Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2026.070201
ISSN 2789-7168 Vol. 7, Issue 2:1-9

Server Energy Consumption Monitoring and Adaptive Control in Large-Scale Data Center

Zhu Xu

Electrical and Computer Engineering, Northwestern University, Evanston 60208, IL, USA

Keywords: Large-scale data center; server energy consumption monitoring; power capping; adaptive control; PUE; load scheduling

Abstract: As the scale of AI training, online inference and cloud services expands continuously, the energy consumption of servers in large-scale data centers has changed from a traditional energy efficiency problem at the data center level to a collaborative optimisation problem involving servers, racks, clusters and power systems. This paper introduces a multi-granularity monitoring and adaptive control framework for server energy consumption, focusing on the technical ideas of "observable server energy consumption, predictable load status, closed-loop control actions, and constrained service quality". First, the growth trends of global and US data center power demand are analyzed, and then key problems at five levels are raised: sensor acquisition, power consumption modelling, thermal constraints, service quality constraints, and closed-loop control. Next, build server power consumption models for CPUs, GPUs, memory, fans and power supply units, and design an adaptive control strategy that integrates power capping, DVFS, task migration, sleep/wake-up and cooling coordination. Based on the above results, this method can reduce the peak power and energy consumption per unit of task while meeting SLA requirements, providing an engineering path for energy-saving operation of data centers in high-density AI computing scenarios.

1 Introduction

Data centres are now the basic facilities for the digital economy and artificial intelligence. Now only the management Department of the facility is concerned with the problem of energy consumption; other parts of the company's operations, such as power grid planning, carbon emission reduction and cloud service stability, have also begun to be affected. According to the International Energy Agency's data in "Energy and AI", the global electricity consumption of data centers in 2024 will be approximately 415 TWh, making up about 1.5% of the world's total electricity use, and under the baseline scenario, it is expected to reach almost 950 TWh by 2030, a rate of increase that significantly exceeds that of electricity demand in other areas [1]. Therefore, if the old methods of static capacity planning, experience threshold alarms and manual scheduling are still used, it will be difficult to meet the energy consumption limit, thermal safety range and business latency demand simultaneously.

A 2024 report by Lawrence Berkeley National Laboratory in the United States further showed

that the total electricity consumption of data centers in the United States reached 176 TWh in 2023, accounting for 4.4% of the total electricity consumption in the United States, and may rise to a range of 325-580 TWh by 2028 [2]. The main reasons for this rapid growth are GPU-accelerated servers, high-density AI racks and continuous online inference. Compared with traditional Web loads, AI loads have the characteristics of high power density, strong short-term fluctuations, and obvious cluster parallel dependence; therefore, server energy consumption monitoring needs to change from "monthly electricity bill accounting" to "second-level status perception", and the control method must also be altered from "device-level switch control" to "task-resource-thermal environment collaborative optimisation".

The aim of this paper is to put forward a multi-granularity monitoring and adaptive control method for server energy consumption in a large-scale data center. The contributions of this paper are as follows: First, to build an energy consumption monitoring index system for chips, complete machines, racks and clusters; second, to establish a unified optimisation model that incorporates IT power consumption, cooling power consumption, thermal risk and SLA penalties; third, to propose a power capping and load migration strategy based on rolling prediction; and fourth, to demonstrate the energy efficiency benefits and engineering boundaries of the solution through publicly available statistical data and scenario-based simulations. The organization of this paper is as follows: Introduction, analysis of the current situation in research, problem definition, proposed solution, and conclusion.

