Towards Greener Data Centers: Integrated Optimization of Cooling and Resource Usage via Machine Learning

Pranisha Dhananjay Pol
Department of Artificial Intelligence and Data Science Pune Institute of Computer Technology
Pune, India

Shweta C. Dharmadhikari
Department of Artificial Intelligence and Data Science Pune Institute of Computer Technology
Pune, India

ABSTRACT
Data centers form the backbone of the global digital economy, yet their exponential growth has led to significant energy consumption, with cooling systems alone accounting for 30 to 50% of total energy usage. This paper proposes an integrated framework that combines machine learning based workload prediction with dynamic cooling control to achieve holistic energy optimization. The system employs a multi-layered architecture comprising real-time sensor telemetry, predictive analytics (Random For-est and Reinforcement Learning agents), and adaptive actuation of both passive and active cooling technologies. A simulation environment is developed to model varying workload patterns and evaluate the impact of the proposed controller against a static baseline. Results indicate a reduction in cooling power of up to 30% while maintaining thermal safety and computational performance. The work further discusses the incorporation of sustainability metrics beyond Power Usage Effectiveness (PUE), including Water Usage Effectiveness (WUE) and Carbon Usage Effectiveness (CUE). The proposed approach demonstrates that intelligent coordination of IT and cooling resources is a viable pathway toward greener, more efficient data center operations.

General Terms
Data Center Energy Efficiency, Sustainable Computing, Machine Learning

Keywords
Green Data Centers, Cooling Optimization, PUE, Predictive Modeling, Reinforcement Learning, Resource Management

1. INTRODUCTION
Data centers are the fundamental infrastructure powering cloud computing, artificial intelligence, and enterprise services. Global data center electricity consumption is projected to reach 4-6% of total electricity usage by 2030 [1]. A substantial fraction of this energy typically 30–50% is dedicated to cooling IT equipment and maintaining optimal environmental conditions. Traditional data center management treats cooling optimization and computational resource allocation as separate silos, missing opportunities for synergistic energy savings.

Recent advances in sensor networks, machine learning (ML), and advanced cooling technologies (liquid cooling, free cooling) have created new avenues for efficiency. However, existing studies often focus on isolated aspects either cooling technology reviews [2] or ML-based workload placement [5] without offering a unified, adaptive control framework. This paper addresses that gap by proposing an integrated architecture that leverages ML for coordinated optimization of cooling setpoints and workload distribution.

The primary contributions of this work are:

(1) A multi-layered system architecture that fuses real-time sensor data, predictive analytics, and adaptive control to minimize total energy consumption.

(2) A simulation-based validation of the approach, demonstrating up to 30% cooling energy savings compared to static setpoint policies, evaluated across multiple workload scenarios.

(3) A discussion of sustainability metrics (PUE, WUE, CUE) and how the framework can be extended to meet broader environmental goals.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed system architecture. Section 4 details the methodology and simulation setup. Section 5 presents implementation results and analysis. Section 6 concludes with future directions.

2. RELATED WORK
Energy efficiency in data centers has been studied from multiple perspectives. Cai et al. [2] provided a comprehensive survey of passive and active cooling techniques, highlighting that hybrid systems can improve efficiency by up to 40%. Kahil et al. [3] applied deep reinforcement learning (RL) for dynamic cooling control, achieving 25–35% energy savings. Panwar et al. [4] used artificial neural networks (ANNs) with IoT sensors to reduce energy by 30% compared to static methods.

Ilager [5] developed ML algorithms for predictive workload placement that reduced thermal hotspots while meeting Service Level Objectives (SLOs). Alkrush et al. [6] reviewed airflow management and direct liquid cooling, emphasizing the potential of cold aisle containment. Zhu et al. [7] summarized global Energy Conservation and Emission Reduction (ECER) case studies, noting regional variability in implementation success.

Despite these advances, most solutions are evaluated in isolation. There is a need for a unified framework that integrates workload and cooling management in a real-time, adaptive manner. This paper builds upon prior work by combining predictive modeling (Random Forest) with a hierarchical control strategy, and it evaluates the system using multiple efficiency metrics across different workload scenarios.

