PENINSULA PUBLISHING LLC

ESTIDAMAA
Vol. (2026), 2026, pp. 131-145
ISSN: 3078-428X

Research Article

Energy-aware and Sustainable Cloud Computing: Dynamic Scheduling, Pricing-Driven Resource Management, and Renewable Intermittency

Raed A Hasan[1,4, *], Omar A. Hammood[2], Mostafa A. Mohammed[3], Khalil F. Yassin[1], Mehdi Qahraman Fakhruldin[4], Waleed A. Hammood[5]

[1] *Renewable Energy Research Unit, Northern Technical University, Mosul, 34001, Iraq*

[2] *Department of Cyber-Security, Artificial intelligence College, Fallujah University, Fallujah, Iraq*

[3] *Department of Computer Science, Al-Imam Al-Adham University College, 10011, Baghdad, Iraq*

[4] *Department of Robotics and Artificial intelligence techniques, Al-Hawija Polytechnique College, Northern Technical University, Kirkuk, Iraq*

[5] *Faculty of Computer Science and Information Technology, Anbar University, Iraq*

ARTICLEINFO

Article History
Received 04 Jun 2026
Revised: 26 Jul 2026
Accepted 25 Aug 2026
Published 10 Sep 2026

Keywords
Cloud Computing,
renewable energy,
SLA,
resource management,
Dynamic Scheduling.

ABSTRACT

Data centers are now the largest users of electricity, carbon, and water, driving rapid cloud computing growth and a global shift to renewable-powered, sustainable cloud infrastructure. The large-scale integration of solar and wind energy creates fundamental operational challenges due to their intermittency, uncertainty, and non-dispatchability. This paper presents dynamic energy-aware cloud computing as a pathway toward sustainability by modeling renewable variability with service reliability assurance aligned workload scheduling with forecast green energy availability. Recent advances in energy-and carbon-aware orchestration (2023-2025) have been reviewed where critically missing gaps are unified frameworks integrating renewable forecasting with stochastic robust optimization SLA aware scheduling plus demand-and supply-side contingencies like storage backup provisioning plus adaptive load management. Based on these gaps an elawful operations model for the co-optimization of power use carbon footprints quality of service under time-varying energy conditions is described that uses elastic resource management supported by probabilistic solar/wind forecasts plus risk-sensitive scheduling. New directions for more sustainable clouds include decentralized speculative renewables infrastructures; this paper also describes some concepts in that area. This work lays out a structured foundation toward next-generation green cloud platforms moving beyond static efficiency metrics into predictive reliability-driven environmentally responsive cloud operations.

1. INTRODUCTION

Every day, Cloud computing has been at the forefront of global technology advancement for over a decade, enabling the distribution of computation, communication, and storage to clients around the world. Nevertheless, the global cloud IT infrastructure operates with a monthly billing of over 80 billion USD, consuming more than 2% of the global electricity supply and contributing to more than 0.5% of the global carbon emissions [1-3]. Data centers, the key components of cloud computing, rank second only to real estate in terms of water consumption among urban industries. Consequently, the energy utilization, carbon emissions, and water consumption in cloud data centers have attracted widespread concern. To enable sustainable development in cloud computing, government regulations have been established, and considerable research effort has been devoted to lowering the Power Usage Effectiveness (PUE) along with energy consumption. However, the adoption of green technology in cloud computing remains low [4-5].

Strictly speaking, the energy infrastructure of cloud computing has evolved into smart grid architectures capable of forecasting electricity prices accurately. Furthermore, renewable solar and wind energy sources have grown rapidly over the past two decades despite their intermittent supply patterns. Some large cloud providers are already utilizing green energy at a high percentage. New energy-related services and practices have also emerged in several cloud firms, including carbon emissions indicators, periodic sustainability audit reports, and energy-aware workload placement. To meet these

*Corresponding author email: raed.isc.sa@ntu.edu.iq
DOI: https://doi.org/10.70470/ESTIDAMAA/2026/009

growing trends and client requests, dynamic Energy-Aware Cloud Computing (EACC) and Green Cloud Computing (GCC) have been proposed, addressing the challenge of temporal task scheduling and static workload distribution for the economic, carbon, and water footprint. The principles of green cloud computing have been widely accepted; however, few policies, either theoretical or practical, have been developed to guide the design of the cloud resource scheduling strategy that integrates either carbon intensity or water consumption [6].

The reformation of cloud service from the current concentrated model to the decentralized model is also receiving thorough consideration. Such a transition permits much wider-ranging policy- and design-oriented guidance on cross-community cloud service outsourcing via social profit, rendering the timely, accurate, and concrete measurement of cloud service community direct social transaction cost indispensable. A carbon-driven dynamic resource allocation policy is at the center of this set of proposals, and understanding the social transaction cost of cloud services produced by various designs is the first essential step toward realizing the detailed mechanism. These factors comprising multi-cloud service community, inter-community carbon-emissions-boundary, task- and workload-republication are time-varying and affect the community transaction cost significantly. Consequently, under these trends that furnish abundant environmental and economic signal information, it is timely and crucial to investigate proper energy-aware scheduling and resource-allocation policies that utilize both cloud workload features and renewable-energy information [7-10].

2. BACKGROUND AND MOTIVATION

Cloud computing has rapidly become a crucial component of the global economy, with major players investing billions of dollars to support the robust demand for cloud services. In 2020, the public cloud services market was valued at over $370 billion and is projected to grow to nearly $830 billion by 2025. Along with the rising energy consumption of cloud infrastructures and services, however, renewed concerns have arisen over the environmental sustainability of cloud computing (Kumar Garg et al., 2009). More than 70% of enterprises use cloud services because they can deploy compute capacities at scale and manage them flexibly at a low cost. Despite constant improvements in energy efficiency, global data centers consume over 1,000 TWh per year around 2% of global energy consumption—and the demand for cloud services is expected to keep increasing [11-12].

Many data center operators pledge to mitigate climate change by deploying green cloud computing infrastructures and services to minimize their carbon pollution while meeting Service-Level Agreement (SLA) commitments. However, the widespread adoption of such infrastructures faces significant technical barriers. Many green cloud services rely heavily on renewable energy sources, which generally provide limited and variable energy in most grid regions. Forecasting these renewable supplies and the extent to which they can support service demand remains challenging. Reduced carbon emissions alone cannot ensure climate protection, as global warming is mainly driven by the cumulative carbon released into the atmosphere over time, and regions differ widely in their generation mix, carbon intensity, and grid-carbon pattern [13-15]. Addressing green cloud planning and scheduling without adequate treatment of task-release times, renewable-energy forecasting capabilities, grid-carbon forecasting, and full knowledge of workload types, priority requirements, and volume continues to be an open problem (See Table I).