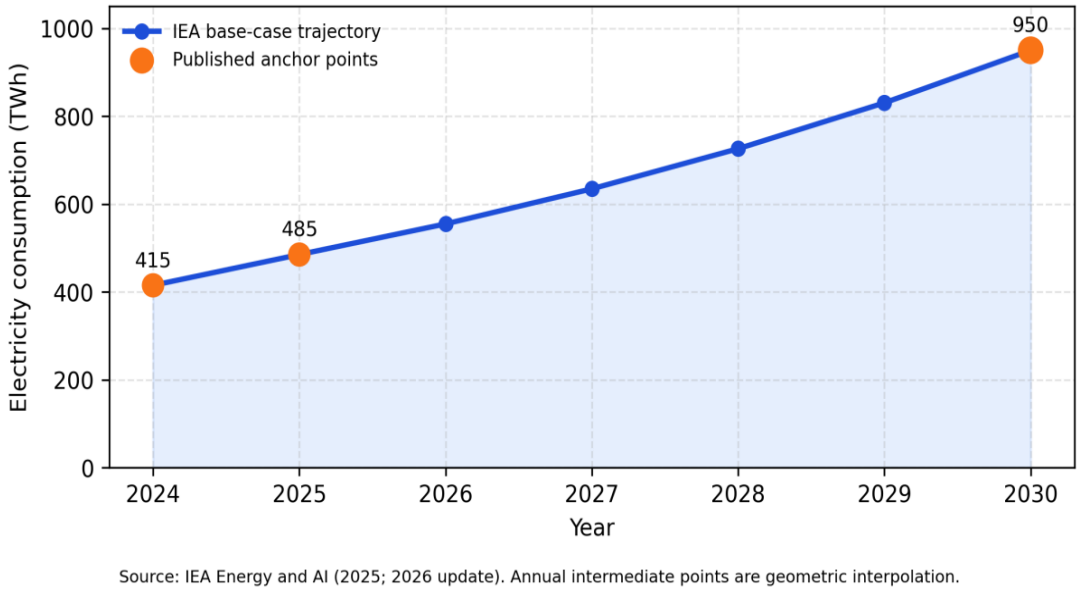


Figure 1. Changes in the Baseline Scenario for Global Data Center Power Demand.

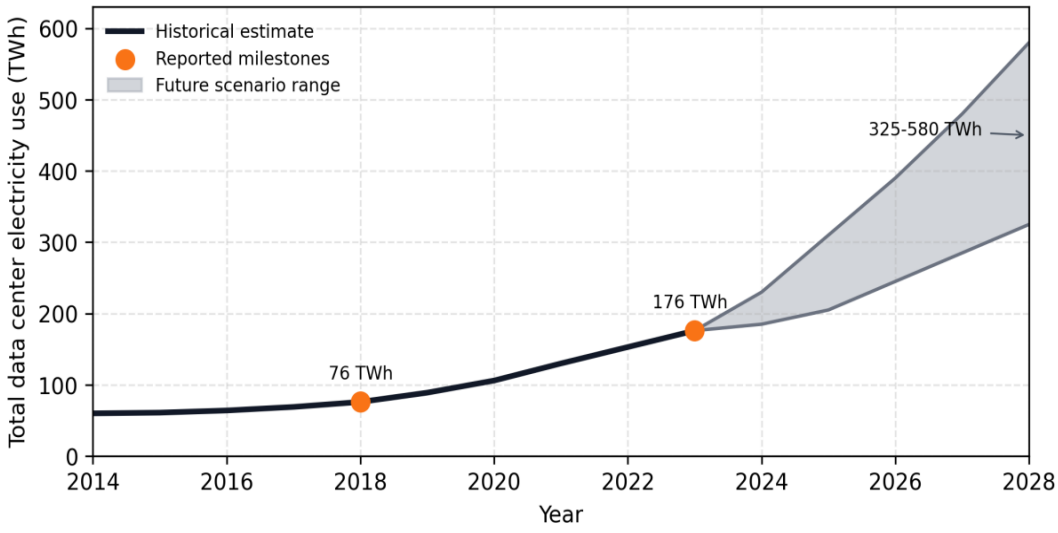
Figure 1 Analysis: Based on publicly available IEA data, the redrawn Figure 1 shows a baseline scenario where global data centre power demand will increase rapidly from 415 TWh in 2024 to about 950 TWh in 2030. 2024, 2025 and 2030 are publicly available anchor points, and the intermediate years are interpolated using a geometric growth rate to show the trend slope. As shown in Figure 1, the growth rate of data center energy consumption has become highly elastic; therefore, with the increase in AI inference and training workloads, server-side power management is now a primary subject of energy-efficiency governance.