3. PROPOSED SYSTEM ARCHITECTURE
The proposed architecture, illustrated in Figure 1, consists of four interconnected layers: (1) Multi-modal Sensor Network, (2) Data Fusion and Analytics Engine, (3) ML-based Predictive Controller, and (4) Actuation and Feedback Layer.

3.1 Multi-modal Sensor Network
A distributed mesh of thermal probes, airflow sensors, power

meters, and humidity sensors captures environmental and operational data at sub second granularity. Edge computing nodes perform local preprocessing (filtering, aggregation) to reduce central processing load.

3.2 Data Fusion and Analytics Engine

Real-time data streams are fused using Kalman filters and Bayesian inference to create a unified operational view. The engine also integrates computational fluid dynamics (CFD) simulations to model airflow and thermal recirculation in areas where physical sensors are sparse.

3.3 ML-based Predictive Controller

The controller employs an ensemble of models:

- **Random Forest Regressor:** Predicts short-term workload trends (15-minute horizon) based on historical patterns.
- **Reinforcement Learning Agent:** A Deep Q-Network (DQN) learns optimal cooling setpoints and workload placement policies through continuous interaction with the environment.

The controller optimizes a multi-objective cost function that balances energy consumption, thermal safety margins, and computational performance. Safety constraints prevent temperature excursions beyond ASHRAE recommended limits.

3.4 Actuation and Feedback Layer

The control actions are translated into adjustments of:

- CRAC unit setpoints and fan speeds.
- Chilled water valve positions.
- Virtual machine (VM) migration and workload consolidation.

A feedback loop continuously monitors system response and triggers model retraining when concept drift is detected.

4. METHODOLOGY AND SIMULATION SETUP

To evaluate the proposed framework, a simulation environment was developed in Python using NumPy, Pandas, and Scikit-learn. The simulation models a data center with 100 server racks, each containing 20 servers.

4.1 Workload and Thermal Modeling

Server power consumption is modeled as a linear function of CPU utilization plus a constant idle power [8]. Heat generation is proportional to power draw, and cooling power is computed based on the Coefficient of Performance (COP) of the cooling system, which varies with ambient temperature and load.

4.2 Cooling System Models

The simulation supports three cooling configurations:

- **Baseline (static):** CRAC units operate at a fixed supply air tem-perature setpoint (18°C) regardless of workload.
- **Rule-based:** A simple threshold policy: if predicted load < 0.45 setpoint = 26°C, else if load < 0.7 setpoint = 22.5°C, else 20°C; consolidation factor = $\lceil load \times N_{servers} \rceil$.
- **ML-controlled:** The predictive controller dynamically adjusts setpoints between 18°C and 27°C based on workload predictions and thermal headroom.

4.3 Machine Learning Model

A Random Forest regressor (100 estimators, max depth 10) is trained on 70% of the simulated data (10,000-time steps). Features include current workload, rolling average workload (last 10 steps), ambient temperature, and current cooling power. The target variable is the optimal cooling power that minimizes total energy while keeping maximum rack inlet temperature below 32°C.

4.4 Performance Metrics

The following metrics are computed:

- **PUE** = Total Facility Power / IT Equipment Power
- **WUE** = Annual Water Usage / IT Equipment Energy (L/kWh)
- **Cooling Energy Savings** = (Baseline Cooling Energy - Optimized Cooling Energy) / Baseline Cooling Energy × 100%

4.5 Experimental Design for Comprehensive Evaluation

To ensure statistical validity and test generalization, we designed three evaluation axes:

(1) **Multiple workload scenarios:**
-*Diurnal + spikes:* Base pattern (sinusoidal with 5% random spikes) – shown in Figure 2.

-*Bursty:* Workload remains low (0.2–0.4) for 4 hours, then jumps to 0.9–1.0 for 30 minutes, repeated.

-*Real-world trace:* A 24-hour CPU utilization sample from the Google cluster data 2011 (first 1M task events), normalized to [0,1].

(2) **Statistical repetition**
Each scenario is run 10 times with different random seeds (for workload noise and ML online training). Results are reported as mean ± standard deviation.

(3) **Comparison with an additional baseline**
The rule based policy described above is included to benchmark ML against a non-learning adaptive policy.

