

Project Management Path and Key Technologies for Energy Efficiency Optimization in AI Data Centers

Chenjie Gao

Department of Computer Science and Engineering, University of California-Santa Cruz, Santa Cruz, 95064, CA, US

Keywords: Artificial intelligence data center; online large model inference; energy efficiency optimization; stage gate management; power profile; capacity control

Abstract: To address the uncertainties in heterogeneous data adaptation, power evidence boundaries, and deployment/promotion judgment of energy-saving solutions for online large-scale model inference resource pools, this study constructs a phased-gate access path for energy efficiency optimization in AI data centers. Based on real request trajectories from BurstGPT and publicly available H100 inference power profiles, the study employs token/configuration/power field adaptation review, profiled trajectory replay, risk-constrained capacity control, and distributed fault stress testing to compare fixed peak capacity, reactive scaling, direct deployment, risk-constrained direct deployment, and the phased-gate strategy. Results show an adaptation coverage rate of 63.02%. Under a capacity insufficiency limit of 1%, the phased-gate strategy achieves a relative energy saving of 16.55% with a capacity insufficiency rate of 0.54%, and a pilot-to-out-of-time energy saving deviation of 0.53 percentage points. The study demonstrates that the phased-gate path does not guarantee the highest energy saving rate, but it can identify incompatible scenarios and high-risk solutions, providing verifiable evidence for the piloting and promotion of energy-saving technologies for online inference resource pools.

1 Introduction

Artificial intelligence applications have gradually extended from offline training to online service scenarios such as intelligent question answering, code assistance, content generation and knowledge retrieval. Online large model inference is usually deployed in GPU resource pools, and the arrival of requests has obvious peak and valley variations. The length of the input prompt, the number of output tokens and the concurrency pressure also vary greatly. In order to avoid insufficient capacity during peak periods, data centers often reserve nodes according to high percentile load, but this method will cause nodes to be idle and basic power consumption to accumulate during low load periods. The energy pressure on the inference side has become a real problem in the operation and management of data centers [1].

Existing research has shown that inference energy consumption is not solely determined by the model parameter size. Output token length, hardware type, inference framework, batch settings, and

actual utilization all affect the unit inference cost [2]. The online inference process includes two stages: prefill and decode. The former is affected by input tokens and memory access, while the latter is related to output tokens, KV-Cache read/write, and batch strategy. Therefore, energy efficiency optimization cannot be judged solely based on the total power of the node or the energy saving rate of a single window. It also needs to simultaneously meet the operational boundaries such as request completion rate, capacity shortage ratio, peak queuing size, and node state adjustment frequency.

Existing research mainly focuses on three aspects: energy saving in cloud data centers, resource scheduling, and inference energy consumption estimation. Research on energy saving in cloud data centers emphasizes resource integration, dynamic scheduling, and software layer optimization [3]; research on data center sustainability further points out that energy efficiency evaluation needs to clarify the IT side, facility side, and life cycle boundary [4]; research on GPU workload scheduling focuses on the differences between training and inference tasks in terms of resource consumption, queue waiting, and scheduling granularity [5]. These studies provide a foundation, but there are still shortcomings from the perspective of engineering implementation.

Some studies directly select capacity schemes based on energy saving rates during the development phase, lacking independent pilot projects and out-of-time verification. Some simulation studies use manually set average power, service capacity, or task duration, failing to address the adaptation boundary between real requests and measured power profiles. GPU inference power consumption is affected by computing core utilization, memory read/write, prefill stage, decode steps, and cache state. If studies only use the number of requests per unit time or average power range for mapping, it will create logical jumps at the physical level.

Based on this, this study limits its research object to homogeneous GPU resource pools that support online large-scale model inference tasks, and investigates how energy efficiency optimization projects in artificial intelligence data centers can complete data review, technical access, virtual pilot testing, and promotion judgment before going live. This study does not claim to simulate millisecond-level GPU power consumption, nor does it present the results of virtual pilot testing as the results of real data center deployment. Instead, it establishes a reviewable project management path to ensure that energy-saving solutions are tested for load adaptation, power evidence, risk boundaries, and distribution discontinuities before going live.

