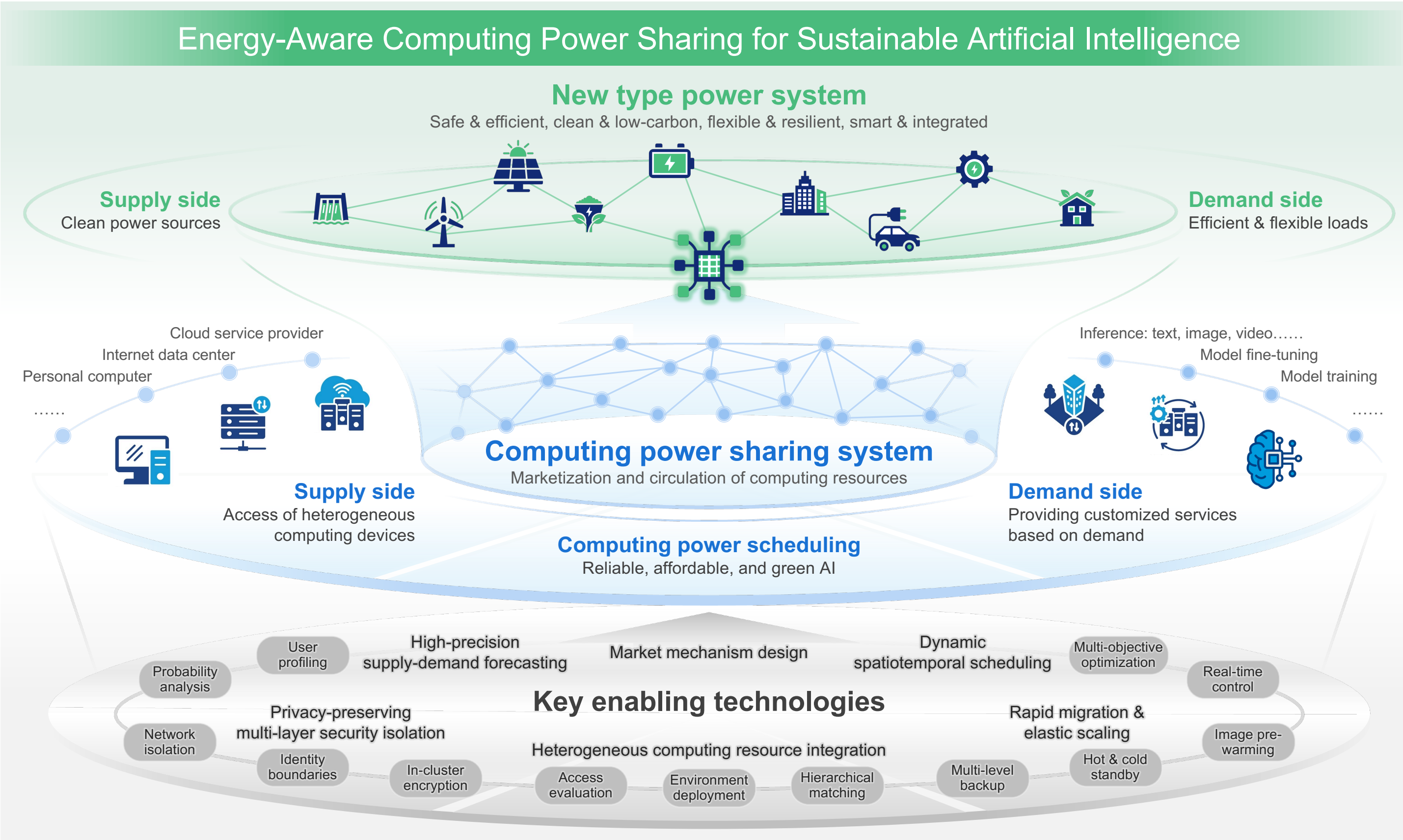

Figure 1. Energy-aware computing power sharing for sustainable AI.

workloads.[8] These tasks, including text, code, image, and video generation using trained models, will become dominant types of demand in computing power sharing systems. Furthermore, with the advancement of edge AI, computing power sharing systems are expected to integrate emerging intelligent terminal devices, including intelligent connected vehicles, low-altitude aircraft, embodied robots, smart buildings, and wearable smart devices.

With the above resources and participants, the computing power sharing system can engage in spatiotemporally flexible interactions with power systems. From a spatial perspective, computing workloads can be dispatched to regions with abundant renewable energy generation or sufficient computing resources, allowing data transmission to replace long-distance power transmission. From a temporal perspective, computing loads can serve as demand-side flexibility resources to harness renewable energy or off-peak electricity, thereby reducing carbon emissions and electricity costs. Under this "computility-electricity coordination" paradigm, the computing power sharing system supports reliable, affordable, and low-carbon AI development.