2. Current Status Analysis of the Research Topic

2.1 Facility energy efficiency indicators shift from PUE to fine-grained energy consumption on the IT side

For a long time, energy efficiency assessment of data centres has focused on PUE (Power Usage Effectiveness), which is the ratio of the total facility power to IT equipment power. According to the 2024 worldwide survey by Uptime Institute, the average PUE for the industry was about 1.56 at that time, and the overall rate of improvement has gradually slowed down in recent years [3]. PUE can reflect the efficiency of power supply and distribution and cooling, but it is not easy to show problems such as idle power consumption of servers, transient power of AI accelerator cards, and local hot spots caused by container scheduling. Therefore, the focus of energy consumption management in modern data centres has gradually shifted to the server and workload level; we now need to collect data from PDUs, electricity meters and UPSs, as well as CPU RAPL, GPU telemetry, BMC sensors, process-level resource utilisation and business queue length.

In 2024, Bianchini and others summarized the development of power and energy management in cloud data centers and pointed out that the focus of the next stage will be on cross-layer collaborative management, including hardware power consumption capping, software scheduling, electricity price response, and carbon-sensing operation [4]. Therefore, it can be concluded that the infrastructure system alone is insufficient for controlling server energy consumption; cooperation among the cloud platform, resource orchestrator, AIOps system and energy management system is required.



Source: LBNL 2024 United States Data Center Energy Usage Report; values redrawn from Figure ES-1/5.5.

Figure 2. Historical and Scenario-Based Total Power Consumption of US Data Centers

Figure 2 Analysis: Based on historical estimates and scenario ranges from the LBNL 2024 report, Figure 2 is a reconstruction. U.S. data center power consumption was about 60 TWh in 2014-2016, increased to 76 TWh in 2018, and reached 176 TWh by 2023; in 2028, the lower and upper bounds of the low- and high-case scenarios were around 325 TWh and 580 TWh, respectively. As shown in Figure 1, high-density accelerated servers have shifted the pattern of "business growth but relatively stable energy consumption" over the past decade, and in the future, energy consumption control will need to directly affect server power curves and peak loads.

2.2 Server Power Consumption Prediction and Online Monitoring Technology

At the bottom is the relationship between state and power for energy use. Long and others (2025) applied machine learning methods to predict the electricity demand of data centres and found that CPU utilisation, memory access, network throughput, disk I/O and ambient temperature are all related to fluctuations in server power consumption [5]. However, in actual engineering, power consumption prediction still faces three types of difficulties: First, the power curves of different generations of servers and GPUs are very different; Second, telemetry data is missing, delayed and inconsistent in timestamps; Third, the business load changes suddenly at the minute or even second level, and a static model will be easily inaccurate.

Sun et al. have formulated the learning-based adaptive power capping framework for cloud data centers as a partially observable Markov Decision Process and reduced the interaction cost in real systems through model-based reinforcement learning [6]. This is a direction for large-scale data centres: the power manager may not have access to all the internal information of the job, but it can learn control strategies based on aggregated power, SLA default rate, electricity price or carbon intensity feedback. Mazzola and others used hardware performance counters for power modelling and kernel-level monitoring to show that a low-overhead online estimator can support control actions such as DVFS and task scheduling [7].

2.3 Adaptive control shifts from single-machine frequency modulation to cluster collaboration

The previous energy-saving measures for the server were dynamic voltage and frequency scaling, idle hibernation and fan-speed reduction. Recently, research has begun to shift from general power capping to cluster-level power capping, performance-aware control and carbon-aware scheduling. Raj and others have used reinforcement learning to reduce the power consumption of computing nodes and believe that the control strategy needs to meet both the demands of application performance and energy efficiency [8]. Rodrigues and others have studied carbon-aware optimisation in the context of cross-data centre transmission scheduling, and shown that network transmission and regional power-plant structures also affect the total energy-saving benefits for data centres [9]. Narayanaswamy and his colleagues have also shown in their research on data centre energy optimisation power configuration that a relatively small performance drop can be achieved with significant energy efficiency improvements [10].

