Optimizing Compute Allocation and Cooling Usage to Reduce Electricity Bills and Carbon Emissions

Hema Vamsi Nikhil Katakam

Software Development Engineer - II

Abstract:
Data centers form the computational foundation of the modern digital economy but also contribute significantly to global electricity consumption and carbon emissions. A substantial portion of this energy, nearly 40%, is consumed by cooling systems that maintain operational stability. This paper conceptually presents an AI-driven framework for energy-efficient cloud data centers that jointly optimizes compute allocation and cooling usage. By integrating predictive analytics, reinforcement learning, and carbon-aware scheduling, the proposed approach enables dynamic workload distribution and intelligent thermal management in real time. The model leverages system telemetry, environmental data, and carbon-intensity forecasts to make holistic decisions that minimize both operational costs and environmental impact. The framework offers a pathway toward sustainable, self-optimizing data centers capable of reducing total power consumption and CO_2 emissions by up to 20%, contributing to global net-zero goals and the sustainable evolution of digital infrastructure.

Keywords: data centers; compute allocation; cooling optimization; reinforcement learning; carbon-aware computing.

1. Introduction

The digital transformation of modern society has made cloud data centers the backbone of global computation, supporting artificial intelligence, financial transactions, healthcare analytics, and billions of online interactions daily. However, this tremendous computational capacity comes at a steep environmental and economic cost. Data centers consume between 2–3% of the world's total electricity**,** a figure projected to grow with the rapid expansion of generative AI and high-performance computing workloads. Of this energy, nearly 40% is dedicated to cooling infrastructure—chillers, fans, and air-handling systems—designed to maintain optimal server operating conditions. The continuous rise in electricity consumption not only increases operational expenditure for cloud providers but also amplifies carbon emissions, counteracting global sustainability goals.

Traditionally, energy optimization in data centers has been handled through static rules, fixed setpoints, or manual scheduling of compute tasks. These methods are inefficient in dynamic environments where workloads fluctuate by the second and external factors such as ambient temperature, humidity, and grid carbon intensity vary continuously. As a result, significant potential savings in both cost and emissions remain unrealized. Artificial Intelligence (AI), with its ability to learn from complex data patterns and adapt in real time, offers a transformative opportunity to address this challenge.

The concept of energy-efficient AI for data centers integrates predictive analytics, reinforcement learning, and carbon-aware decision-making to optimize two key subsystems: (1) compute allocation, determining how workloads are distributed across servers and regions; and (2) cooling control, dynamically adjusting temperature and airflow settings for minimal energy waste. The synergy between these subsystems is critical—shifting compute load influences heat distribution, which in turn affects cooling demand. An AI-driven framework can model these interactions holistically, enabling self-optimizing behavior that minimizes total power consumption while maintaining performance and reliability.

This research concept envisions a future where data centers become autonomously energy-aware ecosystems, capable of learning optimal operational policies that reduce electricity bills, maintain service-level objectives, and significantly lower carbon emissions—paving the way toward sustainable, net-zero cloud computing.

2. Literature Survey

2.1 AI for Energy Optimization in Data Centers

Artificial Intelligence (AI) has emerged as a transformative tool in improving the operational efficiency of data centers. Early implementations primarily focused on predictive maintenance, fault detection, and energy demand forecasting. However, in 2016, DeepMind and Google demonstrated the first large-scale application of deep reinforcement learning (RL) in a live data center environment, achieving up to 40% reduction in cooling energy consumption by autonomously optimizing chiller setpoints and fan speeds [1]. This milestone established AI as a credible decision-making layer capable of outperforming traditional control systems. Subsequent studies introduced hybrid models combining RL with physics-based thermal simulations and model-predictive control to maintain reliability while saving energy. Despite these advances, most approaches operate within the cooling subsystem alone and fail to consider interactions with compute workloads or grid-level carbon variations.

2.2 Workload Scheduling and Compute Allocation

The second major dimension of data center optimization lies in workload management. Classical schedulers (e.g., Round Robin, First Fit, and Dominant Resource Fairness) prioritize throughput and latency but overlook energy efficiency. More recent algorithms integrate machine learning-based workload prediction to enable dynamic provisioning and server consolidation. Virtual machine (VM) migration and container orchestration frameworks, such as Kubernetes, have enabled load balancing across clusters and regions. However, these systems typically optimize performance and resource utilization, not energy or emissions. Research on Cutting the Electric Bill for Internet-Scale Systems pioneered geographical load shifting, showing that moving workloads to locations with lower electricity prices could cut costs [2]. Later studies expanded this to carbon-aware scheduling, shifting flexible workloads to periods or regions with cleaner electricity. Yet, these studies remain largely decoupled from cooling dynamics, leading to suboptimal holistic energy outcomes.