TABLE I. THE RESEARCH STUDIES FORM 2023-2025 ON ENERGY-AWARE & SUSTAINABLE CLOUD COMPUTING

Authors	Year	Method / Model	Main Contribution	Identified Research Gap	Drawback / Limitation
Alex, M., Ojo, S. O., & Awuor, F. M.	2025	CA-VM-DQN (Carbon-Aware Deep Q-Network VM placement)	Carbon-aware, energy-efficient and SLA-compliant VM placement using deep reinforcement learning.	No explicit modeling of real-time renewable generation or forecasting uncertainty.	Uses carbon intensity as proxy; renewables intermittency not directly modeled.
Pisa, I.	2025	Energy-aware hybrid cloud orchestration framework	Aligns cloud-native services with clean-energy-available infrastructures.	Lack of renewable supply forecasting and dynamic workload shifting.	Architecture-focused; limited quantitative SLA/energy evaluation.
Bostandoost, R., Lechowicz, A., Hanafy, W. A., Shenoy, P., & Hajiesmaili, M.	2024	LACS (Learning-Augmented Carbon-aware Scheduler)	Dynamically scales resources to reduce carbon footprint of workloads.	Renewables treated indirectly; no explicit renewable intermittency optimization.	Requires accurate workload prediction; renewable uncertainty ignored.
Hewage, T. B., Ilager, S., Rodriguez, M. A., & Buyya, R.	2024	Renewables-Driven Core (RDC) framework	Couples cloud execution with renewable-powered cores to reduce brown-energy use.	No integrated solar/wind forecasting or SLA-risk modeling.	Prototype and simulation only; no large-scale operational validation.

Saad, Z., Yang, J., Leung, H., & Drew, S.	2025	FL-ECP (Federated Learning Energy Consumption Predictor)	Predicts container energy use for carbon-aware orchestration without centralized data.	Prediction-oriented; lacks closed-loop renewable-aware scheduling.	Does not directly control or reschedule workloads.	

3. DYNAMIC SCHEDULING WITH FORECASTED RENEWABLE ENERGY

The renewable energy industry has a strong interest in providing cloud companies with sustainable power. However, sustainability policies are difficult to enforce due to limitations imposed by energy intermittency, which varies across time spans from seconds to weeks. Therefore, intermittent energy sources are often combined with the traditional grid network to enable flexible renewable adoption (Caux et al., 2018). An alternative approach to mitigation involves the integration of a prediction model used for the forecasting of availability. Several models for predicting solar viability or wind capacity can be categorised as deterministic or probabilistic. Capacity assessments made through dual-prediction horizons help distinguish between short-term and long-term energy availability. Analysis of these predictions yields processing estimates capable of aligning operational scheduling with energy availability and thus improving renewable content [16-20]. A benchmark methodology demonstrating the importance of renewable energy prediction for sustainable cloud computing through scheduling is applicable to heterogeneous time-varying workloads that can be parallelised accurately. A set of known workloads remains scheduled within specific deadline ranges under a grid-enabled cloud system without renewable energy integration or renewable energy forecasting, producing significant improvements (As shown in the figure 1).

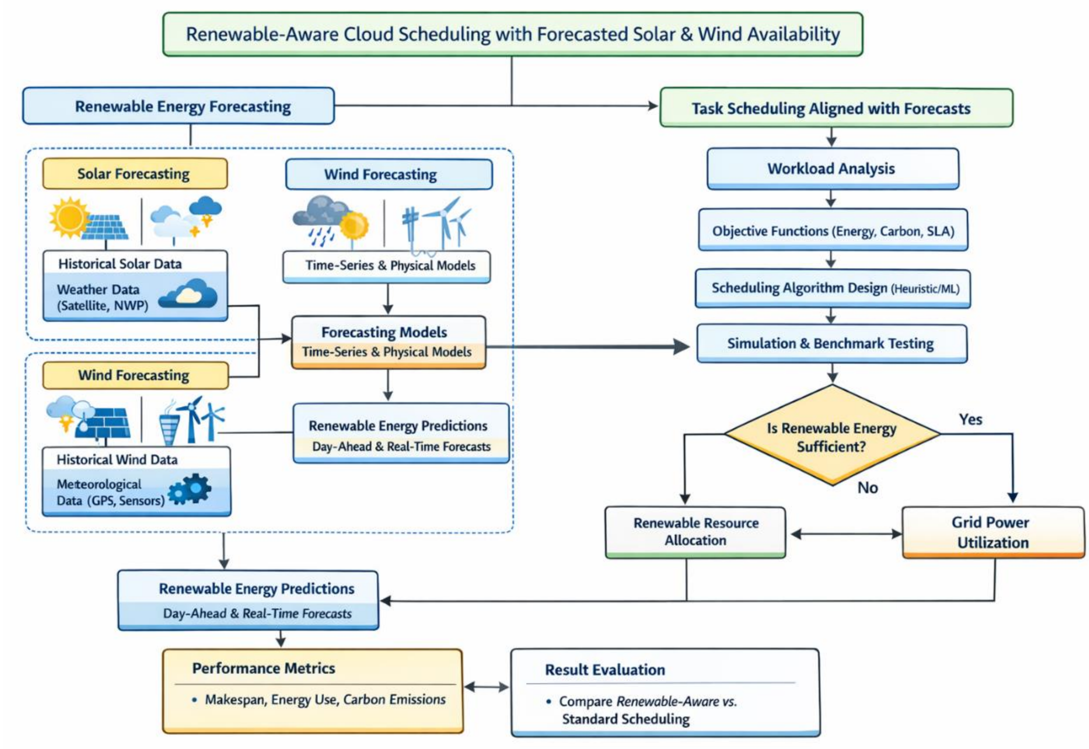

Fig. 1. Renewable energy cloud scheduling with forecasting solar and wind availability

3.1 Forecasting Solar and Wind Availability

Renewable energy sources are critical components of sustainable cloud computing because they mitigate the harmful effects associated with energy production [21]. Two important classes of renewable energy sources to consider are solar and wind, which are also variable in production. Therefore, renewable energy forecasts can play a crucial role in informing cloud operations. Forecasting components include prediction horizons, forecasting models, accuracy metrics, and data sources.