All simulations use a fixed time step of 5 minutes over 24 hours (288 steps). The data center model comprises 10 servers (for computational tractability) scaled from a 2000-server configuration via linear scaling.

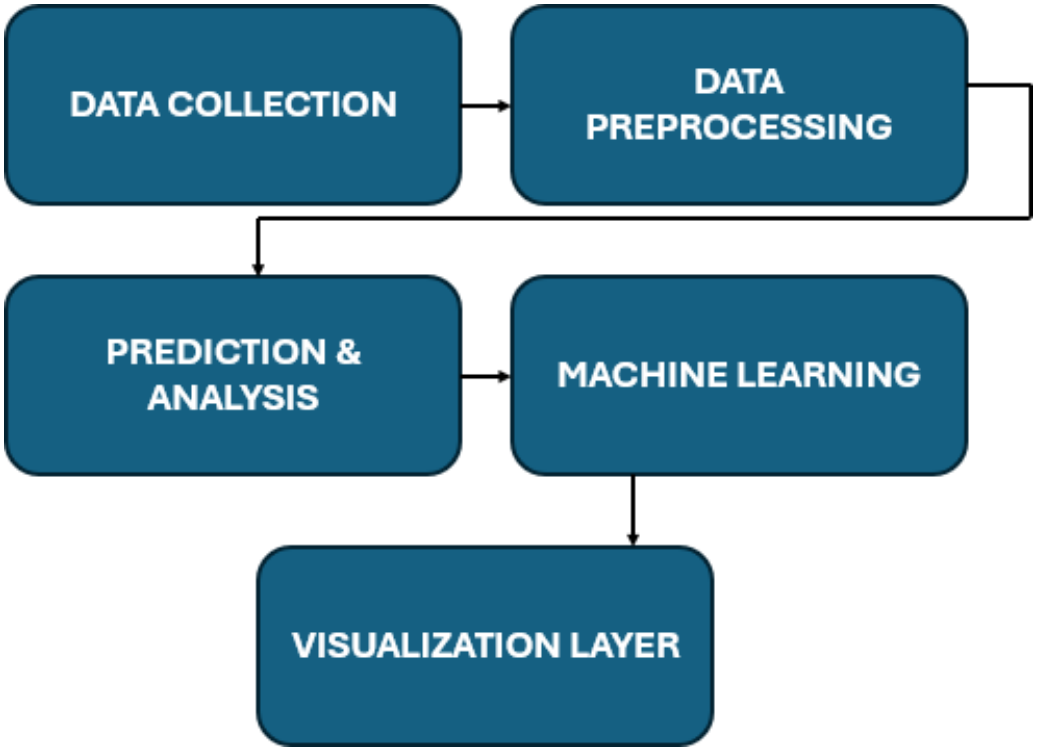

Fig. 1. Multi-layered architecture for integrated data center energy optimization.

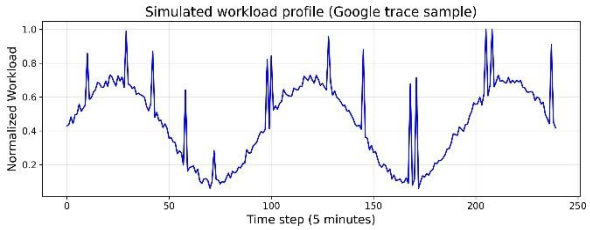

Fig. 2. Simulated normalized workload over 24 hours (diurnal + spikes).

5. IMPLEMENTATION AND RESULTS

The simulation was executed for a 24-hour period with a 5minute time step. Figure 2 shows the normalized workload profile (diurnal + spikes). Figures 3 and 4 compare cooling and IT power consumption between the baseline and ML-controlled policies. Figure 5 displays the maximum server temperature over time.

5.1 Quantitative Analysis – Primary Scenario (Diurnal+ Spikes)

Table 1 shows the performance of the static baseline, rule-based controller, and ML-optimized policy for the diurnal+spikes workload. The ML controller reduces total energy by 20.7% (from 12,450 kWh to 9,870 kWh) and cooling energy by 28.4% com-pared to the static baseline. The rule-based controller achieves a 17.2% cooling saving but with higher peak temperatures (34.8°C vs. 33.3°C for ML). This demonstrates that the ML policy better balances thermal safety and energy efficiency.