2. Relevant Research and Problem Definition

Online inference energy consumption estimation can be broadly classified into three categories. Parameter scale or computational regression methods are convenient for cross-model comparison, but they are difficult to reflect the differences in output tokens, batches and GPU utilization; token-level energy consumption estimation is close to the inference process, but requires fine-grained logs; measured power profile methods can preserve hardware power fluctuations, but the range of application is strict and cannot be arbitrarily extrapolated. Luccioni et al. emphasized in their model environmental impact study that the system boundary and measurement scope must be clearly defined for running energy consumption estimation [6]. This study uses measured power profiles as the basis for project admission evidence, but does not extend the profile to request scenarios where fields cannot be verified.

From the perspective of data center management, energy-saving solutions must not only answer whether power can be saved, but also answer within which request range power can be saved, at what capacity risk power can be saved, and whether it is still acceptable after load changes. Digital service energy consumption is related to infrastructure, operating conditions and resource allocation[7], and ICT energy consumption accounting is also affected by boundary caliber[8].

Therefore, this study clearly distinguishes node-level IT power from the impact of the whole machine, network, cooling, power conversion and life cycle. The electronic waste problem caused by hardware manufacturing and equipment upgrades is also related to the expansion of AI infrastructure[9], but it is not included in the main experiment of this study.

The core issue of this study is not to propose the most complex scaling algorithm, but to determine whether an energy-saving solution meets the conditions for evaluation, pilot testing, and scalability before deployment. To this end, this study distinguishes phase gates from the ordinary training-validation-testing process. Phase gates are not merely about dividing the dataset; rather, they require each phase to generate clear management actions, technical evidence, and admission conclusions: M1 defines the scope of evaluable requests and scenarios where scalability is not possible; M2 generates candidate solutions and reasons for elimination; M3 makes a decision to approve, reject, or require manual review; and M4 provides scalability recommendations and risk boundaries.

3. Project Management Approach and Research Design

3.1 Data Sources and Underlying Facts

This study employs two types of heterogeneous public datasets. The request side uses BurstGPT, derived from real-world Azure OpenAI service logs, which reflects the temporal structure and burst characteristics of online load. The power side uses a public inference power profile based on the NVIDIA H100, grounded in the vLLM benchmark, with a power sampling resolution of 0.1 seconds. It is important to clarify that the two datasets originate from different operating environments and are not joint logs from the same enterprise or data center.

Therefore, this study does not interpret the simulation results as absolute energy savings for the BurstGPT source platform or any production data center. All energy consumption results in this study are expressed as profiled relative energy savings under a unified H100 power profile library, used to compare the risk and benefit differences of different capacity strategies based on the same power evidence. This processing ensures the verifiability of strategy ranking but does not extrapolate to untested hardware or load scenarios.

3.2 M1-M4 Stage Gate Paths and Strategy Grouping

This study constructs a phase gate path from M1 to M4. M1 reviews real load evidence and traceable power evidence; M2 compares candidate capacity strategies within the scope of the adaptation request; M3 uses an independent window to verify whether candidate schemes meet the admission criteria; and M4 conducts out-of-time validation and distributed fault pressure testing under fixed parameters.

Table 1. Management Path for AI Data Center Energy Efficiency Optimization Project

Stage	Management tasks	Technical movements	Admission results
M1	Project initiation and evidence review	Token, configuration, and power field adaptation	Scope of application and boundaries of non-promotion
M2	Feasibility assessment of candidate solutions	Risk Constraint Replay and Strategy Comparison	Candidate List and Reasons for Elimination
M3	Virtual pilot acceptance	Independent window testing, no parameter readjustment required.	Pass, Reject, or Manual Review
M4	Promotion Judgment and Review	Out-of-time verification and distribution fault testing	Promotion Recommendations and Risk Levels

As shown in Table 1, unlike the ordinary validation set used for parameter selection, the M3 stage in this study serves as the engineering admission function. Even if a candidate solution has a high energy saving rate during the development period, it will be rejected or enter manual review if the adaptation coverage, insufficient capacity ratio, completion rate, number of node adjustments, or stress test results do not meet the pre-registration boundaries, and will no longer automatically enter the M4 promotion judgment.

