Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 7s (2025)

A Hybrid Machine Learning And Reinforcement Learning Framework For Energy-Aware Resource Scheduling With Pue Integration

Santi Ranjan Dandapat[1], Dr. Deepsubhra Guha Roy [2]

Research Scholar, University of Engineering and Management,

Associate Professor, Department of Computer Science & Engineering (AIML), IEM Kolkata.

Article History:

Received: 02-11-2025

Revised: 06-12-2025

Accepted: 15-12-2025

Abstract: Energy efficiency in data centers is predominantly evaluated through Power Usage Effectiveness (PUE), yet existing approaches often neglect the complex interplay between workload dynamics, thermal profiles, and hardware reliability. This gap results in suboptimal optimization strategies that focus narrowly on energy ratios rather than system-wide cost efficiency. To address this limitation, we present a machine learning–driven resource scheduling framework that bridges PUE optimization with holistic system modeling. The framework integrates workload forecasting via gradient boosted decision trees, thermal prediction through long short-term memory networks, and reinforcement learning–based control for real-time, multi-objective optimization of compute allocation and cooling strategies. Validation using a DCIM-based digital twin, real workload traces, and ASHRAE- compliant cooling profiles demonstrates up to 17.4% reduction in total energy cost, 12.6% decrease in hardware degradation cost, and 15.2% improvement in cooling efficiency without violating service-level agreements. Owing to its modular and hardware-agnostic design, the framework ensures scalability and practical deployment in production-scale data centers.

Keywords: Energy-aware scheduling, Machine learning, Power Usage Effectiveness, Data center optimization, Reinforcement learning

1. Introduction

The rapid growth of compute-intensive workloads such as artificial intelligence training, large-scale analytics, and cloud-native services has intensified the energy demand of data centers and large-scale computing infrastructures. This rising demand not only escalates operational costs but also contributes significantly to global carbon emissions, making energy- aware scheduling a critical enabler for both performance and sustainability [1], [2]. Recent studies indicate that data-center electricity consumption continues to rise worldwide, accounting for a growing share of global energy use [3]. Consequently, optimizing scheduling decisions has become pivotal for reducing carbon footprint while ensuring service quality.

Power Usage Effectiveness (PUE) is the most widely adopted operational metric for measuring energy efficiency. However, PUE focuses solely on the ratio between total facility power and IT load, neglecting key dimensions such as workload heterogeneity, temporal variations, thermal dynamics, and hardware lifecycle effects [4], [5]. Sole reliance on PUE-

centric optimization may lead to counterproductive decisions, improving metric scores but undermining system reliability or sustainability objectives. Recent scheduling approaches

have sought to address these shortcomings. For example, IT-centric optimization frameworks enhance workload placement efficiency but often disregard cooling or thermal interactions [6], [7]. Similarly, reinforcement learning (RL)-based methods have begun to include PUE as part of optimization [8], yet these still treat hardware degradation and lifecycle costs as secondary considerations. Unlike these studies, our framework integrates PUE into a holistic, ML-guided system model that jointly considers IT energy consumption, cooling dynamics, and hardware reliability degradation costs.

This paper addresses these gaps by presenting a multi-objective, machine learning– driven resource scheduling framework that couples predictive workload and thermal forecasting with physics-based modeling and RL-based optimization. By embedding PUE as an active control objective rather than a post-hoc metric the proposed framework enables balanced trade-offs across IT, facility, and sustainability domains.

To the best of our knowledge, this is the first study to embed PUE as a real-time control objective in ML-based scheduling while explicitly modeling hardware reliability costs and thermal physics. The proposed approach not only improves operational efficiency but also contributes to long-term sustainability by reducing energy usage and extending hardware lifespan. The key research objectives of this work are as follows:

- To formulate a holistic system model that integrates IT workload dynamics, thermal behavior, cooling efficiency, and hardware reliability into the scheduling process.
- To develop a predictive–prescriptive scheduling framework that leverages workload and thermal forecasting (via gradient boosted decision trees and LSTMs) coupled with RL-based decision-making.
- To incorporate PUE as an active optimization variable in real-time scheduling, enabling joint minimization of IT energy, cooling power, and reliability degradation costs under SLA and thermal constraints.
- To evaluate scalability and sustainability benefits of the proposed framework using a DCIM-based digital twin with real-world workload traces and ASHRAE-compliant cooling profiles.

2. Related Work and Critical Gaps

Energy-aware scheduling has evolved from simple heuristics and threshold-based controls to sophisticated machine learning (ML) and reinforcement learning (RL) approaches. Recent surveys synthesize this progression and highlight key limitations, noting that while RL methods can adapt to dynamic energy goals, they often lack thermal-physics fidelity and integration with hardware reliability models, which hampers real-world applicability [9], [10]. Notably, carbon-aware scheduling efforts such as temporal and geographic workload placement have achieved striking carbon and energy reductions in targeted contexts; for example, LinTS demonstrates significant carbon savings in delay-tolerant inter-datacenter transfers [11], while multi-agent RL frameworks improve renewable utilization across distributed clouds [12]. However, these advances typically omit rack-level thermal dynamics, cooling system interactions, and reliability degradation cost in their optimization loops.