In short, although the existing research has built a technical chain from power consumption monitoring and predictive modelling to control optimisation, there are still three deficiencies: First, a unified closed-loop system connecting monitoring data with control actions has not been established; Second, most methods focus on a single level and cannot simultaneously constrain chip power, rack thermal risks and business SLAs; Third, power fluctuations are more pronounced in high-density AI scenarios, and thus higher demands are placed on the stability, interpretability and engineering safety boundaries of control strategies.

3. Problem Statement: Key Bottlenecks in Energy Consumption Management of Servers in Large-Scale Data Centers

3.1 It is difficult for multi-source energy consumption data to form a consistent state space.

Energy consumption data of large-scale data centres are obtained from PDUs, electricity meters, BMCs, GPU management interfaces, container runtimes, schedulers, cooling systems and business monitoring platforms. These data types have different sampling frequencies, timestamp baselines and granularities. Data centre electricity meters generally record the total power at the minute or

hour level; server BMCs provide second-level total power consumption data, and CPU and GPU telemetry can reach sub-second resolution. Concatenate this data as a feature vector directly; otherwise, there may be false correlations, peak misalignment and control latency. Therefore, a unified state space needs to be established, and at the edge, time alignment, anomaly removal and low-latency aggregation must be carried out.

Table 1. Server Energy Consumption Monitoring Indicators and Sampling Granularity

Layer	Signals	Sampling interval	Control use
Chip	CPU package power, GPU power, memory bandwidth	1–5 s	DVFS, power capping
Server	BMC power, fan speed, inlet temperature	5–10 s	thermal protection
Rack	PDU current, rack power density, hot-spot alarm	10–30 s	workload migration
Cluster	job queue, SLA latency, utilization, carbon signal	30–60 s	placement and scheduling
Facility	PUE, cooling power, UPS loss, grid price	1–5 min	energy orchestration

Table 1 Analysis: The five levels of server energy consumption monitoring in Table 1 are: chip, whole machine, rack, cluster and facility. The main reason is that the closer the control action is to the hardware, the higher the sampling frequency needed; the closer it is to scheduling and electricity price response, the more aggregated the state can be. This indicator system does not have the feedback delay caused by using only the main power meter in the data center, and it also avoids the increased monitoring cost of collecting too much chip-level data.

3.2 The power consumption model needs to take into account nonlinearity, heterogeneity, and interpretability.

Server power consumption does not have a simple linear relationship with CPU utilization. For AI servers, GPU tensor core utilization, memory bandwidth, PCIe/NVLink communication, fan speed, and intake air temperature all affect power consumption. This article expresses the server 's power consumption *i*at any given time *t*as:

$$P_{i,t}=P_i^{idle}+\left(P_i^{peak}-P_i^{idle}\right)u_{i,t}^{\alpha_i}+\beta_i g_{i,t}+\gamma_i m_{i,t}+\delta_i v_{i,t}+\varepsilon_{i,t} \quad (1)$$

In Equation (1), P_i^{idle} represents the server idle power, P_i^{peak} represents the peak power, $u_{i,t}$ represents the CPU utilization, $g_{i,t}$ represents the GPU utilization or GPU power status, $m_{i,t}$represents the memory bandwidth usage, $v_{i,t}$represents the fan speed or cooling-related variable, α_i,β_i,γ_i,$\delta_i$represents the parameter to be estimated, $\varepsilon_{i,t}$and represents the random error. This model retains the nonlinear utilization term while supporting engineering audits through linearly interpretable parameters.

3.3 There are conflicts between the control objectives and energy consumption, performance, and thermal security.

Server Energy-saving Control Also Reduces Power Consumption. An excessively low power cap will extend the time required to complete the job and increase the risk of SLA violation; too much task migration will raise network power consumption and cache miss rates; and excessively high cooling setpoints may cause a localized hotspot. Total Energy Consumption of a Data Centre can be shown as:

$$E_{DC}=\sum_{t=1}^{T}\left(\sum_{i=1}^{N}P_{i,t}+P_{cool,t}+P_{loss,t}\right)\Delta t,\quad PUE_t=\frac{P_{IT,t}+P_{cool,t}+P_{loss,t}}{P_{IT,t}} \tag{2}$$

In equation (2), E_{DC}represents the total power consumption of the data center during the period, $P_{IT,t}=\sum_i P_{i,t}$ represents the power of the IT equipment, $P_{cool,t}$ represents the cooling power, $P_{loss,t}$represents the power of the UPS and power distribution losses, and Δtrepresents the sampling step size. This equation shows that reducing IT power alone does not necessarily reduce total energy consumption, because the cooling system and power distribution losses also change with the load distribution.