This study sets up five strategies. B1 is a fixed peak capacity strategy, configuring nodes according to the peak demand of the development phase adaptation requests. B2 is a reactive scaling strategy, adjusting nodes based on the current queue and load status. B3 is a direct deployment strategy to maximize energy saving during development, selecting the parameter with the highest energy saving rate during development. B4 is a risk-constrained direct deployment strategy, selecting the parameter with the highest energy saving rate from candidate parameters whose capacity shortage rate does not exceed 1% during development, but without independent acceptance testing by M3. Proposed is a phase-gate strategy, using the same capacity control technology system as B3 and B4, but the parameters must pass the M2 feasibility assessment and M3 virtual pilot acceptance testing. B4 is used to control for differences in risk preferences, avoiding the misinterpretation of risk aversion as a phase-gate effect.

4 Key Technologies and Methods

4.1 Token/Configuration/Power Field Adaptation Review

Phase M1 does not use the number of requests per unit time or coarse-grained average power range to establish a mapping. Adaptation conditions include: the input token length falls within the measured profile coverage range; the output token length or decode step count falls within the measured profile coverage range; the batch size, concurrency, or equivalent load level can be confirmed; the GPU model, inference framework, precision settings, and runtime configuration are traceable; the power profile sampling granularity can reflect the inference task execution process; requests or profiles with missing key fields must not be included in the main experiment.

Adaptation coverage is defined as:

$$Coverage = \frac{N_{compatible}}{N_{all}}$$

Here, $N_{compatible}$ represents the number of requests that meet the adaptation conditions for the Token, configuration, and power fields, and N_{all} represents the total number of requests included in the audit. Unadapted requests are categorized according to input or output length exceeding the range, inability to confirm concurrent configuration, insufficient power profile metadata, and GPU configuration mismatch. Unadapted requests are included in the technical adaptation report as M1 risk identification results. For business scenarios with long prompts, long outputs, or lacking corresponding power profiles, measured profiles should be supplemented before proceeding to the M2 evaluation.

4.2 Power Profile Library, Processing Boundaries, and Trajectory-Driven Playback

In the M2 phase, a configuration-level measured power profile library is established. Each profile unit must include at least the GPU model, inference framework, precision settings, input/output token range, concurrency or batch configuration, sampling granularity, power time series, and runtime. Average power profiles lacking token and configuration fields are not used for request-level power consumption mapping.

The total energy consumption of the node pool is accumulated from the actual power consumption over time:

$$E_{pool} = \sum_{j=1}^q \sum_{i=1}^{m_j} P_{j,i} \Delta t_{j,i}$$

Where is the power $P_{j,i}$ at the i -th j sampling point of the i - $\Delta t_{j,i}$ th profile, i and is the sampling interval. The energy consumption per completed request and the energy consumption per unit adaptation token are obtained by dividing the total energy consumption by the number of completed requests and the number of completed tokens, respectively.

The processing boundary for a single node C is not manually set, but is derived from the measured inference profile that has passed the adaptation review. Each profile unit needs to record the input/output token range, concurrency or batch settings, runtime, and the number of completed requests or equivalent token processing volume. This study uses the maximum stable number of completed requests or equivalent token processing volume for this configuration unit under conditions of no capacity shortage as the processing boundary; profiles lacking relevant fields are not used for C calculation.

the first k type of reasoning profile is represented as:

$$C_k = \frac{N_k^{complete}}{T_k}$$

Wherein, $N_k^{complete}$ represents k the number of requests or equivalent tokens completed by the i -th profile within the runtime, and T_k represents the profile duration. During node pool replay, different request types are preferentially matched to their corresponding values C_k ; cross-configuration average processing capacity must not be used to replace verifiable processing boundaries.

4.3 Risk-Constrained Capacity Control and Stress Testing

This study chose prediction interval and hysteresis control not because of their highest algorithmic complexity, but because their structural parameters are interpretable, acceptance boundaries are clear, and failure causes are traceable, making them suitable for pre-launch assessment of energy-saving projects. Let the future window load forecast be $\hat{D}_{t+h}$, the rolling error standard deviation be σ_{t+h} , and the upper bound of demand be:

$$U_{t+h} = \hat{D}_{t+h} + z\sigma_{t+h}$$

The control process sets expansion thresholds, reduction thresholds, and cooldown times. A node is added when two consecutive windows reach the expansion threshold, and a node is released when three consecutive windows fall below the reduction threshold, after which it enters the shortest cooldown period. This rule is used to reduce frequent state switching caused by short-term fluctuations.