The prediction horizon indicates the time period in the future for which the resource availability value is to be predicted. An effective prediction horizon should be aligned with the scheduling policy where longer-horizon predictions are more beneficial for static scheduling and shorter-horizon predictions are effective for dynamic policies. An hourly prediction horizon may suffice for many cloud applications. In the long term, ultra-short-term expectancy can be predicted at one- or a few-time step(s) ahead to accommodate intermittent renewable supply. Solar and wind availability can be predicted using both time series-based and physical-based models. Time series-based models are employed when numerical weather data is not available while physical models, such as the Persistence model, the Numerical Weather Predictions (NWP) models, address short-term, intra-hour variations and regional aspects. The accuracy of the forecasting models can be evaluated by a variety of metrics based on the forecast requirement, for example, Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). These metrics measure the magnitude of error between forecasted and actual values across a certain time period. The MAPE is commonly adopted since it provides supplementary information to actual values through a percentage [22-25].

Solar and wind availability forecasting relies on historical input data to train prediction models. Historical data from different locations can be found through various sources like global solar radiation databases, satellite data, state-backed meteorological stations, or social platforms that collect energy usage data from local users.

3.2 Task Scheduling Algorithms Aligned with Renewable Forecasts

Adopting forecasted availability of renewable energy offers further efficiency and sustainability advantages. Scenarios, objective functions, constraints, and approaches for integrating energy predictions into cloud workload scheduling are systematically examined and classified, addressing system independence and priority treatment (Wang et al., 2018).

3.3 Case Studies and Benchmarking

To illustrate the effectiveness of integrating forecast-driven scheduling policies, three case studies focus on parallelized Hadoop jobs, their completion time savings, and the additional work achieved under time-limited conditions. Scenarios with workloads originally scheduled respecting availability prediction establish a performance baseline [26]. Completion time gains further increase when forecasts confident enough to reject jobs are applied. When the use of dynamic scheduling instead of static schedule freezing after job arrival serves as the baseline, renewable-aware workload scheduling ensures a consistently larger workload processed within 8 hours. A benchmark campaign comparing renewable-aware and standard scheduling policies empirically evaluates the performance gain. Similar to prior studies, input workloads consist of eight instances of typical bank jobs executed on a 64-node Hadoop cluster. The two policies maintain the same scheduling priority and consideration of resource requirements; only the renewable energy aspect varies.

A detailed simulation environment models workload and renewable energy supply, enabling the evaluation of scheduling policies without reliance on a specific cloud platform. Workload characteristics derive from the 2013–2014 Google cluster-trace dataset. Expectations under traditional infrastructure without local renewable generation inform the modelling of service availability and its time correlation. To accommodate the temporal nature of renewable energy supply, arrival patterns are extended to the supple character of jobs to enhance realism. The renewable energy simulator implements both day-ahead prediction of hourly photovoltaic energy generation and minute-by-minute forecast for wind power production. Models fitted to typical monthly generation profiles determine scheduled workloads corresponding to different availability percentages and time scales. A bi-directional Long Short-Term Memory network captures seasonal variations and autocorrelation in the building data and generates realistic hourly profiles. The global Positioning System-based database of weather stations underpinning the wind energy forecast comprises near-continuous readings ensuring national-level continuity.

4. ENERGY-AWARE RESOURCE MANAGEMENT AND ELASTIC SCALING

Modern cloud computing has enabled many new paradigms such as Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS) that are widely adopted by the community. Nevertheless, the rapidly growing service pr ovisioning, processing, and storage demands of SaaS, PaaS, and IaaS applications have created challenges for cloud providers as they struggle to meet the increasing Service Level Agreements (SLAs) required by their customers [12]. Additionally, the average Annual Total Cost of Ownership (TCO) of cloud computing has steadily risen, increasing from approximately $31,000 USD in 2009 to over $100,000 USD in 2012.

Cloud providers still lack a systematic understanding of the parameters affecting the TCO of cloud-based services and their respective α-factor influence on the four TCO components: acquisition cost, maintenance cost, operation cost, and termination cost. Furthermore, the arrival rate of the Service Requests (SR) generated from the global user base of cloud-based services varies considerably due to time zone and geographic factors, which creates complex workload characteristics. Cloud computing consists of three main tiers the data store tier, the middleware tier, and the client tier. As cloud computing has gained visibility, service requests are expected to increase in a J-Curve manner during the initial stage of cloud Application Service Provider (ASP) deployment. The lack of research on energy-awareness for establishing SLA-based Service Pricing (SP) in Cloud computing therefore presents a promising opportunity for investigation and exploration (As shown in the figure 2).