KEY ENABLING TECHNOLOGIES

The energy-aware computing power sharing system relies on the key enabling technologies outlined below (see Figure 1).

Heterogeneous computing resource integration

The system retrieves device hardware specifications to evaluate and classify devices for varying computational demands. It then verifies the environmental configurations of computing resources, automates deployment tailored to specific task requirements, and deploys the image to the corresponding device, which then executes the respective computational tasks.

Privacy-preserving multi-layer security isolation

A defense-in-depth architecture is necessary for security isolation,[9] including sandbox isolation, Cilium kernel-level network isolation, RBAC identity boundaries, mTLS in-cluster encryption, BGP traffic ingress control, and multi-region cluster failover control. Together, these safeguards ensure comprehensive privacy preservation for infrastructure and data.

Rapid migration and elastic scaling

The hot and cold standby mechanisms enable millisecond-level switching between active and queued states, achieving high reliability on heterogeneous hardware and effectively supporting elastic scaling demands. Concurrently, the image pre-warming mechanism enables rapid cross-region scheduling responsiveness, while the multi-level backup architecture efficiently reduces latency induced by image upload, download, and retrieval. These mechanisms address the problem inherent in existing idle computing services (e.g., spot instances from Amazon Elastic Compute Cloud), specifically the lack of stability guarantee.

High-precision supply-demand forecasting

Advanced time-series forecasting algorithms combined with multi-dimensional user profiling and probabilistic analysis enable accurate prediction of both the spatiotemporal availability of distributed computing nodes and the dynamics of computing demand. High-precision forecasting with risk quantification provides a critical foundation for dynamically aligning volatile workloads with resources.

Dynamic spatiotemporal scheduling

Based on the nonlinear performance–power characteristics of various computing resources, the system performs multi-objective spatiotemporal scheduling optimization to balance service quality, electricity costs, and carbon emissions.[4,10] Furthermore, a rolling horizon control framework is employed to support real-time optimal decision-making.

Market mechanism design

A harmonized market architecture is essential to incentivize participants and guide the spatiotemporal allocation of computing loads toward low-carbon electricity, with mechanisms including energy-aware pricing at high spatiotemporal resolution, incentives such as demand-response compensation, and integration of computing-related emissions into carbon markets.

STAKEHOLDER BENEFITS

The energy-aware computing power sharing delivers simultaneous benefits to all stakeholders, thereby promoting sustainable AI development.

AI developers and small & medium-sized enterprises (SMEs): Accessible and affordable computing power

By aggregating distributed computing resources, the sharing system reduces AI users' structural dependence on hyperscale cloud operators and dedicated clusters. More SMEs, AI startups, and scientific researchers can access enterprise-grade GPU computing power on demand, without heavy upfront capital expenditure on AI infrastructure or high costs of AI computing services. This helps mitigate inequality arising from AI development.

AI computing hardware owners: Converting underutilized assets into productive resources

The sharing system leverages underutilized computing infrastructure to meet emerging demand and generates revenue for AI hardware owners. Meanwhile, meeting part of AI computing demand through existing infrastructure helps reduce the need for newly built dedicated data centers, thereby avoiding additional embodied carbon, supply chain pressure, equipment costs, and material consumption. This supports a more circular and resource-efficient digital economy.

Power system operators: Reducing power infrastructure investments and providing flexible energy resources

The sharing system can partially meet the growing demand for computing power, thereby reducing the need for power infrastructure investments otherwise required for building additional data centers. Furthermore, the sharing system can aggregate distributed and low-intensity computing resources as demand-side flexible energy resources to support the safe, efficient, and low-carbon operation of power systems.

PRACTICES AND PERSPECTIVES

The Tsinghua research team incubated a startup, ComNergy Tech. Ltd., to advance the concept of energy-aware computing power sharing from theory to practice. The computing power sharing system consists of the following platforms, which have undergone large-scale engineering validation beyond proof-of-concept research.

The supply-side platform for computing resource access (suanleme.cn) enables individuals to connect their GPUs to the computing power sharing system, supporting automatic environment configuration (validated across 94% of personal devices), workload sandboxing, and idle-time monetization. As of May 2026, the platform has aggregated computing resources from 42 AI data centers spanning nine Chinese provinces and over 700,000 distributed computing nodes, including AI servers, high-performance personal computers, and internet café computers. The total computing accessed power has exceeded 10.6 EFLOPS