Publicly available data does not provide request-specific KV-Cache hit rates or cache reuse status; therefore, this study does not treat these as independent control variables. To avoid parameterizing unobservable factors, this study only uses experimental profiles under the same framework, similar input/output token ranges, and the same concurrency or batch conditions, and considers cache differences as part of the internal fluctuations and adaptation errors within the profile. This approach cannot replace real cache modeling, but it avoids assumptions without field support.

Phase M3 employs an acceptance rule that requires all conditions to be met simultaneously. This study sets the capacity shortage ratio cap at 1%, not as a uniform standard for all data centers, but as a strict risk boundary for scenarios involving paid APIs, online interactions, and high-availability inference services. In these types of services, request rejection, long queuing times, or inference failures can impact SLAs, user retention, and service reputation, thus necessitating the inclusion of the capacity shortage ratio as an acceptance criterion. If the operator is operating in low-priority batch inference or deferred processing scenarios, this boundary can be relaxed, and a more energy-efficient but riskier strategy may be chosen.

In addition to verification based on natural time, the M4 phase includes four types of distributed fault pressure tests: input token right shift, output token right shift, burst window extension, and trough disappearance. Energy saving deviation and capacity insufficiency deviation are defined as follows:

$$D_E = |\Delta E^{pilot} - \Delta E^{holdout}|$$

$$D_C = |Breach^{pilot} - Breach^{holdout}|$$

Statistical analysis employed moving block bootstrap, paired permutation test, and Holm correction.

5. Research Results

5.1 M1 Compatibility Review Results

After adapting the token length, output length, concurrency configuration, GPU configuration, inference framework, and power sampling granularity, the adaptation coverage rate is 63.02%. Among them, continuous time windows are maintained during the development period, pilot period, and out-of-time verification period to ensure that there is no time overlap between M2, M3, and M4.

Unsuitable requests mainly include those with input or output token lengths exceeding the range, inability to confirm concurrency or batch configurations, insufficient power profile metadata, and mismatched GPU or inference framework configurations. Subsequent M2, M3, and M4 checks will only be performed on a subset of requests that passed the M1 review; unsuitable requests will not be included in the calculation of energy saving rate, insufficient capacity ratio, and completion rate. The M1 compatibility review results are shown in Table 2.

Table 2 M1 Compatibility Review Results

Index	Numerical values
Compatibility coverage /%	63.02
Input or output token length exceeds range /%	48.62
Concurrency or batch configuration cannot be confirmed /%	21.75
Insufficient power profile metadata /%	18.35
GPU or inference framework configuration mismatch /%	11.28

5.2 Screening Results of M2 and M3 Stage Gates

The power profile library is set to 12 configuration-level measured inference profiles. During development, 36 candidate parameters were generated. In the M2 phase, 24 groups passed the preliminary feasibility assessment, and 12 groups were eliminated. Among them, 5 groups were eliminated because the insufficient capacity request ratio exceeded 1.5%, 3 groups were eliminated because the energy saving was less than 5%, 2 groups were eliminated because the number of node

state adjustments exceeded the limit, and 2 groups were eliminated because the queue peak value was too high under pressure pre-screening.

During the development period, the cumulative profiled energy consumption of B1 was 152.48 MWh, with a capacity shortfall rate of 0.22%; the relative energy saving rate of B2 profiled energy saving was 16.76%, with a capacity shortfall rate of 2.61%; the energy saving rate of B3 was 21.52%, with a capacity shortfall rate of 1.03%; and the energy saving rate of B4 under the 1% risk constraint during the development period was 19.43%, with a capacity shortfall rate of 0.84%. The cumulative profiled energy consumption of the stage gate candidate schemes was 123.41-129.86 MWh, with a capacity shortfall rate of 0.41%-0.98%.

