Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Enhancing Cloud Sustainability through AI-driven Energy Management and Resource Consolidation

Chitiz Tayal
Senior Director, Data & AI, Axtria Inc

Abstract - The increasing cloud computing and artificial intelligence workloads have accelerated the need for energy and the emission of greenhouse gases from data centres. This work proposes an adaptive resource consolidation and predictive energy management integrated to adapt the power consumption while maintaining a defined Quality of Service (QoS). The design and implementation of simulation engines that are energy aware and the research that pertains to scheduling workloads that are carbon aware are discussed, followed by an integrated AI infrastructure that combines predictive workload energy demand, carbon intensity, and the dynamic consolidation of virtual machines and containers, which is closed with a reproducible evaluation framework based on well-known simulation tools. The combination of artificial intelligence consolidation and carbon-sensitive workload distribution, which is explored in literature and in the simulation with the defined case studies, acts in line with the stipulations in the service level agreements to reduce energy and carbon emissions significantly. The study closes with a practical exposition of the balance of yielded results, suggested improvements to the work, and prospective investigations, namely, coordination across multiple clouds and remaining resources.

Keywords - Cloud computing, green computing, energy management, resource consolidation, AI, carbon-aware scheduling.

I. INTRODUCTION

The critical role of cloud data centres in delivering various digital services is accompanied by their considerable demand for electricity. Recent industry assessments show that data centres not only emit high levels of greenhouse gases, but also show that they emit high levels of greenhouse gases [9]. Furthermore, the rapid decarbonisation of these centres is critical to the proliferation of cloud-enabled AI and the computational power it demands [6]. Efficient management and decarbonisation of these centres are critical to reduce the anticipated, significant CO_2-equivalent emissions from “hyper scale” expansion, and the provision of AI cloud services will bring the world by the end of this decade [8].

Since the focus of green computing research seeks to advance hardware efficiency, transformative cooling, and facility-level upgrades, the more untapped potential of software-layer strategies, particularly dynamic workload placement and workload consolidation, should not be overlooked, especially because these strategies can be implemented on current infrastructure [1]. The focus of this paper is AI-enabled energy management and resource consolidation, which encompasses the consolidation of virtual machines and containers, to bring cloud computing eco- sustainability. The paper integrates the pertinent literature, outlines a modular AI framework, suggests some assessment criteria, and highlights anticipated pros and cons.

II. BACKGROUND AND RELATED WORK

Simulation platforms like GreenCloud and CloudSim have contributed to research on energy efficiency with reproducible experiments, benchmarks, and power consumption models across packets, servers, and racks. GreenCloud, especially, is a packet-level energy-aware simulator extensively used in evaluating VM consolidation, DVFS, and network-aware policies [1]. Surveys on the energy efficiency in cloud computing and techniques classify approaches as infrastructure improvements, virtualization and consolidation, workload scheduling, and energy-aware software design [7]. A comprehensive review on the subject also emphasizes software-level mechanisms like containerization, server consolidation, and intelligent scheduling, such as prominent means to reduce energy per computation [2]. Most of these mechanisms result in resource underutilization, latency, and QoS trade-offs, thus the need for multi- objective optimization.

Recent research depicts the incorporation of carbon-awareness and data-driven algorithm selection, as unscheduled delay- tolerant tasks are scheduled for times and locations with lower carbon intensity. Work in hot carbon and similar venues outlines meta-algorithms and reinforcement learning techniques to select carbon- efficient scheduling policies for multi-site clouds [4]. Results from integrating AI into energy systems, particularly for demand forecasting, control, and optimization, suggest its suitability for cloud energy management as well [11]. Predictive

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

algorithms for workload patterns can correlate costs with energy and determine the carbon intensity of energy costs, making placement and consolidation decisions that minimize the total environmental cost and performance cost.

III. PROBLEM STATEMENT AND OBJECTIVES

Problem. Cloud operators must fulfill highly variable workloads with strict quality-of-service (QoS) requirements while minimizing energy consumption and carbon emissions. Static heuristics for consolidating and scheduling workloads struggle to adapt to non-stationary workload patterns and external signals, such as real-time carbon intensity of electricity and the availability of renewable generation [11].

Objectives.