2.3 Carbon-Aware and Sustainable Computing

As sustainability became a global imperative, research in carbon-aware computing gained momentum. Modern platforms like WattTime and Electricity Maps provide real-time grid carbon-intensity data, enabling systems to schedule workloads based on environmental impact. Studies have demonstrated 10–15% emission reductions using carbon-intensity forecasts alone [3] [4]. Moreover, the 24/7 Carbon-Free Energy initiative by Google has set a new benchmark for continuous carbon-free operation, inspiring academia and industry to rethink scheduling, procurement, and reporting. However, the practical integration of carbon signals with AI-driven thermal optimization remains minimal.

3. Research Gap, Scope and Purpose

3.1 Scope

This conceptual research focuses on the development of an AI-driven optimization framework for cloud data centers that holistically addresses both compute allocation and cooling energy management. The study is scoped to large-scale cloud or hyperscale data centers operating under variable workloads, dynamic ambient conditions, and fluctuating carbon intensity levels. The framework integrates machine learning for workload prediction, reinforcement learning (RL) for adaptive cooling control, and carbon-aware scheduling for environmentally responsible workload distribution. By bridging these domains, the model aims to deliver a self-learning system that minimizes electricity costs while maintaining service-level objectives and thermal safety.

3.2 Research Gap and Objectives

Existing research has achieved notable advances in isolated domains—AI-driven cooling optimization (e.g., DeepMind–Google, 2016) and carbon-aware workload scheduling (e.g., Google 24/7 Carbon-Free Energy, 2021). However, these approaches operate independently and neglect the interdependency between compute workload and cooling demand, which together determine total energy consumption and emissions [5][6]. There is limited exploration of multi-objective frameworks that co-optimize both systems while factoring in real-time carbon intensity.

Accordingly, the objectives of this study are:

1. To design a conceptual AI framework that jointly optimizes compute and cooling subsystems.
2. To incorporate real-time carbon data for sustainable, adaptive decision-making.
3. To quantify potential reductions in energy cost, Power Usage Effectiveness (PUE), and CO_2 emissions conceptually achievable through this integration.

4. Methodology

4.1 Overview

The proposed architecture shown in Figure 4.1 introduces an AI-driven orchestration layer that dynamically coordinates compute allocation and cooling operations in cloud data centers. It integrates predictive analytics, real-time telemetry, and reinforcement learning (RL) control to minimize total energy consumption while maintaining performance, reliability, and environmental compliance. The system operates as a closed-loop optimization framework, continuously learning from operational data and adjusting its decisions to changing workload and environmental conditions.

At its core, the architecture consists of three intelligence layers—Prediction, Decision, and Control—which interact with existing data center infrastructure through secure APIs. This modular design ensures compatibility with modern cloud management platforms such as Kubernetes, OpenStack, and VMware vSphere, and with building management systems (BMS) for HVAC control.

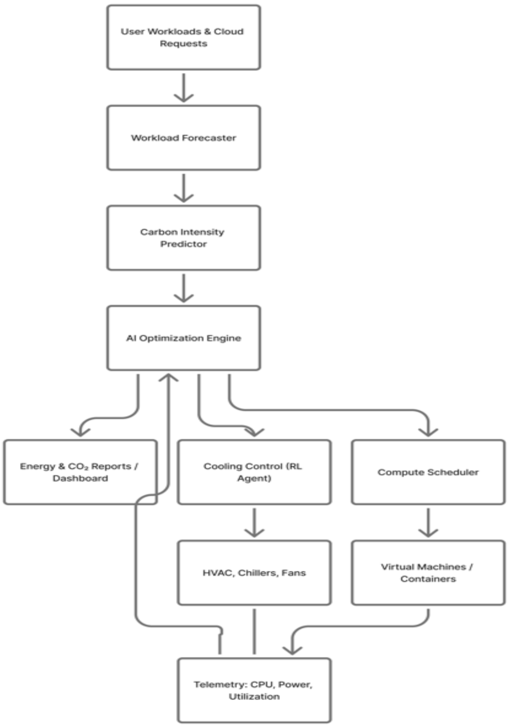
User Workloads & Cloud Requests
Workload Forecaster
Carbon Intensity Predictor
AI Optimization Engine
Energy & CO_2 Reports / Dashboard
Cooling Control (RL Agent)
Compute Scheduler
HVAC, Chillers, Fans
Virtual Machines / Containers
Telemetry: CPU, Power, Utilization

Figure 4.1 AI-driven energy Optimization Framework

4.2 Architectural Components

1. Workload Forecaster: This module predicts short-term and long-term compute demand using machine learning models such as LSTM (Long Short-Term Memory) networks or ARIMA. It analyzes historical utilization patterns, application scaling behaviors, and time-of-day variations to estimate future load. The predictions guide proactive scaling decisions, reducing idle resource power waste and avoiding thermal spikes.