Twenty-four candidate schemes were selected for M3. During the pilot period, eight schemes simultaneously met the criteria for energy saving, capacity shortage, completion rate, peak queue, and number of node adjustments, while the remaining 16 schemes failed acceptance. Scheme B3 had a capacity shortage request rate of 1.31% during the pilot period, failing to meet the strict capacity boundary. Scheme B4 had a capacity shortage rate of 0.94% during the pilot period, but the peak queue in the stress pre-check was close to the manual review threshold. The final stage gate group selected a prediction window of 60 minutes, a prediction upper bound confidence level of 0.90, an expansion threshold of 0.82, a reduction threshold of 0.58, a minimum number of retained nodes of 2, and a cooling time of 10 minutes. This combination achieved a profiled relative energy saving rate of 17.08%, a capacity shortage rate of 0.56%, and a completion rate of 99.44% during the pilot period, passing the M3 acceptance. The selection results for candidate schemes M2 and M3 are shown in Table 3.

Table 3 The selection results for candidate schemes M2 and M3

Plan	Relative energy saving rate during development period /%	Pilot period capacity insufficient /%	Pilot period completion rate / %	Number of node adjustments	M3 conclusion
B1	—	0.23	99.77	0	Baseline
B2	16.76	2.66	97.34	55	reject
B3	21.52	1.31	98.69	41	reject
B4	19.43	0.94	99.06	29	Manual review
Proposed	17.08	0.56	99.44	18	pass

5.3 M4 out-of-time verification results

During the out-of-time verification period, the cumulative profiled energy consumption of B1 was 74.62 MWh, with a capacity shortfall rate of 0.21%; the energy saving rate of B2 was 17.03%, with a capacity shortfall rate of 2.73%; the energy saving rate of B3 was 19.80%, with a capacity shortfall rate of 1.38%; the energy saving rate of B4 was 17.78%, with a capacity shortfall rate of 0.96%; the energy saving rate of the stage gate group was 16.55%, with a capacity shortfall rate of 0.54%, and the completion rate was 99.46%.

Both B4 and the stage gate group are below the 1% capacity deficit boundary, but B4 is closer to the upper limit. Their pilot-out-of-time energy saving deviations are 1.62 and 0.53 percentage points, respectively, and their capacity deficit deviations are 0.38 and 0.02 percentage points, respectively. This result indicates that, under consistent risk constraints, the stage gate path reduces the pilot-out-of-time deviation. Out-of-time validation results for M4 are shown in Table 4.

Table 4 Out-of-time validation results for M4

Plan	Cumulative profiled energy consumption/mwh	Relative energy saving rate /%	Insufficient capacity percentage / %	Completion rate / %	Number of node adjustments
B1	74.62	—	0.21	99.79	0
B2	61.91	17.03	2.73	97.27	62
B3	59.84	19.80	1.38	98.62	44
B4	61.35	17.78	0.96	99.04	31
Proposed	62.27	16.55	0.54	99.46	19

5.4 Results of Distributed Fault Pressure Tests

As shown in Figure 1, in the scenario of right-shifting input tokens, the capacity insufficiency rate for B3 is 2.26%, for B4 it is 1.66%, and for the stage gate group it is 0.91%. In the scenario of right-shifting output tokens, the capacity insufficiency rate for B3 is 2.48%, for B4 it is 1.84%, and for the stage gate group it is 0.98%. In the scenario of extended burst window, the capacity insufficiency rate for B3 rises to 2.91%, for B4 it is 2.06%, and for the stage gate group it is 1.12%. In the scenario of trough disappearance, the capacity insufficiency rate for B3 is 1.72%, for B4 it is 1.21%, and for the stage gate group it is 0.74%.

Stress tests show that the phased gate group does not meet the 1% boundary in all impact scenarios, but its capacity underperformance ratio and the number of node adjustments are generally lower than those of B3 and B4. The phased gate path is more suitable for promotion and judgment under strict service boundaries, rather than replacing high-energy-saving solutions under all risk preferences.