- Develop an artificial intelligence- based energy management module that predicts workload demand and carbon intensity index signals, and advises energy-aware placement and consolidation actions.
- Include adaptive VM/container consolidation, which considers the tradeoff between energy savings and service-level agreements (SLAs).
- We will measure end-to-end performance (energy, carbon emissions, QoS, and consolidation overhead) in a controlled simulation, and show reproducible gains in performance compared to baseline heuristics.

IV. PROPOSED AI-DRIVEN FRAMEWORK

Monitoring & Telemetry Layer. A continually operational placement layer responsible for gathering resource usage (CPU, memory, I/O), per-host power estimates (or PHY measurements), network utilization, and external signals like real- time grid carbon intensity and price of electricity [13]. Telemetry also extends to tracking SLA metrics such as response time and error rates.

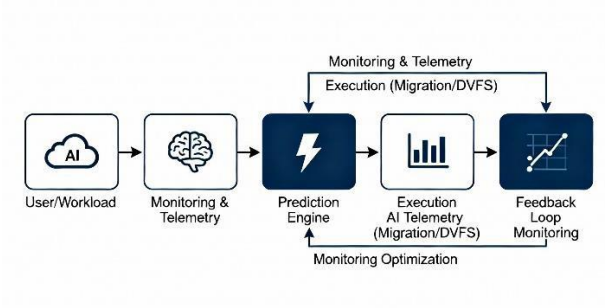
Prediction Engine. Based on time series forecasting (such as an LSTM/ Transformer regression or gradient-boosted trees, depending on the quantity of data), one can predict the short-term workload demand per application or tenant. Also, short-term predictions for carbon intensity and energy price [12]. The predictions occur as a horizon of minutes to hours to facilitate workload consolidation and/or shifting.

Policy Optimization Module (AI Planner). Utilizing predictions from the Prediction Engine Module will optimize placement and consolidation actions using a multi-objective approach. Candidate methods include:

1.
Reinforcement Learning (RL) that has a reward based on a combination of energy reduction and SLA penalties that discount rewards based on SLA use.
2.
Model Predictive Control (MPC) solves a constrained optimization over a forecasting horizon.
3.
A Meta-algorithm that will choose among a set of heuristic policies (e.g., bin-packing, energy-aware first-fit) using a bandit or follow-the-leader approach that will optimize in terms of energy and SLA.

It could also use carbon-aware objectives to prefer placement in lower carbon-intensity areas of the grid for non-latency-sensitive workloads.

Execution & Safe-Rollback Layer. Orchestrates the live migrations of workloads, automatic rescheduling of containers, DVFS (Dynamic Voltage and Frequency Scaling) changes, and power state transitions. Incorporates safety-monitoring of current states, and rollback strategies that attempt to bind the overhead of SLA violations and migration overhead [10]. Actions taken by the execution policy are reversible and bounded by the cost of a migration budget by throttling the execution policy.



Design Choices and Rationale

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Aspect	Description
Prediction Horizon	Short horizons (5–60 minutes) are effective for consolidation decisions as they limit stale actions and reduce unnecessary migrations.
Hybrid AI Choice	Use simpler machine learning models (e.g., gradient-boosted trees) for reliable demand forecasts when training data is limited. Use reinforcement learning (RL) or model predictive control (MPC) for policy optimization, where simulation can bootstrap safe-to-deploy policies.
Carbon Signal Integration	By incorporating carbon intensity or renewable energy availability data, the system can shift delay-tolerant tasks across time or geography to reduce emissions without affecting latency-sensitive workloads.

V. RESOURCE CONSOLIDATION STRATEGIES

Resource consolidation may reduce the number of active physical hosts by packing VMs/containers, enabling the powering down of idle servers or putting them in low- power states. The mechanisms for resource consolidation include the following:

- **Dynamic Threshold-based Consolidation:** Resource thresholds relad to the utilization of the host will trigger consolidation. While a simple approach, it is sensitive to the chosen thresholds, and appropriate thresholds may be difficult to determine [1].
- **Workload-aware Consolidation:** Resource consolidation happens based on workload characterization to minimize collocating workloads that interfere with each other. Workload-aware consolidation is also able to consolidate workloads that complement one another [2].
- **Predictive Consolidation:** Resource consolidation will use demand forecasting for workloads in order to avoid migrations for short spikes and to help with planning consolidations at low-risk times [4].
- **Carbon-aware Resource Consolidation:** Resource consolidation will prioritize consolidations that rely on less energy consumption and offload flexible workloads to cleaner energy regions or times when carbon intensity is high.