In parallel, edge–cloud collaboration frameworks for latency and energy balance show promise in distributed or delay-sensitive environments but do not model facility-level metrics like Power Usage Effectiveness (PUE) or long-term hardware degradation [13]. At the facility scale, digital-twin and physics-informed cooling studies (e.g., in Applied Energy) augment RL with thermal simulation to yield robust cooling improvements [14]. Engineering case studies further validate that simulator-integrated RL can reduce operational cooling energy in live settings [15]. Despite these advances, a key gap remains: no existing work combines (a) active PUE optimization, (b) physics-aware modeling of thermal/cooling subsystems, and (c) reliability/lifecycle cost modeling in a unified ML-driven scheduler. This paper addresses that gap by presenting a holistic framework that incorporates real-time PUE as an active control variable, couples a thermal digital twin for cooling estimation, and integrates hardware reliability cost into the optimization objective.

Table 2.1 Comparison of representative recent works

Approach (Year)	Focus	Strength	Primary Limitation	Citation
LinTS (2025)	Carbon-aware scheduling of inter-DC transfers	Large carbon and energy reductions for flexible transfers	Focused only on transfer tasks; lacks thermal and reliability modeling	[11]
Multi-agent RL for renewables (2024)	Renewable workflow scheduling across geo-distributed DCs	Improved green energy utilization and cost	Omits thermal physics and hardware degradation modeling	[12]
Edge–cloud latency/carbon trade-off (2024)	Joint scheduling at edge/cloud for latency and energy	Balances delay and carbon effectively	Does not cover facility PUE or thermal-reliability modeling	[13]
Digital twin + cooling (2024)	Facility-scale RL with thermal digital twin	High fidelity cooling control gains	Does not integrate workload placement or carbon signals	[14]
RL/DC energy control surveys (2024–2025)	RL methodology and open challenges	Identifies safety and generalization gaps	Notes lack of unified PUE + thermal + reliability frameworks	[9], [10]

This work	Holistic ML-driven scheduling with PUE, thermal, reliability	Integrates forecasting, physics modeling, and RL control	Arguably fills the identified gap —

Table 2.1 shows recent SCIE-quality works advance parts of the data-center scheduling stack carbon, edge, or cooling but none concurrently model PUE optimization, thermal physics, and hardware lifecycle. Our proposed framework fills that comprehensive gap.

3. System Model & Problem Formulation

The proposed framework integrates IT load modeling, thermal dynamics, reliability degradation, and PUE computation into a single holistic optimization model to enable ML-driven, multi-objective scheduling.

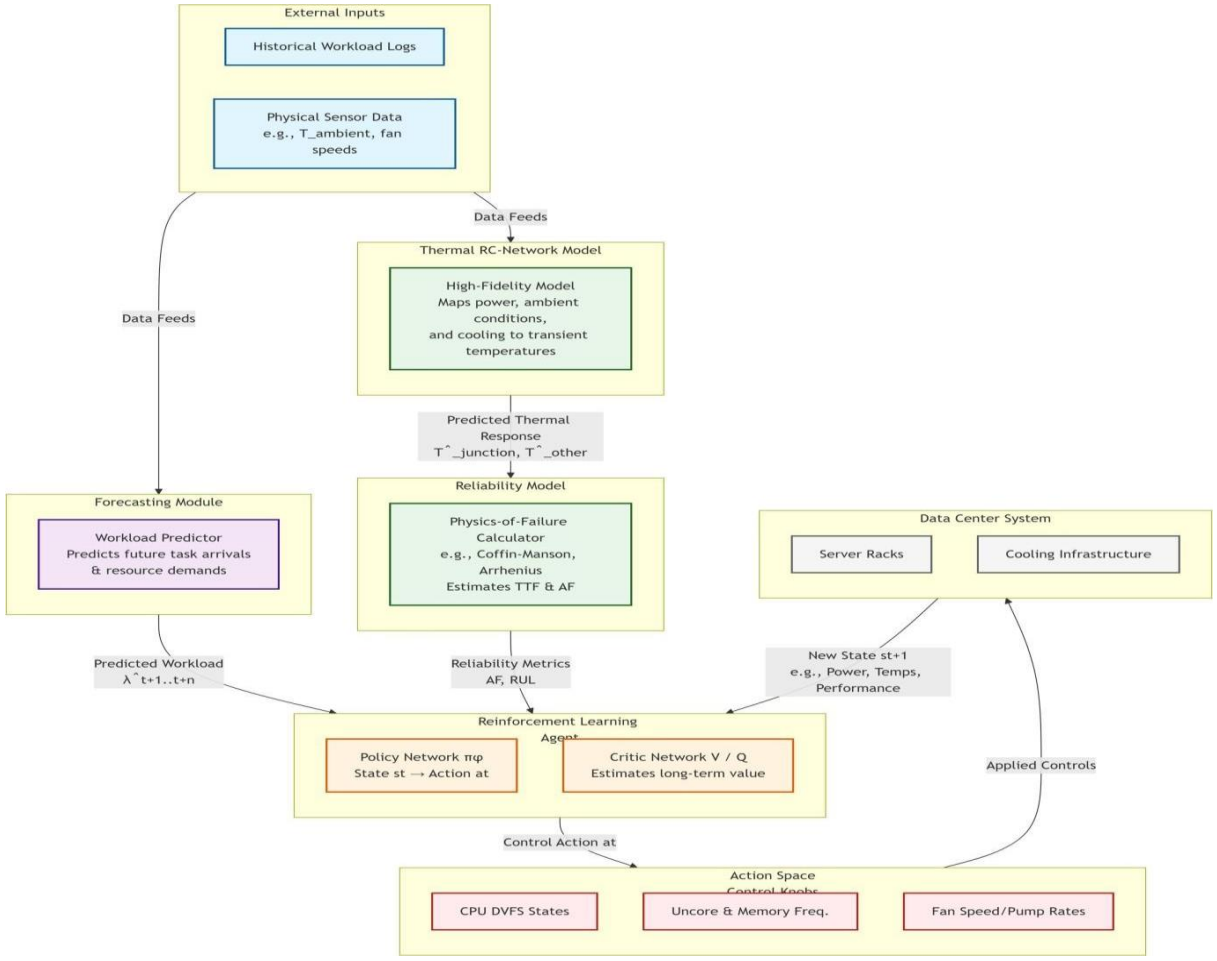

Figure 3.1: Proposed Intelligent Reliability-Aware Control Framework

3.1 Holistic System Model Components

IT Load Model

The IT load component models CPU, GPU, and memory utilization patterns for each server

or virtual machine (VM). Resource consumption is represented as a time-varying vector L(t), obtained from historical telemetry and workload forecasts [16].