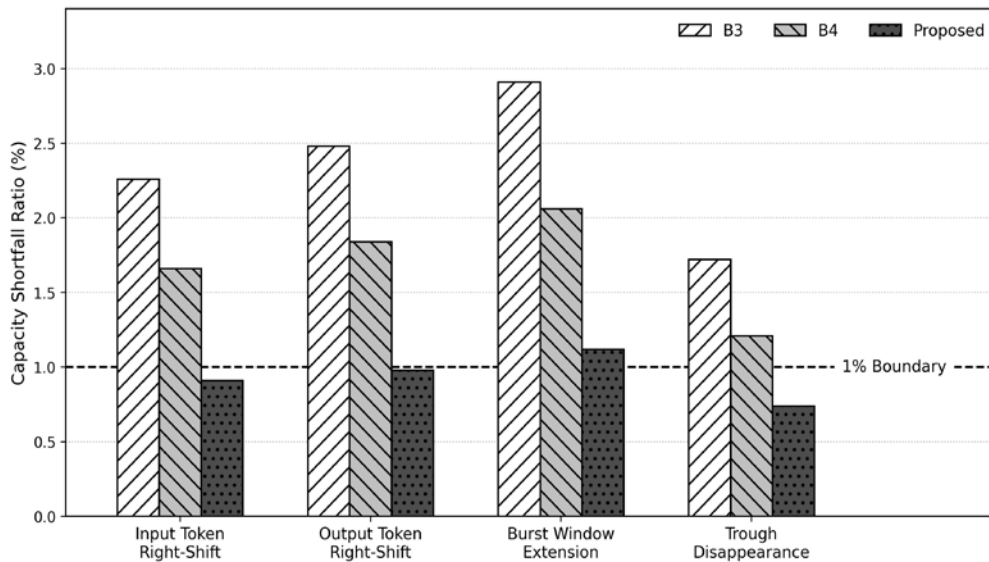


Figure 1. Comparison of the proportion of insufficient capacity under different strategies in distributed fault pressure testing.

6. Discussion

This study constructs an adaptation review system based on tokens, output length, concurrency configuration, GPU configuration, and power sampling granularity, avoiding the problem of directly mapping GPU dynamic power consumption to the number of requests or average power

range. The adaptation coverage rate is 63.02%, indicating only partial overlap between the actual request trajectory and the publicly available power profile. This does not indicate a research failure, but rather that before energy-saving projects go live, it is essential to answer which requests can be supported by existing power evidence. For data centers with a high proportion of long Prompt summaries, long contextual questions, or complex inferences, long tokens and high-concurrency power profiles should be prioritized for supplementation [10].

Since the BurstGPT and H100 inference power profiles come from different operating environments, all energy consumption results in this study are presented as profiled relative results under a unified H100 inference power profile library. These results can be used to compare capacity policy differences based on the same power evidence, but cannot be interpreted as absolute energy savings on Azure source platforms or real production data centers. This approach avoids piecing together heterogeneous public data into a single platform's measured logs and also preserves the verifiability of relative policy evaluations [11].

B3 achieved higher energy savings than the phased gate group in out-of-time validation, but also had a higher undercapacity rate. B4, using the same 1% risk boundary as Proposed during development, showed out-of-time results that its profiled energy savings were still higher than the phased gate group, but it had higher pilot-out-of-time bias and stress test risks. Therefore, this study does not claim that the phased gate strategy is optimal for all business objectives. If operators can accept a higher undercapacity rate, B3 or B4 may be more attractive for energy savings; if the business involves paid APIs, online interactions, or high-availability inference services, the phased gate group is more suitable for inclusion in the rollout decision in terms of stability and boundary explainability [12].

M3 should not be interpreted as a standard validation set selection. Standard model selection focuses on parameter performance, while stage gate acceptance also includes adaptation coverage, risk boundaries, completion rate, node adjustments, stress testing, incompatibility risk logs, and manual review conditions. Even if a solution has a high energy efficiency during development, it may still be rejected due to insufficient capacity exceeding limits or instability under distribution shocks[13].

This study still has limitations. The publicly available power profile cannot fully reflect the KV-Cache hit rate, millisecond-level computing unit utilization, and memory access fluctuations. Although this study limits the cache differences to the internal fluctuations and adaptation errors of the same configuration profile, it cannot replace the per-request cache state log. The node pool results are a virtual pilot and do not equal real data center deployment; node-level IT power cannot replace the energy consumption of the whole machine, network, power supply, and cooling. The environmental impact of AI servers is also affected by infrastructure efficiency, energy structure, and hardware lifecycle [14]. Further verification using logs from the same platform is needed for real multi-node verification.