The adaptive schemes that combine predictive workload forecasts and resource consolidation policies, as shown in the literature and simulations, may outperform static heuristics in terms of energy and service level agreement metrics [13]. The AI Planner is designed to select consolidation targets and determine migration schedules while minimizing migration overhead and impacts to service level agreements for the programmer.

VI. EVALUATION PLAN

The paper suggests a validation process that includes simulation-based evaluations and (if possible) small-scale validation on a real test-bed.

Tools: GreenCloud and CloudSim would be appropriate for this evaluation; GreenCloud enables us to model power consumption at the packet level, and CloudSim gives experimental leverage when working with VM placements [5]. Employing both GreenCloud and CloudSim provides cross-validation of our results.

Workloads: It will test applications typical of both latency-sensitive web services (i.e., HTTP microservices), batch analytics jobs (i.e., MapReduce-like jobs), and ML training/ serving workloads [11]. This can also mean workloads that represent diurnal patterns, bursty arrivals, and event-driven processing spikes. To operationalize the carbon-aware evaluation, add real-world traces of carbon intensity to the mix (e.g., data from electrical grid operators and APIs).
Baselines: It will provide four baselines:
(a) no consolidation (i.e., a round-robin policy), (b) consolidation at some static threshold, (c) heuristic energy-aware consolidation, and (d) some state-of-the-art carbon-aware scheduler/approach from the literature [9].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Metrics:

- Energy consumption (in terms of kWh) as well as accessible CO_2- equivalent estimates (using carbon- intensity traces).
- Quality of Service (QoS) metrics and indicators to be monitored: average latency, 95th percentile latency, and service level agreement (SLA) violation rate.
- Number and overhead of migrations, including network traffic and migration-induced latency.
- Cost (if pricing or time-of-use electricity is a factor).

Experiment Scenarios:

- Normal day with variable demand.
- A day with high demand and frequent spikes.
- High carbon-intensity window (e.g., the evening mix of the grid) to put pressure on carbon-aware decisions.

Anticipated consequences of current literature: Previous research and empirical investigations demonstrate that optimized AI-guided consolidation and carbon-aware scheduling can significantly lower energy use and emissions while minimizing SLA

violations with appropriate tuning [10]. Findings from survey research suggest that these applications are viable and are increasingly being used.

VII. DISCUSSION - TRADE-OFFS AND PRACTICAL CONSIDERATIONS

Comparative Migration Costs and Savings. The process of live migration utilizes bandwidth and CPU resources, which may result in performance deterioration in the short term; the AI planner must weigh energy savings against the cost of migration and SLA risk [4].

Prediction Errors. Predictions can fail, leading to migrations that are unnecessary or worse, workloads that are under- provisioned; improving robustness through a conservative buffer or ensemble of forecasts may also provide more reliable migration decisions [13].

Data Privacy and Security. Telemetry and control planes are now the subjects of need for protection, and revealing decisions on workload placement may provide revealing information about the tenant's use of IaaS; this becomes a privacy concern. They would need some protection of telemetry, as well as orchestrating workloads based on role-based decisions.

Operational Integration. Implementing this system in production will require adequate integration with the cloud orchestration environments (e.g., Kubernetes or OpenStack), as well as involve extensive and careful testing [11]. Only then could the system be rolled out gradually, or through canary policies to minimize disruption.

Economic Incentives & Interaction with the Grid. When the cost of electricity and the carbon intensity vary, the best practice for cloud operators is to leverage their workloads economically. Clearly, this pushes incentive alignment, but also into regulatory and contractual areas.

VIII. CONCLUSION AND FUTURE WORK

This paper has put forward an artificial intelligence-driven framework. It combines predictive energy management, a framework for adaptive resource consolidation to enable cloud sustainability. By leveraging the combination of workload forecasting with carbon-intensity awareness and multi- objective policy optimization, cloud operators should be enabled to make significant strides in reducing their energy consumption and emissions, while still ensuring QoS is at an acceptable level. Through evaluation in simulated environments using GreenCloud/CloudSim and real carbon traces, a replicable framework allows for the benefits to be quantified.