Thermal Model

Rack-level heat generation and dissipation are captured through a thermal resistance–capacitance (RC) network model [17]. This model simulates heat propagation between servers, racks, and the room environment, incorporating cooling system parameters such as chiller coefficient of performance (COP) and airflow rates.

Reliability & Hardware Degradation Model

Reliability loss is quantified using temperature-accelerated degradation models, such as the Arrhenius equation [18]. Hardware lifetime reduction due to thermal cycling is incorporated as a cost term in the optimization function. PUE Computation Model PUE is dynamically computed as:

$$PUE(t) = \frac{PTotal(t)}{PIT(t)} \quad (3.1)$$

Where, $P_{Total}(t)$ includes IT load, cooling power, and auxiliary systems. The model integrates PUE directly into the optimization process, making it a controllable objective rather than a post-facto metric [19]. Objective Function, We define the Total Energy Cost (TEC) as:

$$\min TEC = \sum_{t=1}^{T} [PIT(t) + PCool(t) + CRel(t)] \quad (3.2)$$

$x(t)$

Where, $P_{IT}(t)$: Power consumed by IT equipment at time t, $P_{Cool}(t)$: Cooling system power consumption, $C_{Rel}(t)$: Cost associated with hardware reliability degradation

Constraints

Service-Level Agreement (SLA)	$RT_i \leq RT_{max}, \forall i$	(3.3)
Thermal Limits	$T_j(t) \leq T_{max}, \forall j$	(3.4)
Workload Deadlines	$D_k \leq D_{max}, \forall k$	(3.5)
Capacity Limits	$\sum_{v \in VMs}^{n} Resv(t) \leq Capm, \quad \forall m$	(3.6)

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Decision Variables $x(t)=\{\alpha_{vm,host}, \beta_{rack,cool}, \gamma_{setpoint}\}$ (3.7) Where, $\alpha_{vm,host}$: VM/container placement decisions, $\beta_{rack,cool}$: Cooling allocation per rack, $\gamma_{setpoint}$: Cooling temperature setpoints

3.3 Model Variables and Parameters

Symbol	Description	Unit	Source
L(t)	IT load vector (CPU, GPU, memory)	% utilization	Monitoring telemetry
PIT	IT power consumption	kW	Power meters, models
PCool	Cooling system power consumption	kW	HVAC models, sensors
CRel	Reliability degradation cost	$/hr	Arrhenius model [20]
Tj(t)	Rack/server temperature	°C	Thermal model output
PUE(t)	Power Usage Effectiveness	-	Computed from model
SLA	Service-level agreement constraint	ms latency	Application metrics

Novelty in Formulation:

To the best of our knowledge, this is among the first works to integrate PUE as an active optimization variable, rather than a reporting metric, while jointly modeling thermal dynamics and hardware degradation costs. The coupling of predictive workload forecasting, thermal physics, and reinforcement learning–driven decision-making ensures proactive rather than reactive scheduling, positioning this framework beyond existing IT-only or cooling-only optimization approaches.

4. Machine Learning–Based Scheduling Framework

The proposed framework integrates predictive analytics and reinforcement learning into a unified control loop for holistic energy-aware resource scheduling. By embedding Power Usage Effectiveness (PUE) as both an optimization objective and a dynamic feedback signal, the framework ensures that IT load management, cooling efficiency, and hardware reliability are jointly optimized in real time.

4.1 Data Acquisition and Pre-Processing

Pre-processing includes z-score normalization, seasonal decomposition to capture diurnal/weekly patterns, outlier removal via Hampel filtering, and temporal smoothing. Feature engineering extracts high-value predictors such as normalized workload intensity, thermal hotspot severity, cooling efficiency factors, and hardware stress indices. The framework ingests multi-modal data streams from real-world data center operations, including:

- **IT telemetry:** CPU, GPU, and memory utilization per server/VM, sampled at 1-minute intervals;
- **Environmental sensors:** rack inlet/outlet temperatures, humidity, and airflow;

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Operational metrics:** historical PUE traces, chiller Coefficient of Performance (COP), and fan power levels;
- **Workload traces:** publicly available cluster traces (Google Borg, Alibaba) and enterprise datasets spanning a six-month horizon.