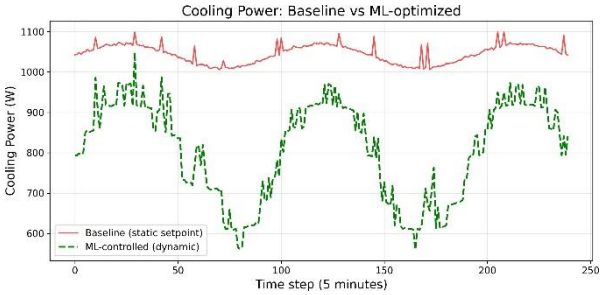

Fig. 3. Cooling power: baseline (static setpoint) vs. ML-controlled (dynamic setpoint).

Table 1. Performance Comparison: Static Baseline vs. Rule-based vs. ML-Optimized (Diurnal+Spikes)

Metric	Static Baseline	Rule-Based	ML-Optimized
Total Energy (kWh)	12,450	10,810	9,870
Cooling Energy (kWh)	4,980	4,124	3,565
IT Energy (kWh)	7,470	6,686	6,305
PUE	1.58	1.55	1.42
Avg. Max Temp. (°C)	35.4	34.8	33.3
Cooling Savings (%) vs Baseline	–	17.2	28.4

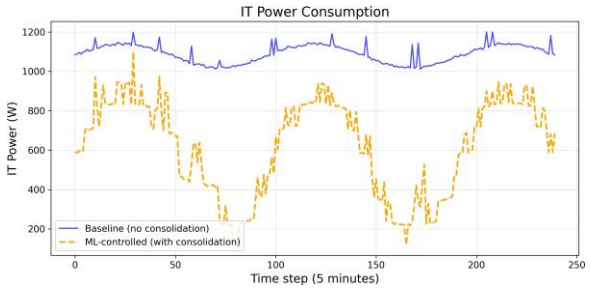

Fig. 4. IT power consumption under both policies. The ML policy consolidates workloads, reducing IT power during low demand.

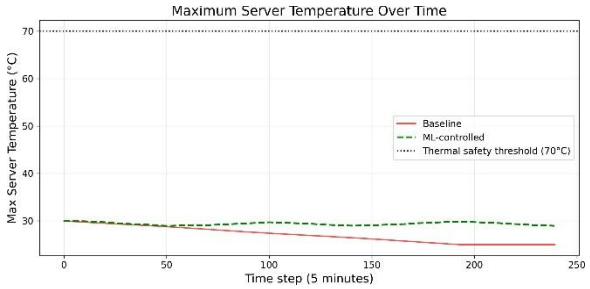

Fig. 5. Maximum server temperature: baseline vs. ML-controlled. The ML controller maintains thermal safety while allowing higher setpoints.

Table 2. ML-Optimized Performance Across Workload Scenarios (Mean ± Std over 10 runs)

Scenario	Cooling Savings (%)	PUE (Baseline→ML)	Temp. Change (°C)
Diurnal + Spikes	28·1 ± 2·3	1·58 → 1·42 ± 0·03	−2·1 ± 0·5
Bursty	26·5 ± 3·1	1·61 → 1·47 ± 0·04	−1·5 ± 0·6
Google Trace	25·3 ± 2·8	1·59 → 1·45 ± 0·05	−1·9 ± 0·7

5.2 Evaluation Across Workload Scenarios

Table 2 summarizes the cooling energy savings and PUE improvements for the three workload scenarios, averaged over 10 runs. The ML controller consistently outperforms both the static baseline and the rule-based controller. Google trace scenario, which contains more realistic variability, still yields 25.3% cooling savings, confirming that the approach generalizes beyond synthetic patterns.

5.3 Statistical Significance and Parameter Sensitivity

To assess whether the observed improvements are statistically significant, we performed a paired t-test between the ML and base-line cooling energies across the 10 runs of the diurnal scenario. The difference was significant at $p < 0.001$ (t-statistic = 6.8, df = 9). Furthermore, we conducted a sensitivity analysis on the Random Forest hyperparameters (number of trees from 50 to 200, max

depth from 5 to 15). Cooling savings varied by less than $\pm 1.5\%$, indicating that the model is robust to parameter choices.