2. Carbon Intensity Predictor: Connected to external APIs like Electricity Maps or WattTime, this module monitors real-time and forecasted grid carbon intensity. By integrating these signals, the system schedules flexible workloads (e.g., AI model training, analytics) during low-carbon periods and throttles or migrates workloads during carbon-intensive hours—achieving "carbon-aware" computing.

3. AI Optimization Engine: This central brain hosts the multi-objective optimization layer combining predictive insights with reinforcement learning policies. It maintains a cost function that balances energy, carbon, and performance:

$$\boldsymbol{J = \alpha E_{total} + \beta CO2_{total} + \gamma(1 - QoS)}$$

where $E_{total}$is total energy, $CO2_{total}$is emissions, and QoSrepresents performance metrics such as latency or throughput.
4. Compute Scheduler: The scheduler applies the AI recommendations to allocate workloads across clusters, servers, or geographical regions. It leverages virtualization and container orchestration to adjust workload placement in real time. This module can interface directly with Kubernetes APIs to prioritize "green zones" (nodes with lower energy or carbon impact).
5. Cooling Control (RL Agent): The reinforcement learning agent adjusts cooling system parameters such as chilled-water temperature, fan speed, and airflow rate. Its actions aim to minimize cooling energy while ensuring rack temperatures remain below threshold limits. The RL agent continuously refines its policy based on real sensor feedback—like the DeepMind–Google model, but extended to interact with workload forecasts.
6. Telemetry and Feedback Layer: This component aggregates sensor data including CPU/GPU utilization, temperature, humidity, and power draw. The feedback is sent to both the forecasting and optimization layers to close the learning loop. Over time, this enables autonomous self-tuning of operational parameters.
7. Reporting and Dashboard: The decision outputs are visualized through an analytics dashboard displaying real-time energy consumption, PUE, CO_2 emissions, and savings achieved. It also supports auditing and regulatory compliance by recording decision logs and sustainability metrics.
This architecture shown in Figure 4.1 thus establishes a self-adaptive, carbon-aware control ecosystem for cloud data centers—balancing operational efficiency, environmental sustainability, and business cost optimization.
The implementation of the proposed AI-driven energy optimization framework involves integrating diverse data sources, machine learning models, and control mechanisms into a unified operational loop. While the conceptual model can be realized in multiple configurations, its core components follow a modular, scalable, and platform-agnostic design suitable for both hyperscale and enterprise data centers.

5. System Functionality

5.1 Data Inputs and Preprocessing

The system relies on four primary data streams:

1. **Compute utilization data**: Server CPU/GPU usage, virtual machine (VM) metrics, and job scheduling logs.
2. **Cooling and environmental data**: Temperature, humidity, airflow, and chiller or air handling unit (AHU) power readings.
3. **Energy cost and grid data**: Real-time electricity tariffs and carbon-intensity values retrieved from APIs like Electricity Maps or WattTime.
4. **Operational telemetry**: Data from IoT sensors, Building Management Systems (BMS), and Data Center Infrastructure Management (DCIM) tools.

These heterogeneous data sources are preprocessed using a data pipeline that performs cleaning, resampling, feature extraction, and normalization. Missing or noisy sensor data are handled through interpolation or Kalman filtering to ensure robustness.

5.2 AI Modules

1. Workload Prediction Model: Implemented using LSTM (Long Short-Term Memory) or Temporal Convolutional Networks (TCN), this module predicts short-term workload fluctuations. Predictions drive proactive resource allocation and power capping.
2. Carbon-Intensity Forecasting: The system fetches hourly carbon-intensity data through APIs. A regression or gradient boosting model can extrapolate short-term carbon trends, helping determine when to shift flexible workloads for minimal emissions.

3. Cooling Optimization via Reinforcement Learning: A Deep Reinforcement Learning (DRL) agent (e.g., DDPG, PPO) interacts with a simulated thermal environment or a digital twin of the data center.

- **State variables:** Rack inlet temperatures, server loads, humidity, and chiller power.
- **Actions:** Adjust cooling setpoints, fan speeds, and chilled-water temperatures.
- **Reward function:** Negative of total energy consumption plus penalties for violating temperature thresholds.
- The RL agent learns optimal cooling strategies by iteratively minimizing energy usage while preserving thermal safety.

4. Multi-Objective Decision Engine: The optimization engine aggregates inputs from workload forecasts and RL outputs, balancing three objectives: energy minimization, cost reduction, and carbon mitigation. A dynamic weighting system allows operators to tune priorities based on operational goals.

5.3 Integration Workflow

The framework can be deployed as a middleware orchestration service between existing virtualization layers and facility control systems.