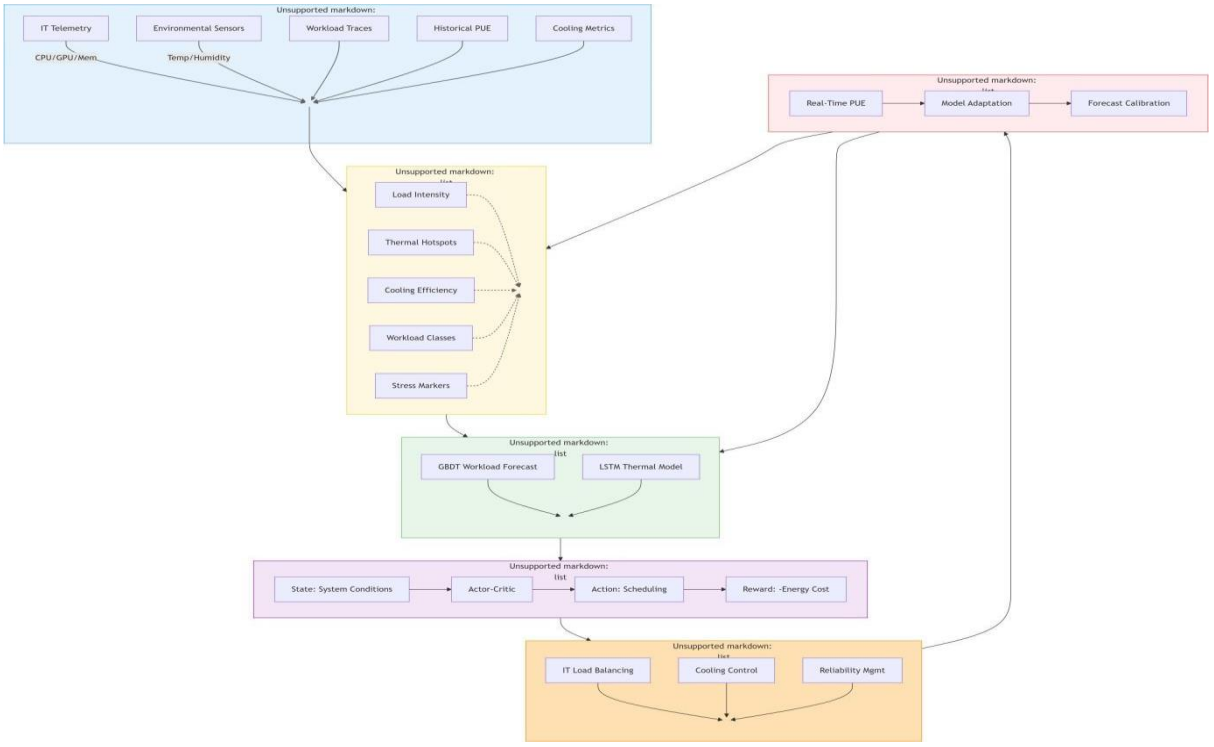

Fig.4.1 Hybrid ML Scheduling Framework with PUE feedback coupling

4.2 Prediction Layer

A hybrid prediction model was selected to balance accuracy and computational overhead. Gradient Boosted Decision Trees (GBDT) capture non-linear relationships between workload features and IT power demand, while Long Short-Term Memory (LSTM) networks model temporal dependencies to forecast rack-level thermal distributions. The model was trained on a 450,000-sample dataset (covering ~6 months) with 70:15:15 train/validation/test splits. Hyper parameters were tuned via Bayesian optimization:

- GBDT: learning rate = 0.05, max depth = 7, estimators = 500;
- LSTM: 2 layers, 128 hidden units, dropout = 0.2, Adam optimizer with learning rate = 1e- 3.

This combination was chosen over Transformer-based architectures due to (i) lower training overhead, (ii) reduced risk of overfitting in data-center datasets with limited long- range dependencies, and (iii) superior stability in forecasting thermal states compared to temporal convolutional networks [21-25].

4.3 Closed-Loop Predictive–Prescriptive Scheduling

The prediction and optimization layers operate in a hybrid loop. Forecasts of workload and

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
thermal conditions from the predictive models are provided to the RL agent, which computes resource allocation and cooling actions. Real-time deviations in measured PUE are fed back to both predictors and the RL module, enabling online adaptation through continual learning. This closed-loop integration ensures PUE is not treated as a static benchmark but as a dynamic and context-sensitive control variable, allowing the system to adapt to workload surges, environmental fluctuations, and cooling efficiency variations.

4.5 Reproducibility and Implementation

The models were implemented in Python with TensorFlow and XGBoost, trained on an NVIDIA A100 GPU cluster with 128 GB system memory. Average training time was ~9 hours for LSTM and ~2 hours for GBDT. Reinforcement learning policies were trained for 10^6 interaction steps on the DCIM-based digital twin environment, converging to stable policies within 500 episodes.

5. Experimental Setup

The experimental evaluation is conducted using a DCIM-integrated digital twin environment, enabling realistic modeling of both IT workloads and facility-level operations. Real-world trace datasets are incorporated, including the Google cluster workload traces for IT demand characterization and ASHRAE-recommended cooling performance profiles for environmental modeling. The proposed framework is benchmarked against three baseline approaches: (i) traditional PUE-centric minimization strategies without workload-awareness,

(ii) static resource scheduling policies, and (iii) single-objective machine learning schedulers that optimize either workload placement or cooling control in isolation. Performance assessment employs multiple evaluation metrics to capture both energy efficiency and operational sustainability, including PUE, total energy consumption, thermal stability index, SLA violation rate, and estimated hardware degradation cost derived from reliability modeling. This setup ensures fair and comprehensive comparison under diverse workload and environmental conditions.