7. Conclusion

This study focuses on the energy efficiency optimization project of online inference resource pools in artificial intelligence data centers, constructing M1-M4 stage gate paths for pre-deployment technical access. The study combines real request structures with measured GPU inference power profiles, but does not use coarse-grained request counts or average power ranges for physical mapping. Instead, it uses token length, output length, concurrency configuration, GPU type, inference framework, and power sampling granularity as adaptation conditions.

The results show that the phased gate strategy does not guarantee the highest energy saving rate. Within the natural time out-of-window period, the profiled energy saving rates of B3 and B4 are

higher than those of the phased gate group; however, under strict capacity risk boundaries and distribution fault stress tests, the phased gate group exhibits a lower undercapacity ratio, smaller pilot-out-of-time bias, and a clearer generalization boundary. The contribution of this study is that it advances the energy efficiency optimization of AI data centers from a single strategy comparison to a pre-launch admission, risk screening, and generalization judgment process. This approach does not replace specific capacity algorithms but provides an auditable, rejectable, and reviewable engineering decision-making mechanism for energy-saving technology solutions.

References

- [1] De Vries A. *The growing energy footprint of artificial intelligence*[J]. *Joule*, 2023, 7(10): 2191-2194. DOI: 10.1016/j.joule.2023.09.004.
- [2] Desislavov R, Martínez-Plumed F, Hernández-Orallo J. *Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning*[J]. *Sustainable Computing: Informatics and Systems*, 2023, 38: 100857. DOI: 10.1016/j.suscom.2023.100857.
- [3] Katal A, Dahiya S, Choudhury T. *Energy efficiency in cloud computing data centers: A survey on software technologies*[J]. *Cluster Computing*, 2023, 26(3): 1845-1875. DOI: 10.1007/s10586-022-03713-0.
- [4] Wang, B. (2025). *Methods of Load Optimization for Computer Systems Based on Physical Principles*.
- [5] Liu, X. (2026). *Construction of Consumer Data Privacy Protection System Based On Blockchain Technology*. *Procedia Computer Science*, 282, 2082-2091.
- [6] Jing, X. (2025, November). *In-Depth Analysis of the Optimization Decision Model of Financial Investment Portfolio Based on Machine Learning Algorithms in Python Scikit-Learn Library*. In *International Conference on Computational Technologies for Research in Data Analytics* (pp. 130-140). Cham: Springer Nature Switzerland.
- [7] Yu, X. (2026). *Exploration of Multi-Channel Conversion Path Optimization Methods Based on A/B Testing*.
- [8] Liu, X. (2026). *Research on the Application of Artificial Intelligence in Consumer Privacy Data Protection*. *Procedia Computer Science*, 279, 956-965.
- [9] Guo, X. (2025, April). *Research on Financial Trading Algorithms Based on Deep Reinforcement Learning*. In *2025 IEEE 3rd International Conference on Control, Electronics and Computer Technology (ICCECT)* (pp. 1711-1715). IEEE.
- [10] Guo, X. (2025, March). *Research on Blockchain-Based Financial AI Algorithm Integration Methods and Systems*. In *2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE)* (pp. 816-821). IEEE.
- [11] Yuan, Y. (2026). *Research on Memory Management and Dynamic Reasoning Methods for Intelligent Agents Based on Large Language Models*.
- [12] Yang, Y. (2026). *Causal Inference-based Identification of Incremental Effects in Digital Advertising and Optimization Pathways for Resource Allocation*. *Advances in Computer and Communication*, 7(2).
- [13] Sun, L. (2026). *Consistency Optimization Strategies for Distributed Advertising Systems Under Cross-regional Deployment*. *Advances in Computer and Communication*, 7(2).
- [14] Han, W. (2025, November). *An Empirical Study on Building a High-precision Forecasting System for Financial Market Trends Based on the PyTorch Deep Learning Framework*. In *International Conference on Computational Technologies for Research in Data Analytics* (pp. 518-528). Cham: Springer Nature Switzerland.