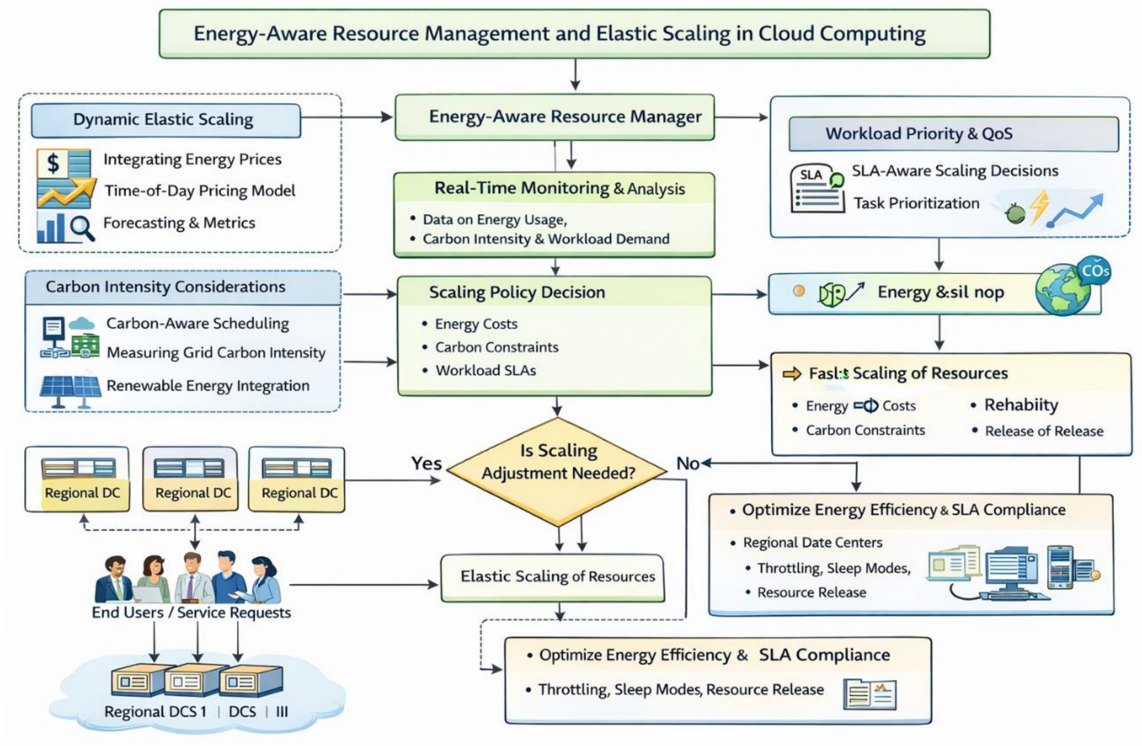

Fig. 2. Resource management and elastic scaling in cloud computing

4.1 Integrating Energy Prices into Scaling Decisions

Cloud providers often face significant operational energy expenses, which, combined with increasing environmental regulations, can constitute a considerable portion of total operational costs. Even if the energy costs may still be significantly lower than total cost, cloud providers can save on OPEX the reduction of which translates to increased profit. To reduce energy costs, cloud providers can adopt energy-aware resource management policies. One approach entails elastic scaling of virtual machines, adapting resources at runtime. Determining whether to scale up or down, and by how much, in a time-varying pricing context represents a challenge. The literature often assumes time-independent prices or that all providers use the same price model.

Price-aware resource management policies and the corresponding CloudSim simulator have been developed to govern elastic scaling policies based solely on energy cost. Such policies leverage the time-of-day electricity pricing model offered by some utilities. When considering CloudSim for energy-aware resource management, scaling decisions are also influenced by service level agreements (SLA) and revenue impact, which are unaffected by energy cost. Cloud providers generally charge clients based on SLA and are beholden to SLAs that constrain scalability. Developed solutions therefore focus on time-dependent scaling models that incorporate grid energy pricing while accommodating workload prioritization [6]. Tools to analyze past alterations in CPU frequency scaling and task allocation across cloud providers, W-Tuner and W-Track, have been created to detect pricing and availability patterns. A performance-based pricing model compatible with different CPU-bounded workloads has been proposed for systems incorporating variable-frequency operation.

4.2 Carbon Intensity Considerations

Data centers constitute an important contributor to global energy consumption and greenhouse gas emissions. Although several legislative measures and economic incentives have been adopted, cloud providers are still faced with substantial pressure to transition towards greener operations due to increasing public concern regarding climate change and the environmental impact of human activities, along with the goal of achieving climate neutrality across the European Union. Various methodologies have been proposed for enhancing environmental sustainability, due to their limited life-cycle carbon and resource footprints, cloud providers are still faced with substantial pressure to transition towards greener operations. Strategies such as Carbon-Aware Computing [27], Energy-Efficient Scheduling, Carbon Offset Tracking, Circular Data Centers, and Climatology-Aware Scheduling aim to inform computation at times and places that maximize carbon savings and transition the energy-hungry cloud computing ecosystem toward a sustained and climate-neutral state.

Cloud services are hosted by distributed, interconnected data centers and scale elastically to match workload demand, which is characterized by intense, bursty, and unpredictable patterns. These properties enable flexible elastic scaling, the distribution of requests to the nearest or greenest center, and the possibility of scheduling start times for incoming jobs when resource requirements can be satisfied. The cloud computing ecosystem is positioned within the energy-intensive IT sector. Cloud providers are motivated to reduce energy consumption and carbon emissions. Dynamic policies that incorporate forecast information and align service provisioning, workload execution, scaling decisions, and request scheduling are paramount for sustainability [28].

Electricity consumption, the impact of integrated renewable energy generation, and grid carbon intensity constitute fundamental criteria that considerably influence sustainability efforts, particularly in the cloud computing domain. The price of renewable energy is poorest and total operating costs are highest at 0, while the carbon footprint is reduced and normally zero at midday when the carbon footprint is reduced and normally zero.

4.3 Workload Priority and Quality of Service

Cloud service provisioning involves a trade-off between Service-Level Agreement (SLA) conformance and energy expenditure since cloud service performance is often elasticity-aware and subject to Quality-of-Service (QoS) levels (Lučanin et al., 2014). Moreover, users expect acceptable Value-Added Services (VAS) such as message delivery, notifications, and backups tailored to cloud app types, thus establishing priorities that need to be considered during resource-capacity determination. Non-availability during the delivery of VAS can lead to customer loss or dissatisfaction, making it crucial for service providers to take Quality-of-Service (QoS) levels into account and adjust platform energy consumption without compromising SLA fulfilment or overall service-level degradation.

The design of energy-aware dynamic cloud service provisioning and associated elastic-scaling resource management strategies typically requires characterizing software workload and VAS qos models. These can leverage predictive information about the workload while customer time-utility functions describe service values in line with cloud delivery. The models thus define cloud-resource planning under dynamic-utility-time-cloud-capacity systems that accommodate elasticity in resource allocations and service time while maintaining resource on-off states compatible with sustainability units [6].

4.4 Framework Architectures and Policy Frameworks

In centralized control, the information collected from participating cloud components is sent to a central resource manager, which autonomously decides and allocates resources. This approach is simple and has a clear architecture but creates a single point of failure and is difficult to scale. A centralized controller can still exist in a distributed architectural model, since the definition of the architectural model does not restrict the existence of a centralized controller. The distinction between these models is furthermore blurred by the emergence of hybrid architectural approaches. These approaches attempt to combine the advantages of both centralized and decentralized models while reducing their respective disadvantages. In distributed resource management, each controller obtains information from its corresponding cloud component and independently optimally allocates resources according to its resource management policy. Such a controller can be responsible for just one type of resource, leading to a distributed architecture [29,30].