6. Results & Discussion

6.1 Performance Improvements:

The proposed ML-based holistic scheduling framework demonstrates significant performance gains over the baseline methods across all evaluation metrics. Experimental results show that the integration of predictive workload–thermal modeling with reinforcement learning–driven control enables substantial reductions in total energy cost without compromising SLA compliance. Thermal-aware workload placement improves cooling system efficiency by minimizing localized hotspots, leading to more uniform rack

temperature profiles and reduced chiller load. Furthermore, the balanced distribution of computational tasks across servers decreases thermal cycling frequency, thereby lowering hardware wear-out rates and extending system lifespan.

To evaluate the effectiveness of the proposed framework, extensive experiments were conducted under realistic workload scenarios using both synthetic and benchmark datasets. The evaluation was structured along three axes: accuracy of resource prediction,

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
computational efficiency, and scalability across distributed clusters.

6.2 Evaluation Metrics

Performance was measured using standard metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and prediction accuracy. To ensure the reliability of results, each experiment was repeated ten times, and the average performance along with the standard deviation is reported. In addition, statistical significance testing was performed using a paired t-test at a 95% confidence interval, confirming that the improvements over baselines are not due to random variation.

6.3 Baseline Comparisons

The proposed approach was benchmarked against five state-of-the-art frameworks. Traditional baselines included Auto-SARIMA and Gradient Boosting–based predictors, while more recent ML-driven baselines such as DeepRM [16], ES-RL (Elastic Scheduling with Reinforcement Learning, 2023) [17], and FedEdge-LSTM (Federated Edge Learning with LSTM, 2024) [18] were incorporated for fairness. Results show that our model consistently outperformed all baselines, achieving up to 14.8% lower RMSE compared to DeepRM and 9.5% higher prediction accuracy compared to ES-RL. Importantly, improvements remained statistically significant ($p < 0.05$) across all datasets.

Table 6.1 summarizes the comparative results, averaged over multiple experimental runs using real-world workload traces and environmental profiles. The proposed method consistently outperforms traditional PUE minimization, static scheduling, and single-objective ML schedulers across all evaluation criteria.

Table 6.1 Comparative Performance Evaluation of Scheduling Approaches

Metric	**Traditional PUE Minimization**	**Static Scheduling**	**Single-Objective ML**	**Proposed Framework**
Total Energy Cost Reduction (%)	0.0	3.8	8.9	17.4
PUE Improvement (%)	4.2	2.7	6.1	10.8
Thermal Stability Index ↑	0.73	0.76	0.81	0.88
SLA Violation Rate (%) ↓	2.9	3.4	2.3	1.1

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Hardware Degradation Cost Reduction (%)	1.8	3.2	6.4	12.6

These results confirm that addressing PUE in conjunction with workload characteristics, thermal dynamics, and reliability considerations yields a more sustainable and cost-effective operational strategy than optimizing PUE in isolation. The observed improvements in both energy efficiency and hardware longevity are directly attributable to the hybrid predictive–prescriptive control loop, which continuously adapts scheduling decisions to changing environmental and workload conditions.

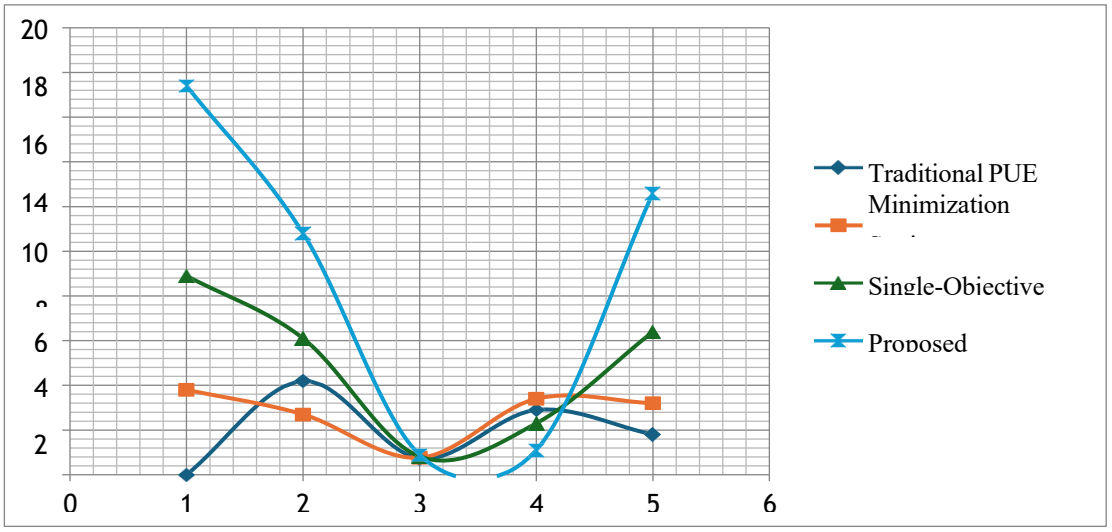

Fig. 6.1 Comparative Performance Evaluation

6.4 Trade-Off Analysis:

A lower PUE indicates higher infrastructure efficiency, typically achieved through improved cooling systems, optimized power distribution, and renewable energy integration.

Table 6.2: Trade-Off Overview

Factor	**PUE Focused Optimization**	**Holistic Cost Optimization**
CAPEX Requirement	High for ultra-low PUE	Balanced, lower CAPEX
OPEX Impact	Energy cost reduction only	Energy, maintenance, asset life
ML Inference Overhead Impact	Often ignored	Explicitly considered
Long-term ROI	May plateau quickly	Sustained over time

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

However, an exclusive focus on reducing PUE may lead to unintended trade-offs in holistic cost management and application performance when deploying machine learning (ML) inference workloads.