The demand-side platform for computing power service (suanli.cn) provides elastic deployment services, GPU cloud hosting, bare-metal servers, and batch services to AI developers and enterprises. As of May 2026, the platform has served over 18,000 registered users, spanning individuals, AI model developers, AI inference companies, AI-generated content communities, scientific computing groups, and university research teams. The platform has dispatched over 4.2 million computing tasks and maintained a service availability exceeding 99.99%, defined as the system's uptime ratio for accepting incoming requests. In August 2025, the platform processed 3 billion PDF pages of pre-training data for ModelBest, an AI company specializing in on-device LLMs. These intermittent batch workloads possess partly latency tolerance, allowing them to avoid computation during peak periods. Under the elastic on-demand model, which bills only during task active periods, the hourly cost fell from 1.53 CNY (the full rental rate for an NVIDIA RTX 4090 GPU) to an average of 0.49 CNY—a 68% reduction. In another case, since October 2025, the platform has provided text-to-image computing services to LiblibAI, an AI image-generation platform. These interactive workloads operate at an average utilization of roughly 40%. Under the same model, the cost fell to an average of 0.77 CNY—a 50% reduction.

The computing power scheduling platform helps IDC operators, cloud vendors, and government agencies efficiently manage their computing resources. Our team has contributed to developing several provincial-level platforms. Using the platforms and regional time-of-use tariffs, we simulated 20 interactive workloads dispatched across 40 NVIDIA RTX 4090 GPUs in Chongqing, Fujian, and Henan. The results demonstrate an 18% reduction in computing costs and a 20% increase in renewable energy utilization.[4] Currently deployed in the Qinghai pilot, this platform promises even greater potential for cost savings and renewable energy adoption, given the region's abundant renewable resources.

Our practices have demonstrated that energy-aware computing power sharing offers a viable pathway toward sustainable AI. AI computing power can be transformed from being high-barrier, high-cost, and centralized to easily accessible, low-cost, and ubiquitous. In the future, scaling up such sharing practices across broader regions may accelerate the global adoption of AI and broaden access to its benefits. Such a globally interconnected computing power sharing network could enable new forms of international collaboration on computing and energy resources, fostering a paradigm distinct from traditional frameworks.

REFERENCES

1. China Academy of Information and Communications Technology. (2026). Blue Book on Advanced Computing and Computing Power Development Index (2025). Available online: https://www.caict.ac.cn/kxyj/qwfb/bps/202603/t20260306_715320.htm
2. International Energy Agency. (2026). Key Questions on Energy and AI. International Energy Agency. *IEA*. Available online: https://www.iea.org/reports/key-questions-on-energy-and-ai
3. Ge X.X., Du B.J., Du E.S., et al. (2026). Sustainable Artificial Intelligence: Power Through Computility–Electricity–Carbon Coordination. *IEEE Energy Sustaina. Mag.* **2**:89–102. DOI:10.1109/ESM.2025.3633634
4. Fu Z., Zheng B.W., Liu X.C., et al. (2026). A Computing-Power Coordination Method and Practice Based on Idle Computing Resource Scheduling. *Auto. Elect. Power Sys.* Accepted.
5. Zhang Y.B., Tang H., Li H.X., et al. (2025). Integration and interaction of next-generation AI-focused data centers with smart grids and district energy systems: The state-of-the-art, opportunities and challenges. *Renew. Sustain. Energy Rev.* **224**:116097. DOI:10.1016/j.rser.2025.116097
6. Colangelo P., Coskun A.K., Megrue J., et al. (2026). AI data centres as grid-interactive assets. *Nat. Energy* **11**:254–261. DOI:10.1038/s41560-025-01927-1
7. Sun J.J., Fu Z., Ma J.H., et al. (2026). Sharing Computing Power: Enhancing Energy Efficiency and Reducing Carbon Emissions. *IEEE Energy Sustaina. Mag.* **2**:103–119. DOI:10.1109/ESM.2025.3627176
8. Morgan T.P. (2024). The Battle Begins for AI Inference Compute in the Datacenter. The Next Platform, September 2024. Available online: https://www.nextplatform.com/2024/09/10/the-battle-begins-for-ai-inference-compute-in-the-datacenter/#:~:text=applications%20or%20to%20create%20entirely,processes%20with%20modest%20packaging%20needs
9. Theodoropoulos T., Rosa L, Benzaid C, et al. (2023). Security in cloud-native services: A survey. *J. Cybersecur. Priv.* **3**:758–793. DOI:10.3390/jcp3040034
10. Liu X.C., Luo Z.Y., Zhang T., et al. (2025). Ergonomics in Energy Use: Bridging Energy System-Oriented Flexibility and Human-Oriented Service Quality. *Engineering*. DOI:10.1016/j.eng.2025.12.002

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the Ministry of Industry and Information Technology: Computing Infrastructure Enhancement and Open Bidding for Selecting the Best Candidates (Cross-Domain Task Orchestration System for Heterogeneous Computing Power), Beijing Natural Science Foundation (QY25032), and the Tsinghua University Undergraduate Academic Research Promotion Program.

DECLARATION OF INTERESTS

The authors declare no competing interests.