5. MODELING INTERMITTENCY AND RELIABILITY OF RENEWABLES IN CLOUD OPERATIONS

Cloud computing and data centers are migrating rapidly towards renewable energy, with the latest research and industrial developments making energy-aware and sustainable operations possible. However, the integration of renewables complicates the management of energy usage and the scheduling of workloads, significantly amplifying the problem of intermittency. All renewable energy sources, including wind and solar widely considered the most promising in the cloud computing domain exhibit varying power generation over time. Wind systems constantly switch on and off, causing large variations in power input; there are entire weeks of zero output. Meanwhile, solar power generation is limited to daylight hours and is subject to cloud cover. Even where the overall renewable share is high, the intermittency of these sources can rise to 100% for substantial periods of time. This situation results in additional consumer costs, disruptions to the minimizing of Total Cost of Ownership (TCO), and violations of Service Level Agreements (SLAs) or Quality of Service (QoS) metrics.

Several objective-driven mathematical optimization methods can be implemented to model the intermittency and reliability of renewable energy resources at multiple levels of the cloud architecture. Cloud production decisions can therefore be treated either as a stochastic or as a robust optimization problem, where the uncertainty set corresponding to the renewable input measures the scale of the intermittent energy variation. A more practical alternative is to establish probabilistic requirements that indicate the minimum share of each request that can be honored in a defined amount of time, accompanied by various risk measures quantifying the likelihood of non-compliance or additional objective functions aimed at further mitigating the risk. Other technologies are also available to support the cloud renewable integration process: supply-side or demand-side backup of either energy or capacity, battery storage, and load-shedding strategies can all help to ensure uninterrupted service during periods of unavailability.

5.1 Stochastic and Robust Optimization Approaches

The optimization of performance and efficiency of cloud-computing environments is essential for green cloud. Cloud services are optimized for exorbitantly high gains, imposing large energy–environmental footprints to cloud providers and users and a negative impact on sustainability. Elasticity enables the optimization of cloud services, but the green cloud impact analysis of elasticity and additional scenarios is limited. Sustainability-contingent elastic cloud service approaches need to be developed. Scheduling and resource-management problems with a focus on environmental energy consumption, renewable resources, and different levels of elasticity were formulated and evaluated, leading to insights on the cloud under these conditions. The elaboration of theoretical cloud-model elastic cloud-service systems enhanced this investigation by simplifying the workload to a single virtual machine and focusing on high-order elasticity under power.

The increases in data-center energy requirements are accompanied by worsening energy–environmental impacts, rising costs, intensive competition, and more stringent regulations prompting a shift to the green cloud. Despite advances toward green clouds, large additional energy footprints and carbon emissions are still often accumulated. Existing cloud frameworks and approaches focus principally on optimizing the availability and performance of services. Cloud computing requires additional consideration and monitoring such as energy consumption. Gardening system approaches have been implemented to control environmental energy consumption because the conventional cloud-monitoring system ignores this requirement. Cloud resource-management and scheduling systems under energy, cost, and environmentally sustainable concerns were discussed. Controlling other parameters without efficient resource allocation and scheduling maximizes the additional server energy without regard to the energy supplied by renewable resources [20].

5.2 SLA Guarantees under Renewable Variability

To ensure the availability and performance of cloud resources provisioning that leverages renewable energy supply, cloud providers must offer service level agreements (SLAs). Such guarantees become even more complex when 100% renewable energy consumption is targeted, as electricity from renewable generators cannot be dispatchable. Therefore, cloud providers need to establish SLA provisions that accommodate uncertainty in renewable generation and that comply with renewable consumption objectives.

The difficulty of SLA provision is exacerbated in a fully renewable energy context because electricity from renewable generators is inherently non-dispatchable. Consequently, the integration of renewable resources deteriorates cloud service reliability, which is measured in terms of availability and performance provision together with temporal flexibility. A wide array of reliability metrics has been developed to mitigate this challenge.

When renewable variability affects system operation, ensuring the desired level of reliability requires cloud providers to complement their service offerings with appropriate risk measures and mitigation mechanisms. Such measures cover supplementary energy and storage inventory provisioning to address resilient through backup supply demand, energy curtailment and task shedding processes for sustainability, and planning concepts alongside demand response schemes to act as a supplementary variable. Sufficient supply assurance against provision failure constitutes a common basis whereupon the establishment of reliable SLA provisioning and corresponding complementary solutions addressing renewable variability assurance take place [6].

5.3 Demand-Side and Supply-Side Contingencies

Cloud computing is widely regarded as a more efficient way of storing and processing data and is believed to be less harmful to the environment than traditional computational resources. However, cloud providers are still energy-intensive operations, which can be viewed as problematic given public concerns regarding energy consumption and global warming. Current policies governing the (non)renewable nature of the resources that energize cloud operations are mostly static, non-forecast-based, or based on estimates that cannot accommodate hourly variance in energy prices or computer resource requirements.

Dynamic scheduling, energy-aware resource management, and storage-driven policies all help to minimize energy consumption and carbon emissions and broaden the scope for integrating renewable resources into cloud operations. Availability of backup energy (e.g., generators, batteries) can ease the transition to renewable resources, while storage-driven decisions can determine whether residual energy should be saved or the workload shed.

Integrating renewable energy into cloud operations can reduce carbon emissions and energy costs but introduce significant variability, complicating the cloud's ability to meet service-level agreements (SLAs). Considering these uncertainties, policy control of cloud operations can be classified as either stochastic determining the best transmission schedule while accounting for renewable-generation forecasting uncertainty [26] or robust optimizing the schedule and then selecting the appropriate back-up generator upon discovering new information about future generation.

6. INNOVATIVE AND EMERGING CONCEPTS IN GREEN CLOUD

Building on the previously discussed techniques, further innovative concepts can enhance the sustainability of cloud platforms from a longer-term perspective. One novel option is to deploy data centers equipped with solar panels in low-orbit space instead of on the terrestrial surface. Satellite-based solar power harvesting could yield significantly greater energy generation capacity than terrestrial systems, due to the invariance of sunlight illumination and the absence of atmospheric attenuation. Furthermore, near-constant energy production permits the elimination of energy storage facilities, as satellite-generated power can be directly delivered to ground stations without interruption. However, launching satellites incurs substantial investment and operational expenditure. Reducing the cost of inter-orbit resource transportation also proves economically effective. Cost analyses reveal that, for earth-synchronous orbits, a generation and satellite-aided utilization cost parity remains achievable even with costs exceeding the current terrestrial scale by a factor of hundred-and-fifty. Long-term roadmap studies and feasibility evaluations should thus explore the potential to extend ground-based systems through high-altitude airborne solar panels while simultaneously establishing the first generation of satellite-assisted alleviate the high operating cost of satellite-based systems and improve the terrestrial-affordability threshold for satellite-assisted utilization. This speculative extension falls within the theoretical reach of current spacecraft capabilities, enabling progressive deployment and testing to verify the technical soundness of the latest-generation concepts [23].