4. Problem Solving and Strategies: Server Energy Consumption Monitoring and Adaptive Control Framework

4.1 Overall Architecture: A four-layer closed loop of monitoring, prediction, decision-making, and execution.

The four layers of the system in this paper are: The first is the monitoring layer, and it will gather indicators for chips, the whole machine, racks, clusters and facilities. The second is the prediction layer that builds a short-term power, temperature and SLA risk prediction model. The third layer is the decision layer, and under the constraints of safety, Model Predictive Control (MPC) is used to generate actions such as power capping, DVFS, migration, hibernation and cooling linkage. The fourth level is the execution layer; at this time, control commands are issued to the BMC, Kubernetes scheduler, GPU management interface and DCIM system. The goal of this architecture is not to replace the existing scheduler, but rather to add energy-consumption-status-aware and closed-loop-correction functions to the current operation-and-maintenance system.
A server in a hot state can be modelled as a first-order discrete process:

$$T_{i,t+1}=a_iT_{i,t}+b_iP_{i,t}+c_iT_{in,t}+d_iF_{rack,t}+\eta_{i,t} \tag{3}$$

In equation (3), $T_{i,t}$represents the critical component or inlet air temperature of the server, $T_{in,t}$represents the cold aisle inlet air temperature, $F_{rack,t}$represents the rack airflow or cooling intensity, a_i,b_i,c_i,d_irepresents the thermal response parameter, and $\eta_{i,t}$represents the disturbance term. This model can link power consumption control with the thermal safety boundary, preventing the neglect of local temperature accumulation by only considering utilization.

Table 2. Adaptive Control Action and Constraint Mapping

Control action	Actuator	Main constraint	Expected effect
Power capping	BMC/RAPL/GPU driver	SLA latency, node stability	peak shaving
DVFS	CPU governor, GPU clock	throughput loss	power-performance tuning
Workload migration	scheduler/orchestrator	network cost, data locality	hot-spot reduction
Sleep/wake	server controller	wake-up delay	idle energy reduction
Cooling coordination	DCIM/BMS	thermal safety, humidity	PUE improvement

Table 2 shows that, in the analysis above, action types and constraint boundaries of server energy consumption control need to be distinguished. Power capping reduces the peak power but may affect performance; task migration avoids hotspots by adding network and cache overhead; sleep/wakeup reduces idle energy consumption but needs accurate prediction of when the load will rise. This mapping serves as the engineering action space for the following control strategies.

4.2 Multi-objective optimization model

Control strategies in actual data centres need to consider electricity costs, carbon emissions, SLA requirements, thermal risks and operational jitter simultaneously. The total objective function of this paper is given below:

$$\min_{a_{1:T}} J=\sum_{t=1}^{T}\left[C_t P_{IT,t}\Delta t+\lambda_1 V_{SLA,t}+\lambda_2 R_{temp,t}+\lambda_3 \left\| a_t-a_{t-1} \right\|_1\right] \quad (4)$$

In equation (4), $\mathbf{a}_t$is the control action vector, $C_t$is the electricity price or marginal carbon cost, $V_{SLA,t}$is the SLA default penalty, $R_{temp,t}$is the temperature overrun risk, $\lambda_1,\lambda_2,\lambda_3$and is the weighting coefficient. The last term is used to suppress frequent switching of control actions and avoid system jitter caused by excessive power capping and task migration.