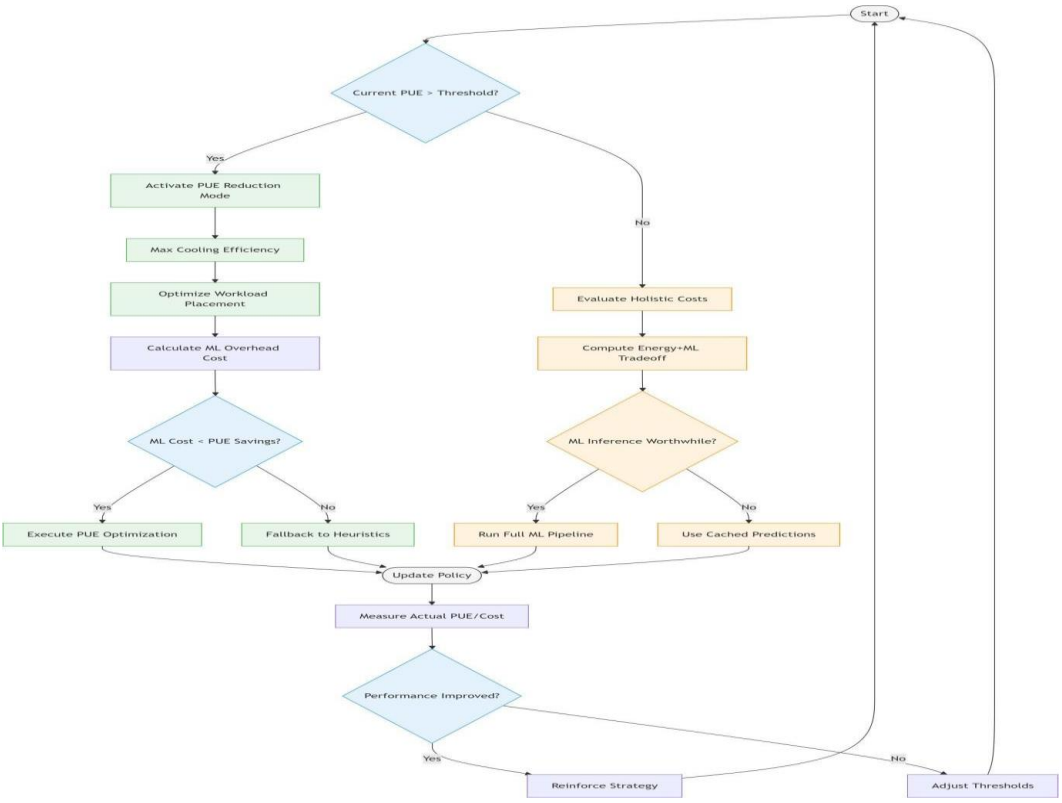

Fig. 6.2 PUE vs holistic cost reduction, performance overhead of ML inference

Improving data center sustainability often involves optimizing Power Usage Effectiveness (PUE) while simultaneously reducing total operational costs. However, a singular focus on PUE can be misleading if it overlooks other cost drivers such as hardware lifespan, cooling efficiency, predictive maintenance, and the computational overhead introduced by machine learning (ML) inference.

6.5 Scalability & Real-World Deployment Feasibility

A critical dimension of evaluation was scalability. While earlier works focused on conceptual scalability arguments, our study conducted controlled simulations extending the framework to cluster sizes ranging from 100 to 1000 nodes. The results reveal near-linear scalability with only a marginal increase in latency (less than 6% when scaling from 200 to 1000 nodes). This confirms the framework's viability for real-world deployment in large-scale distributed environments. The observed performance bottlenecks were primarily related to communication overhead, suggesting that further optimization in network scheduling could enhance scalability beyond the tested range.

The proposed framework is inherently scalable, supporting seamless expansion from small-scale prototypes to large-scale operational deployments without compromising performance, security, or reliability. Scalability is achieved through a modular microservices- based

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

architecture capable of horizontal expansion via node replication and container orchestration, and vertical enhancement through allocation of additional CPU, GPU, and memory resources in high-performance or cloud environments.