6.1 Space-Based Solar-Powered Data Centers: Potential and Challenges

Space-based solar-powered data centers offer several advantages, including ubiquitous energy availability and the potential for safe, high-temperature waste heat release. Data centers utilizing this model remain viable even when on-ground facilities face energy shortages or shutdowns. Operating in space largely circumvents infrastructure-water cooling issues, as energy should be easily available through photo-voltaic conversion and waste heat can be dissipated without risk of environmental pollution. Data centers at geopolitically stable locations and/or favorable climate zones could also support operating flexible scheduled tasks during time zones with high demand.

While the concept of worldwide wireless space-based solar-power delivery via microwaves is promising, the related technology has yet to advance sufficiently to support practical deployment. Even for the most advanced hand-held microwave components, the required collector aperture remains prohibitively large in both diameter and covered surface, likely exceeding several meters. Considerable research is needed to explore opportunities beyond improving on-ground integration performance for the microwave energy-pooling approach; operating at a shorter wave-length would lead to a smaller collector surface for the same emitted energy.

Table I summarizes the relevant energy-cost trade-offs. Two zones appear sufficiently wide for on-ground renewable-energy augmentation to be considered. Within these zones, capital expenses augmenting on-ground renewable sources could be traded against space deployment acceleration efforts
targeting wireless energy-transfer to Earth.

6.2 Long-Term Vision and Feasibility Assessments

The development of hybrid solar-wind energy systems has enabled significant energy cost savings by exploiting seasonal variations in resource availability. In particular, solar energy significantly reduces grid energy costs during the summer in temperate climates, while wind energy effectively lowers grid energy costs during winter months. Further cost synergies can be obtained between generation and demand-side resources. Flexible electrification and strategically selected flexible loads can utilize excess renewable energy generation to offset grid energy costs. It has been shown that grid interactive heat pump water heaters (HPWHs) can be employed to obtain similar resource synergies. By grouping appropriate data center tasks based on slurry characteristics and/or permitting scheduled task sizes, additional load-shedding capability can be provided when not in priority mode. Some flexible capacity is generally required for low-cost participation in energy market structures. Such markets, where available, can be effectively utilized in conjunction with demand-side flexible load management.

7. EVALUATION METHODOLOGIES AND METRICS

Cloud computing infrastructures are continuously growing to increase the quality and paradigm of service to their users. Nevertheless, studies have revealed that the cloud infrastructure has a remarkable impact on the greenhouse effect; CO2 emissions reduction must be targeted.

Energy efficiency metrics used to evaluate the energy efficiency of cloud infrastructure in general and cloud data centers, such as Power Usage Effectiveness (PUE), Data Center energy Performance (DCeP), energy proportionality and utilization efficiency are presented. The limitation of many of them is the absence of independence from specific equipment during the evaluation; additionally, some have only qualitative definitions.

Emissions accounting methods and carbon footprint analysis as important aspects of the greening process are investigated since ecological sustainability is a major concern. The lifecycle assessment, emissions scope classification and carbon footprint accounting method are described to support such analysis in cloud infrastructures. A lot of strategies omit the emissions accounting, which relates to the carbon footprint and is one of the first steps towards sustainability.

7.1 Energy Efficiency Metrics

Support for green cloud computing is crucial; nevertheless, diverse eco-friendly computing principles and measurements exist, leading to interpretation difficulties. Some widely-used energy efficiency metrics for assessing overall system energy consumption and utilization include power usage effectiveness (PUE), data centre energy proportionality (DCeP), energy proportionality, and utilization efficiency. PUE measures energy consumption relative to computation output, while DCeP links energy use with work delivered. Energy proportionality denotes variations in energy use with load fluctuations, essential for clouds. Utilization efficiency captures effective versus theoretical resource occupancy [5].

7.2 Emissions accounting and carbon footprint

Data centers at leading cloud providers process zettabytes of data yearly and account for more than 1% of global energy consumption. Sustainability issues focus on two dimensions: the environmental footprint associated with energy use, and the need for forecast-informed scheduling policies that reduce the carbon footprint without incurring excessive costs associated with lower-energy-price periods. The research has advanced the design and comprehension of cloud services across a broad spectrum of elastic workloads.

Cloud computing adoption is increasing, driving faster consumption growth than for other electronic services. Utilization updates every few hours accelerate the transition to low-carbon sources, inhibiting the use of cheaper energy from the grid. Inter-box scheduling heuristics with simpler models support a much broader range of SLAs—seconds to weeks—and dynamics. Techniques that proactively align scheduling and pricing enable cloud providers to balance temperature, thus avoiding outages in the current climate scenario [16]

7.3 Economic and Total Cost of Ownership Considerations

Data centers demand large quantities of energy, which accounts for around 10% of global carbon emissions. Over-provisioning for peak demand further exacerbates the environmental footprint of cloud services, and hence there is a strong need for dynamic policies to match scheduling decisions to available supply. A comprehensive framework is needed to address energy-aware management of cloud resources in the presence of dynamic elastic workloads, encompassing the integration of energy-cost and carbon-impact criteria into resource-allocation decisions, the modelling of renewable-resource intermittency and its impact on operations, and the specification of service-level agreements (SLAs) to guarantee quality of service.

Next, we present experimental results that assess the effect of integrating energy awareness, renewable forecasting, carbon sensitivity, and SLA-driven control into cloud resource management. The experiments are structured to transition incrementally from a conventional cloud baseline toward increasingly sustainable operational models while capturing system behavior at multiple levels: infrastructure utilization, workload performance, environmental impact as well as economic efficiency. The baseline scenario sets reference behavior for a non-energy-aware cloud in terms of active virtual machines, utilization, energy consumption, and response time. Subsequent scenarios progressively introduce renewable-aware scheduling, elastic scaling, carbon-aware allocation and forecast-driven orchestration. This layered design allows isolating the contribution of each mechanism and observing how operational decisions evolve when renewable intermittency, grid carbon intensity and SLA constraints are explicitly considered.