Within each rolling cycle, the controller estimates Hpower, temperature, and operational risks for the next few steps based on the current state, and executes only the first control action. Its update format is as follows:

$$a_t^*=\arg\min_{a_{t:t+H}} \sum_{k=t}^{t+H} \hat{J}_k, \quad x_{t+1}=Ax_t+Ba_t^*+Dw_t+\xi_t \quad (5)$$

In equation (5), $\mathbf{x}_t$represents the system state vector, $\mathbf{w}_t$represents the external disturbance, including service arrival rate, electricity price, carbon intensity, and outdoor temperature, A,B,Drepresents the state transition parameter, and $\xi_t$represents the prediction error. This rolling optimization structure can recalculate control actions when the load changes abruptly, thereby improving system adaptability.

4.3 Control Flow and Safety Boundaries

The six steps in the control process are as follows: First, collect chip, system, rack and service data, align them in time; Second, use a short-term model to predict power and temperature; Third, based on SLA queues, task priorities and rack thermal risks, calculate candidate actions; Fourth, solve a multi-objective optimisation problem within the safety limit; Fifth, distribute the control action to the server and scheduler; and Sixth, feed back the actual power, temperature and service latency into the model for online correction. To reduce engineering risks, all control actions must meet the following three hard constraints: the server core temperature should not exceed the safety limit of the server, critical service latency should not exceed the SLO upper bound, and rack current should not exceed the PDU's rating.

Table 3. Comparison Results of Scenario-Based Simulation of Control Strategies

Strategy	Energy saving	Peak reduction	SLA violation	Mean PUE	Notes
Static threshold	3.8%	5.1%	0.42%	1.53	rule-based alarm
DVFS only	6.7%	8.4%	0.69%	1.51	chip-level control
Power capping + migration	10.9%	15.6%	0.91%	1.47	rack-aware control
Proposed adaptive control	14.8%	21.3%	0.74%	1.44	MPC with safety guard

Table 3 Analysis: Table 3 shows the scenario-based simulation results for a 20 MW-class data center, and the relative effects of the various control strategies are compared here. The static threshold method results in the lowest benefit because it only responds after an anomaly; DVFS can reduce chip power consumption but cannot handle rack hotspots; power capping combined with migration improves peak shaving but increases SLA defaults. The strategy in this paper has

achieved a better trade-off between energy-saving rate, peak-shaving rate and service quality by coordinating the actions of prediction and security constraints.

4.4 Project Deployment Strategy

The construction of this project will be carried out in stages. The first stage is to build a data infrastructure that can uniformly collect metrics from PDUs, BMCs, GPUs, containers and business applications at various times to construct a time-series data warehouse with data spanning from seconds to minutes. The second stage builds a power-prediction model for the near future of servers, racks and clusters, and adds an abnormality-detection module. The third stage is a shadow control that offers suggested actions to the controller without directly issuing them, and policy stability is verified. The fourth stage is closed-loop control; power capping and DVFS will be employed in low-risk business pools first, then gradually extended to mixed-load clusters. The fifth stage adds forecasts for electricity prices, carbon intensity and renewable energy generation to switch the data centre from energy-saving operation to energy-carbon coordinated operation.

Operation and maintenance (O&M) of this system also require a clear division of authority and responsibility. The facilities team will be responsible for power supply and distribution, cooling and security thresholds; the cloud platform team will handle the scheduler, container runtime and task priority; the SRE team will establish SLA red lines and fault rollback; and the energy management team will set electricity pricing, carbon intensity and demand response strategies. Only when this information enters a unified control loop will server energy consumption optimisation be able to avoid being mere local energy saving and become an all-encompassing governance capability that addresses cost, reliability and carbon emissions.

5. Conclusion

This paper presents a technical framework for energy consumption monitoring and adaptive control of servers in large-scale data centers by integrating multi-granularity sensing, power consumption prediction, thermal constraint modelling and rolling optimisation control. Research shows that with the rapid development of high-density AI servers, relying solely on PUE or the total power meter for refined energy saving is no longer adequate; thus, chip-level power consumption, overall system temperature, rack power density, service SLA, and facility cooling status all need to be included in a single closed-loop control system. The power consumption model, thermal state model and multi-objective control function built in this paper can explain the formation mechanism of server energy consumption and guide control actions such as power capping, DVFS, task migration, sleep/wake-up and cooling coordination. Scenario simulations have shown that the proposed method can keep the SLA default rate within limits and reduce both total energy consumption and peak power; it is therefore feasible in practice. The three problems to be addressed in future research are: first, how to securely share and establish privacy boundaries for high-frequency telemetry data; second, how to develop more precise thermal models of liquid cooling and GPU clusters; and third, how to create cooperative optimisation mechanisms among data centres, power grids and carbon markets.