Future directions should include multi- cloud coordination in the effort to reduce carbon in a global environment, tighter coupling of forecasts of on-site and grid- level renewable generation, and federated approaches for learning to enable policy improvement across operators without the need for sharing raw telemetry. One area of future research to explore is hardware- aware consolidation that accounts for accelerators such as GPUs and TPUs that have differentiated power profiles. Similarly, looking into mixed workloads that take into consideration the underlying hardware will be duly noted. Finally, field trials with cloud service providers and workload contractors will be a critical next step for the research, as it relates to validating operational viability over a prolonged timespan.

REFERENCES

[1] A. Ashraf and I. Porres, “Multi- objective dynamic virtual machine consolidation in the cloud using ant colony system,” *arXiv preprint arXiv:1701.00383*, Jan. 2017.

[2] G. Li, Y. Jiang, W. Yang, C. Huang, and W. Tian, “Self-Adaptive Consolidation of Virtual Machines For Energy-Efficiency in the Cloud,” arXiv preprint arXiv:1604.04482, Apr. 2016.

[3] A. Ashraf, B. Byholm, and I. Porres, “Distributed virtual machine consolidation: A systematic mapping study,” *arXiv preprint arXiv:1803.03094*, Mar. 2018.

[4] N. Akhter, M. Othman, and R. K. Naha, “Evaluation of Energy-efficient VM Consolidation for Cloud Based Data Center - Revisited,” *arXiv preprint arXiv:1812.06255*, Dec. 2018.

[5] V. Garg and B. Jindal, “Resource optimization using predictive virtual machine consolidation approach in cloud environment,” *Intelligent Decision Technologies*, vol. 17, no. 4, pp. 665-681, 2023.

[6] Y. Mehta, V. Lo, V. Mehta, K. Agrawal, C. T. Madabathula, E. Chang, and J. Gao, “Renewable Electricity Management Cloud System for Smart Communities Using Advanced Machine Learning,” *Energies*, vol. 18, no. 6, Art. no. 1418, 2025.

[7] M. Ghorbian, M. Ghobaei-Arani, and L. Esmaeili, “An energy-conscious scheduling framework for serverless edge computing in IoT,” *Journal of Cloud Computing*, vol. 14, Art. no. 52, 2025.

[8] Z. Miao, L. Liu, H. Nan, W. Li, X. Pan, Yang, M. Yu, H. Chen, and Y. Zhao, “Energy and carbon-aware distributed machine learning tasks scheduling scheme for the multi-renewable energy-based edge-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

cloud continuum," *Science and Technology for Energy Transition*, vol. 79, Art. no. 82, 2024.

[9] W. A. Hanafy, L. Wu, D. Irwin, and P. Shenoy, "CarbonFlex: Enabling Carbon- aware Provisioning and Scheduling for Cloud Clusters," *arXiv preprint arXiv:2505.18357*, May 2025.

[10] M. Alex, S. O. Ojo, and F. M. Awuor, "Carbon-Aware, Energy-Efficient, and SLA-Compliant Virtual Machine Placement in Cloud Data Centers Using Deep Q-Networks and Agglomerative Clustering," *Computers*, vol. 14, no. 7, Art. no. 280, Jul. 2025.

[11] T. B. Hewage, S. Ilager, M. A. Rodriguez, and R. Buyya, "A Framework for Carbon-aware Real-Time Workload Management in Clouds using Renewables- driven Cores," *arXiv preprint arXiv:2411.07628*, Nov. 2024.

[12] H. Wu, Y. Chen, C. Zhang, J. Dong, "Loads Prediction and Consolidation of Virtual Machines in Cloud," *Concurrency and Computation: Practice and Experience*, 2023.

[13] P. Naveen, Wong Kiing Ing, M. K. Danquah, A. S. Sidhu, and A. Abu-Siada, "Cloud computing for energy management in smart grid - an application survey," *IOP Conference Series: Materials Science and Engineering*, vol. 121, no. 1, Art. 012010, 2016.