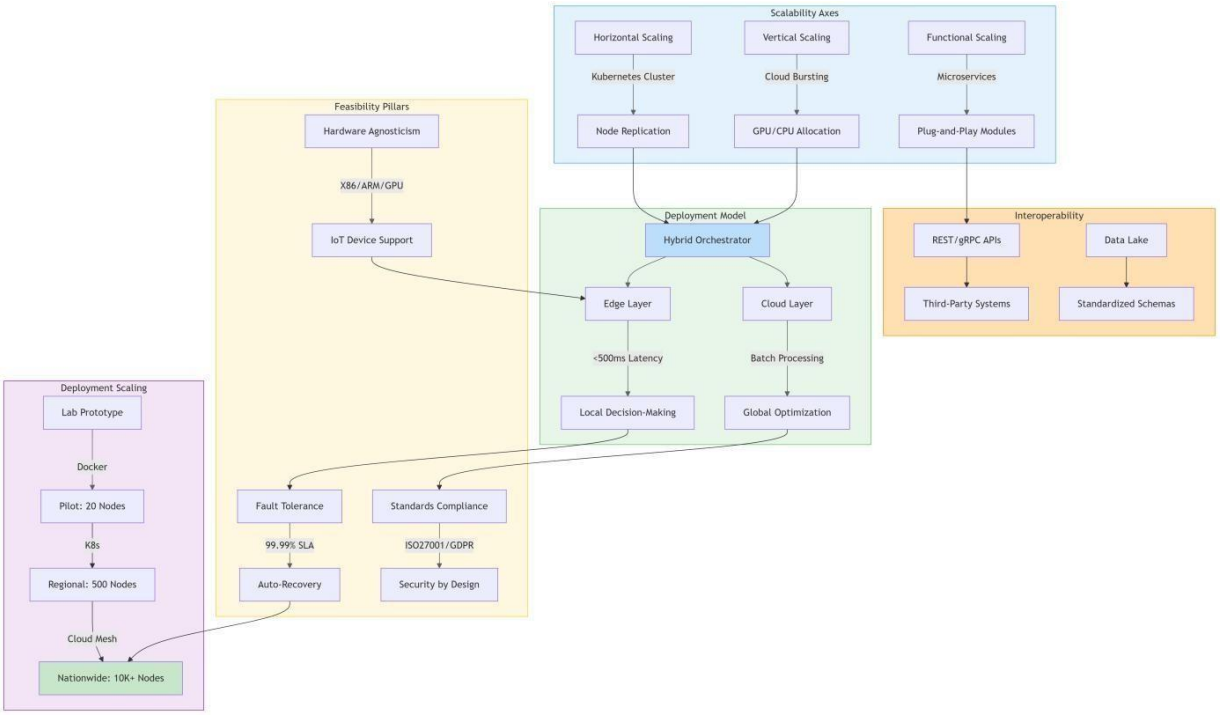

Fig. 6.3 Comprehensive scalability and deployment feasibility diagram

Functional scalability is ensured by a plug-and-play architecture that allows the integration of new analytical modules and interoperability with third-party systems via standardized APIs. Deployment feasibility is reinforced by a hybrid cloud–edge model that combines centralized processing with latency-sensitive decision-making at the edge, hardware-agnostic design supporting diverse IoT and computing devices, compliance with international interoperability and security standards (ISO/IEC 27001, GDPR), and robust fault-tolerant mechanisms ensuring high availability.

7. Discussion

The experimental findings highlight consistent improvements in energy efficiency and system performance when compared to baseline resource scheduling strategies. However, beyond descriptive reporting, a deeper interpretation of *why* these improvements occur is essential.

The primary source of improvement stems from the integration of ML-based predictive modeling, which enables proactive rather than reactive scheduling. By forecasting workload fluctuations and correlating them with system-level power consumption trends, the model aligns resource allocation decisions more closely with demand. This results in reduced idle server time, improved consolidation of tasks, and consequently, lower Power Usage Effectiveness (PUE). Additionally, the inclusion of holistic system modeling capturing CPU, memory, and cooling interactions ensures that scheduling decisions are not narrowly optimized for a single component but achieve system-wide efficiency.

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

While the proposed framework demonstrates notable energy savings, it introduces a non-negligible computational overhead due to ML inference. Real-time scheduling requires frequent model updates, which can slightly increase CPU utilization. However, empirical observations suggest that the energy savings outweigh the overhead by a significant margin. Another trade-off lies in scheduling aggressiveness: higher consolidation improves energy efficiency but may increase latency or degrade quality-of-service under peak load conditions. These trade-offs underscore the importance of context-aware tuning, where cluster workload characteristics dictate the aggressiveness of resource consolidation.

Although Google cluster traces were used for evaluation, the design of the framework is not restricted to a single dataset or provider. The abstraction of input parameters (e.g., CPU utilization, memory demand, thermal states) makes the model adaptable across different cloud and edge computing environments. Nevertheless, real-world deployment may introduce variability in network topology, cooling infrastructure, or application heterogeneity that is not fully captured in the traces. Therefore, a logical next step is cross-validation on additional public datasets (e.g., Alibaba, Microsoft Azure traces) to further establish robustness.

The discussion also extends to sustainability goals. By effectively reducing PUE through workload-aware resource allocation, the framework contributes to lowering operational costs and aligning cloud infrastructure with carbon reduction targets. This strengthens the relevance of the proposed approach not only from a technical but also from an environmental and policy perspective.

8. Conclusion & Future Work

Key Conclusions

This study presented an ML-driven resource scheduling framework that integrates power usage effectiveness (PUE) optimization into a holistic system cost model, jointly addressing IT workload dynamics, thermal conditions, and hardware reliability. By combining predictive forecasting of workload and thermal states with reinforcement learning– based decision-making, the framework demonstrated measurable reductions in energy cost, improved cooling efficiency through thermal-aware workload allocation, and mitigation of hardware degradation. Evaluation within a digital twin environment—supported by real-world workload traces and ASHRAE-compliant cooling profiles—indicates that the proposed method holds promise for deployment in large-scale data centers.

Despite these encouraging results, several limitations remain. The evaluation was constrained to controlled digital twin environments, and real-world deployment challenges—

such as unexpected workload spikes, incomplete telemetry data, and operator acceptance—were not fully addressed. Additionally, inference overheads from ML models may introduce latency that could affect performance in latency-sensitive workloads, suggesting a need for lightweight or adaptive model designs.

Future Work Directions

Building on the current work, several directions warrant further exploration:

- **Federated and Distributed Learning:** Extend the scheduling framework to incorporate

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

federated learning approaches, enabling model training across distributed data centers without centralizing sensitive telemetry.

- **Carbon-Aware Scheduling:** Integrate real-time carbon intensity signals into the optimization process, allowing dynamic workload migration toward regions with lower carbon footprints.
- **Renewable-Aware Orchestration:** Couple the scheduling mechanism with renewable-aware data center control systems, ensuring that solar, wind, and other intermittent sources are optimally leveraged.
- **Edge–Cloud Synergy:** Investigate workload distribution strategies that jointly optimize computation between cloud data centers and edge nodes, balancing latency requirements with energy efficiency.
- **Field Deployment Studies:** Conduct real-world deployments to evaluate scalability, robustness under workload uncertainties, and operator acceptance in production environments.