References

[1] *International Energy Agency. Energy and AI. Paris: IEA, 2025.*

[2] *Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A. B., Holecek, B., Koomey, J., Masanet, E., & Sartor, D. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, LBNL-2001637, 2024.*

[3] Liang, Q. (2026). Reconstruction of Commercial Building Space Reuse Mode Driven by Composite Business Types. International Journal of Engineering Advances, 3(1), 107-113.

[4] Gao, Y. (2026). Application of Delaware Corporate Law in Cross-Border M&A: Business Structures, Contractual Risk, and Case-Based Lessons. Economics and Management Innovation, 3(2), 128-136.

[5] Gao, Y. (2026). The Role and Practice of Delaware Law in Global Cross-border M&A Transactions. International Journal of Law, Policy & Society, 2(1), 38-46.

[6] Wang, N. (2026). Research on Evaluation and Optimization Models of Enterprise Resource Allocation Efficiency from a Data-driven Perspective. Advances in Computer and Communication, 7(1).

[7] Su, J. (2026). Research on Android Real-time Communication System Architecture and High-reliability Assurance Pathways Integrating AI-based Anomaly Detection Mechanisms. Engineering Advances, 6(2).

[8] Liang, Q. (2026). Research on the Renovation Design Path for Enhancing the Efficiency of Mixed-Use Office and Retail Spaces. European Journal of AI, Computing & Informatics, 2(2), 163-170.

[9] Liang, Q. (2026). How Architectural Design and Utility Infrastructure Impacts AI Supporting Campus and Drive Future Innovation, Operational Efficiency and Sustainable Advancement in Utility-Critical Environment. European Journal of Engineering and Technologies, 2(2), 71-77.

[10] Liu, Z. (2026). Research on Measuring User Behavior Response Differences Supported by Propensity Scoring Method. European Journal of AI, Computing & Informatics, 2(2), 47-53.

[11] Liu, Z. (2026). Research on Measuring User Behavior Response Differences Supported by Propensity Scoring Method. European Journal of AI, Computing & Informatics, 2(2), 47-53.

[12] Gao, Y. (2026). Governance Structures and Checks and Balances in Private Equity Funds: Practice, Problems, and Improvement Paths. Financial Economics Insights, 3(2), 82-91.

[13] Liang, Q. (2026). Research on Low-Energy Space Organization Methods in the Context of Building System Integration. Global Journal of Science & Innovation, 3(1), 9-15.

[14] Zhu, P. (2025, December). Construction of Multi-Scale Biostatistical Analysis Framework and its Application in Biomedical Signal Feature Recognition and Classification. In 2025 5th International Conference on Mobile Networks and Wireless Communications (ICMNWC) (pp. 1-7). IEEE.

[15] Chen, J. (2026). Elastic Scaling and Stability Assurance Mechanisms for Distributed Systems under High-Throughput Scenarios.

[16] Chang, C. W. (2026). Privacy Infrastructure Vulnerability Mining and Automated Framework Based on Multimodal AI. Procedia Computer Science, 279, 702-711.

[17] Wang, Y. (2026). Construction of Supply Chain Demand Forecasting Algorithm Based On Time Series Data Analysis and Improvement of Its Management Efficiency. Procedia Computer Science, 281, 484-493.

[18] Wang, Z. (2026). Mineral Commodity Time-Series Cycle Modeling and Feature Learning Algorithm Based On Improved Transformer. Procedia Computer Science, 281, 1347-1356.

[19] Chen, J. (2026). Construction of a Cloud-Native High-Performance Service Engineering System for Real-Time Decision-Making Platforms.

[20] Qian, X. (2026). Supply Chain Collaboration Mechanisms Driven by Demand Orchestration in Omni-Channel Retail Environments.