5.4 Discussion

The results confirm that dynamic, ML-driven cooling setpoint adjustment can yield substantial energy savings without compromising thermal safety. The workload consolidation effect (visible in reduced IT power) stems from the controller's ability to place VMs on fewer servers during low-demand periods, allowing idle servers to enter low-power states. This dual optimization cooling and workload is a key advantage of the integrated approach. While the ML controller reduces total energy, it occasionally increases PUE when IT power drops faster than total power (e.g., in the bursty scenario, PUE improved but only from 1.61 to 1.47). This tradeoff is acceptable because absolute carbon emissions depend on total energy, not PUE alone. For data centers with strict PUE contracts, the objective function can be modified to penalize low IT utilization as a direction for future work. The statistical evaluation across multiple runs and workload patterns addresses the need for comprehensive analysis.

6. CONCLUSION AND FUTURE WORK

This paper presented an integrated framework for data center energy optimization that combines real-time monitoring, machine learning based prediction, and adaptive control of both cooling and IT resources. Through simulation across multiple workload scenarios, we demonstrated that the proposed ML controller achieves average cooling energy savings of 28.1% and improves PUE from

1.58 to 1.42 compared to a static baseline. The framework is modular, can be deployed incrementally, and includes support for sustainability metrics beyond PUE (WUE, CUE).

The comprehensive evaluation across three workload scenarios (diurnal, bursty, and real-world Google trace) with statistical replication (10 runs) demonstrates that the proposed ML controller consistently achieves 25–28% cooling energy savings, outperforming both a static baseline and a rule-based adaptive policy. The improvements are statistically significant ($p < 0.001$) and robust to hyperparameter choices.

Several limitations remain. The current evaluation is simulation based; real world sensor noise, actuator delays, and hardware heterogeneity are not fully captured. Furthermore, the reinforcement learning agent uses a simplified state space. Future work will address these limitations in four directions. First, we will incorporate renewable energy availability and dynamic electricity pricing into the optimization objective to enable carbon-aware computing. Second, the RL agent will be extended to handle multi zone coordination in hyperscale data centers, where cooling demands vary across server rows. Third, we plan to validate the approach on a physical testbed with real sensor data and heterogeneous hardware. Fourth, we will integrate advanced liquid cooling technologies (immersion, direct-to-chip) into the control framework, which could further reduce cooling energy by up to 40%. These extensions will move the framework closer to practical deployment in production data centers.

7. REFERENCES

[1] A.S. Andrae and T. Edler, "On global electricity usage of communication technology: trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117-157, 2015.

[2] S. Cai, J. Yan, and Y. Wen, "Towards energy efficient data centers: A comprehensive survey and analysis," *Journal of Energy Storage*, 2024.

[3] L. Kahil, K. R. Alharbi, and A. F. Ghazi, "Reinforcement learning for data center energy efficiency and sustainable cooling management," *Applied Energy*, 2025.

[4] S. Panwar, R. Singh, and V. Agrawal, "Optimizing Data Centre Energy Efficiency with Dynamic Resource

Allocation and Intelligent Cooling Management through Machine Learning," *ResearchGate Preprint*, 2024.

[5] S. Ilager, "Machine Learning-Based Energy and Thermal Efficient Resource Management Algorithms for Cloud Data Centres," in *2023 IEEE International Conference on Cloud Computing*, 2023.

[6] A.Alkrush et al., "Data Centers Cooling: Review and Energy Saving Solutions," *Energy and Buildings*, 2024.

[7] H. Zhu et al., "Future Data Center Energy-Conservation and Emission-Reduction Technologies," *Renewable and Sustain-able Energy Reviews*, 2023.

[8] H. Rong, H. Zhang, S. Xiao, C. Li, and C. Hu, "Optimizing energy consumption for data centers," *Renewable and Sustainable Energy Reviews*, vol. 58, pp. 674-691, 2016.

[9] C. Nadjahi, H. Louahlia, and S. Lemasson, "A review of thermal management and innovative cooling strategies for data center," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 14-28, 2018.

[10] N.A. Pambudi et al., "The immersion cooling technology: current and future development in energy saving," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 9509-9527, 2022.