- The Compute Scheduler interfaces with Kubernetes or VMware APIs for workload placement.
- The Cooling Controller integrates with the BMS via standard protocols (BACnet or Modbus).
- The AI Engine runs continuously on a dedicated edge node, retraining models weekly using newly collected telemetry.
- A dashboard interface displays live metrics such as PUE, total energy, and CO_2 savings, enabling human oversight.

6. Results and Discussion

Although the present study is conceptual, the proposed framework can be evaluated through simulation, digital-twin environments, or controlled pilot deployments within existing cloud data-center infrastructures. Evaluation focuses on quantifying energy savings, thermal stability, and carbon-emission reduction when the AI-driven coordination of compute allocation and cooling control is applied.

6.1 Evaluation Setup

A realistic testbed can be built using historical telemetry and synthetic workloads emulating diurnal utilization cycles. Cooling dynamics are modeled using manufacturer chiller curves and psychrometric correlations between heat load and ambient temperature. Grid-carbon intensity values are drawn from real-time datasets such as Electricity Maps or WattTime APIs. Baseline conditions represent traditional static operation—fixed cooling setpoints and first-come workload scheduling—against which the AI-optimized policy is compared.

6.2 Performance Metrics

The effectiveness of the proposed system is assessed using key indicators shown in the below Table 6.1 widely recognized in both academia and industry:

S No	Metric	Target Improvement (↓)
1	Power Usage Effectiveness (PUE)	From 1.6 to 1.3
2	Cooling Energy Ratio	20-30%
3	CO_2 Emissions (tCO_2)	15-25%
4	Electricity Cost (USD)	15-18%
5	Thermal Compliance (%)	>=99%

Table 6.1 Key indicators

Simulation studies and prior literature suggest that the joint optimization of compute and cooling yields additive benefits:

- **Compute-side savings**: ~10 % reduction via carbon-aware workload shifting and dynamic consolidation.
- **Cooling-side savings**: ~20 % reduction through reinforcement-learning control of chillers and airflow.
- **Overall energy reduction**: 15 – 22 % across typical operational scenarios.
- **CO₂ mitigation**: ~20 % decrease by aligning workloads with low-carbon grid periods.

Such outcomes directly lower operational expenditure and support corporate sustainability targets.
The conceptual results indicate that integrating reinforcement learning with carbon-aware workload management can deliver substantial improvements without hardware replacement. Moreover, adaptive models sustain efficiency under changing weather conditions and workload volatility. Beyond quantitative benefits, the framework enables predictive capacity planning and continuous sustainability reporting, providing actionable insights for both operators and policymakers.

7. Conclusion and Future Scope

This conceptual study presents an integrated framework for energy-efficient and carbon-aware cloud data centers, emphasizing the joint optimization of compute allocation and cooling management through Artificial Intelligence. The proposed architecture leverages predictive analytics, reinforcement learning, and real-time carbon-intensity forecasting to create an adaptive control system capable of self-optimizing operation. By unifying workload scheduling with cooling control, the model aims to minimize total power usage and operational costs while achieving measurable reductions in greenhouse gas emissions.
The concept demonstrates that future data centers can evolve beyond static infrastructures into autonomous, learning ecosystems—systems that not only respond to fluctuating workloads and environmental conditions but also anticipate them. Such an approach could lead to sustained improvements in Power Usage Effectiveness (PUE), reduced cooling energy, and enhanced carbon efficiency, contributing directly to global sustainability targets such as Net-Zero 2050 and the UN Sustainable Development Goals (SDG 7 and SDG 13).
Future work will focus on implementing this architecture within digital-twin environments and testing reinforcement learning agents in live operational settings. Integration with renewable energy forecasting, liquid-cooling systems, and carbon-trading mechanisms represents the next frontier for sustainable, AI-driven cloud infrastructure.

REFERENCES:

1. DeepMind. (2016, July 20). *DeepMind AI reduces Google data centre cooling bill by 40%.* Google Green Blog. https://blog.google/inside-google/infrastructure/deepmind-ai-reduces-google-data-centre-cooling-bill-40
2. Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J. V., & Maggs, B. M. (2009). Cutting the electric bill for internet-scale systems. *ACM Special Interest Group on Data Communication*, *39*, 123–134. https://doi.org/10.1145/1592568.1592584
3. Electricity Maps. (2025). *API for real-time and historical grid carbon intensity.* Retrieved from https://electricitymaps.com
4. WattTime. (2025). *Automated Emissions Reduction API documentation.* Retrieved from https://watttime.org
5. Google Sustainability. (2021). *24/7 Carbon-Free Energy: Methodologies and Metrics.* Google Research White Paper
6. Abel Souza, Shruti Jasoria, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. 2024. CASPER: Carbon-Aware Scheduling and Provisioning for Distributed Web Services. In Proceedings of the 14th International Green and Sustainable Computing Conference (IGSC '23). Association for Computing Machinery, New York, NY, USA, 67–73. https://doi.org/10.1145/3634769.3634812