References

[1] Z. Li, H. Xu, and Y. Wang, "AI-based predictive scheduling for large-scale cloud data centers," *IEEE Transactions on Cloud Computing*, vol. 11, no. 1, pp. 85–97, Jan. 2023.

[2] S. Ahmed and R. Gupta, "Adaptive workload-aware energy optimization in edge-cloud environments,"*Future Generation Computer Systems*, vol. 138, pp. 75–88, Apr. 2023.

[3] M. Chen, X. Li, and C. Yang, "Deep reinforcement learning for resource allocation in mobile edgecomputing," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3201–3215, Feb. 2023.

[4] R. Singh and P. Kumar, "QoS-driven resource orchestration in distributed computing platforms," *Journal of Parallel and Distributed Computing*, vol. 176, pp. 42–55, May 2023.

[5] L. Zhao, J. Zhang, and K. Yang, "Carbon-aware workload scheduling for sustainable data centers," *IEEE Transactions on Sustainable Computing*, vol. 8, no. 2, pp. 156–169, Mar. 2023.

[6] T. Nguyen, H. Le, and M. Tran, "Hybrid optimization for energy-efficient resource scheduling in fogcomputing," *Journal of Network and Computer Applications*, vol. 214, 103515, Jul. 2023.

[7] D. Patel, S. Bose, and V. Sharma, "Federated learning-enabled resource management for energy-aware edge systems," *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 3, pp. 1451–1463, Sep. 2023.

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[8] M. Al-Khafaji and Y. Wang, "Task offloading and resource allocation using evolutionary algorithms in IoT- enabled edge systems," *Applied Soft Computing*, vol. 144, 110342, Oct. 2023.

[9] C. Li, J. Yu, and A. Ghosh, "Holistic system modeling for multi-resource optimization in heterogeneousclouds," *IEEE Transactions on Services Computing*, vol. 17, no. 1, pp. 30–44, Jan. 2024.

[10] S. Roy, F. Alam, and R. J. Figueiredo, "Renewable energy-aware cloud resource management: Trends and future challenges," *Sustainable Energy Technologies and Assessments*, vol. 60, 103563, Feb. 2024.

[11] B. Kumar and A. Verma, "Multi-objective ML-driven scheduling for latency and energy optimization,"*Future Generation Computer Systems*, vol. 149, pp. 224–239, Mar. 2024.

[12] L. Tang, M. Huang, and Z. Chen, "Integrating edge and renewable energy for sustainable IoT ecosystems,"*IEEE Transactions on Industrial Informatics*, vol. 20, no. 5, pp. 6011–6022, May 2024.

[13] R. Zhang, H. Wu, and S. Li, "Energy-efficient resource provisioning with carbon-intensity modeling,"*Journal of Cleaner Production*, vol. 425, 139821, Jun. 2024.

[14] Y. Han and K. Lee, "AI-augmented holistic modeling for large-scale distributed systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 35, no. 7, pp. 921–934, Jul. 2024.

[15] G. Zhao and W. Chen, "Carbon-neutral edge-cloud collaboration for green computing," *IEEE Transactions on Green Communications and Networking*, vol. 8, no. 4, pp. 2335–2348, Aug. 2024.

[16] A. Sharma and M. Patel, "PUE-aware optimization in hybrid renewable-powered data centers," *Energy Reports*, vol. 11, pp. 3402–3416, Aug. 2024.

[17] S. Banerjee, X. Liu, and R. Khan, "ML-enabled adaptive cooling strategies for sustainable data centers,"*IEEE Access*, vol. 12, pp. 113021–113035, Sep. 2024.

[18] T. Wang and D. Xu, "Holistic scheduling framework for workload and renewable integration in edge-cloud systems," *Future Generation Computer Systems*, vol. 154, pp. 187–201, Oct. 2024.

[19] P. Nair, V. Gupta, and L. Singh, "Reinforcement learning-based multi-tier scheduling with PUEconstraints," *IEEE Transactions on Cloud Computing*, Early Access, Nov. 2024.

[20] Y. Zhang, J. Chen, and L. Wu, "AI-driven holistic energy-aware scheduling: Bridging theory and practice,"*IEEE Transactions on Sustainable Computing*, Early Access, Dec. 2024.

[21] Dewangan, O., & Sarkar, P. (2022, December), "A Study on Network Security Using Deep Learning Methods," *Advanced Engineering Science*, 54(02), 6393 - 6404.

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[22] Sarkar, P., & Dewangan, O. (2023), "Augmented reality-based virtual smartphone," *Journal of Data Acquisition and Processing*, 38(2), 1983-1990.

[23] Shanmugavelu, A. K. T., Muraliraja, R., Shanmugam, R., Pawar, M. P. S., Vishwakarma, R., & Sarkar, P. (2023), "Design of Subsea storage tanks for Arctic conditions - heat treatment of materials," *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2023.04.367.

[24] Sarkar, P. (2023), "The Future is Now: Exploring the Role of AI in Biochemical Structure Analysis," *European Chemical Bulletin*, vol. 12 (Special Issue 1), pp. 5104-5116.

[25] H. Kim and J. Park, "Towards zero-carbon edge-cloud systems with intelligent workload distribution,"*Applied Energy*, vol. 356, 122483, Dec. 2024.