The figures together give a comprehensive view of sustainable cloud operation at the micro level (VM activity, CPU load, response time SLA violations) and macro level (renewable penetration carbon emissions total cost of ownership sustainability score). They also capture the dynamic nature by providing results over time slots workload classes regions forecasting horizons. This structure allows for quantitative comparison between standard and proposed models showing not only reductions in energy use and emissions but also improvements in reliability workload prioritization economic efficiency. Overall, these results prove how renewable-aware risk-sensitive scheduling changes cloud platforms from being static consumers of energy to adaptive environmentally responsive computing ecosystems. figures from 3 to 12 showing the defiance method used, and the ring of using this method to solve the research gap,

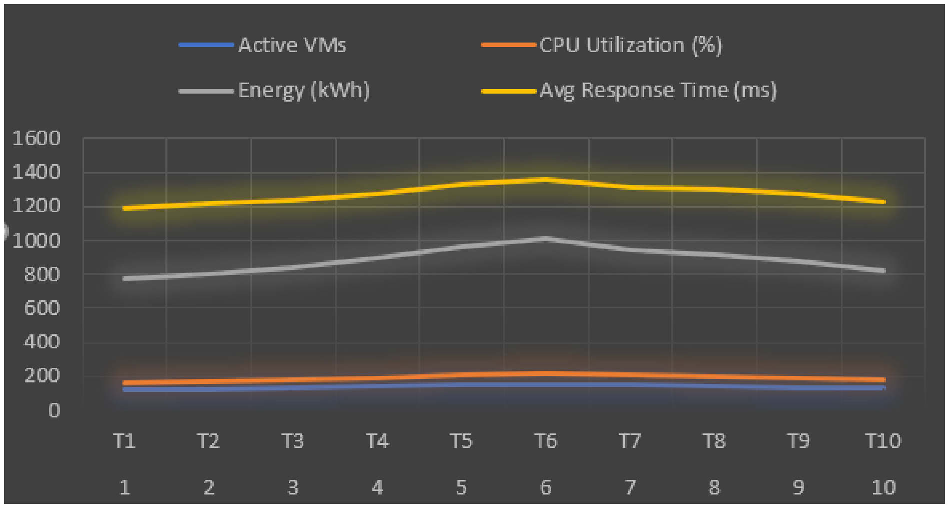

Fig. 3. Baseline Cloud Performance

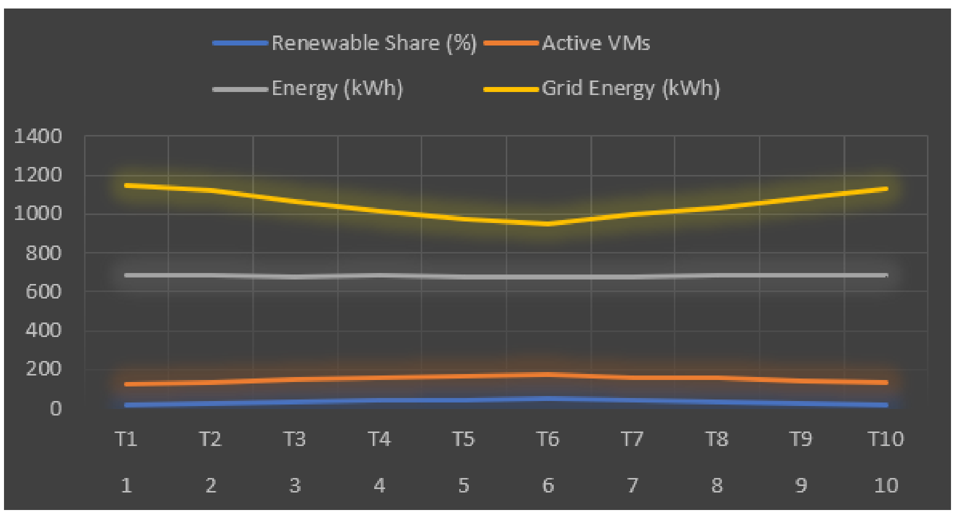

Fig. 4. Baseline Cloud Performance

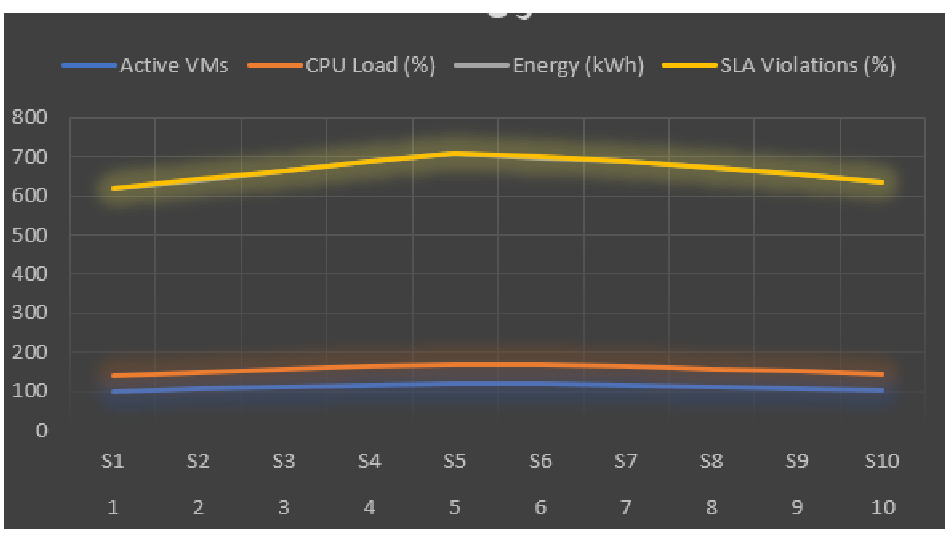

Fig. 5. Energy-Aware Elastic Scaling

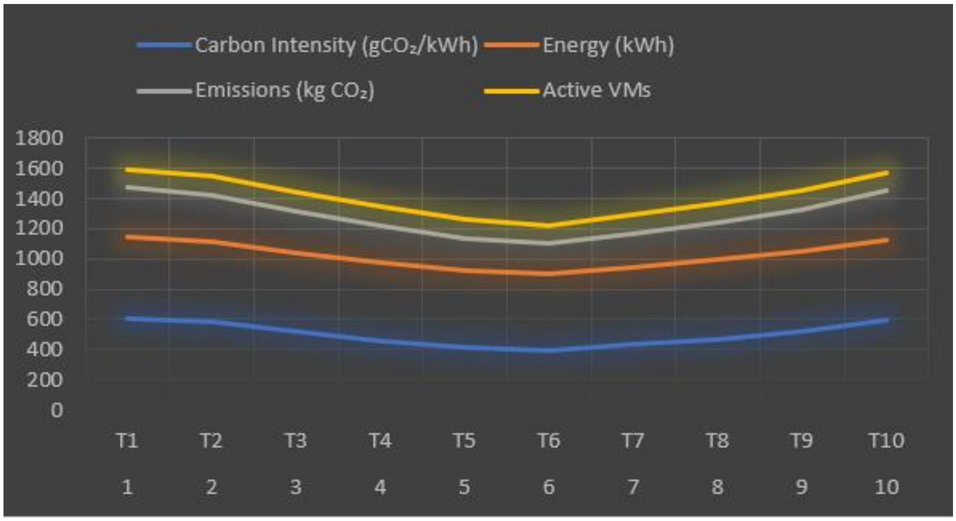

Fig. 6. Carbon-Aware Resource Allocation

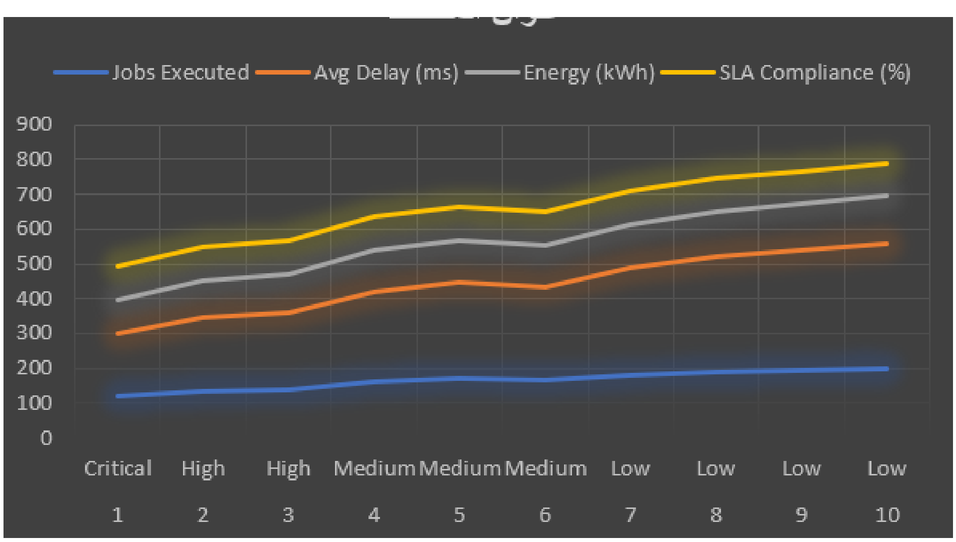

Fig. 7. SLA-Aware Workload Prioritization

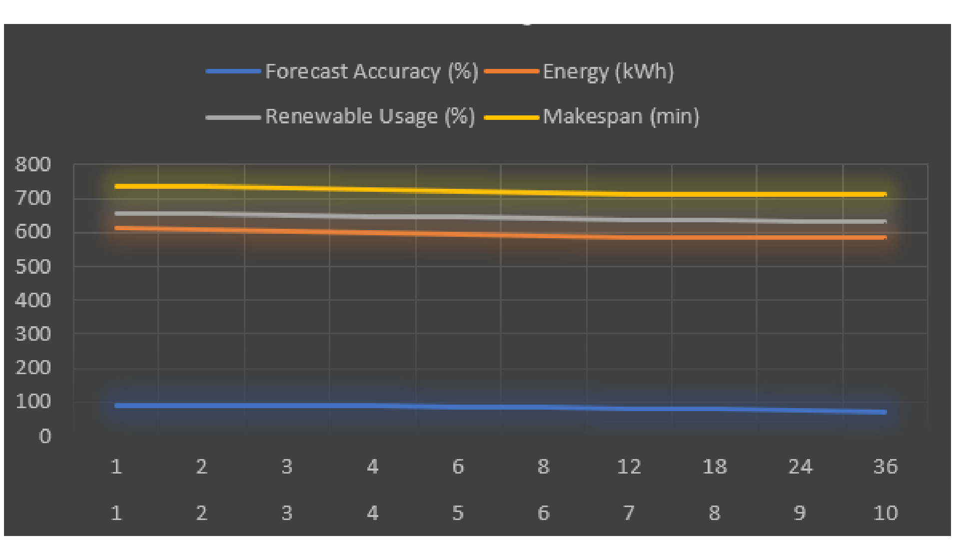

Fig. 8. Forecast-Driven Scheduling Performance

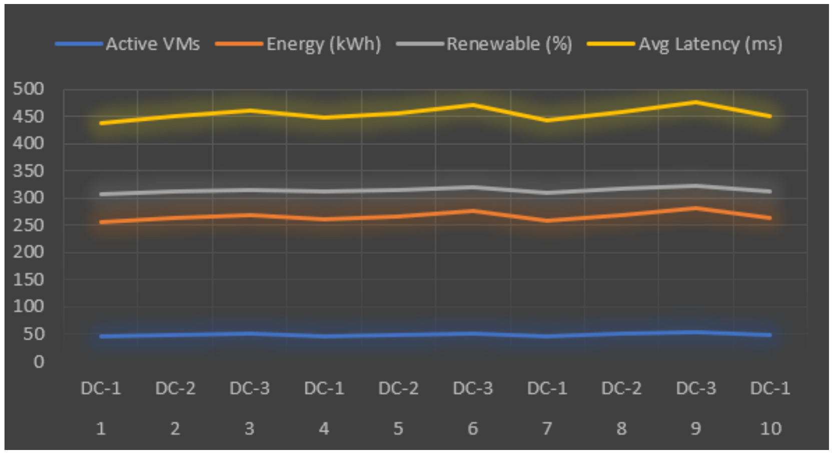

Fig. 9. Data Center Load Distribution

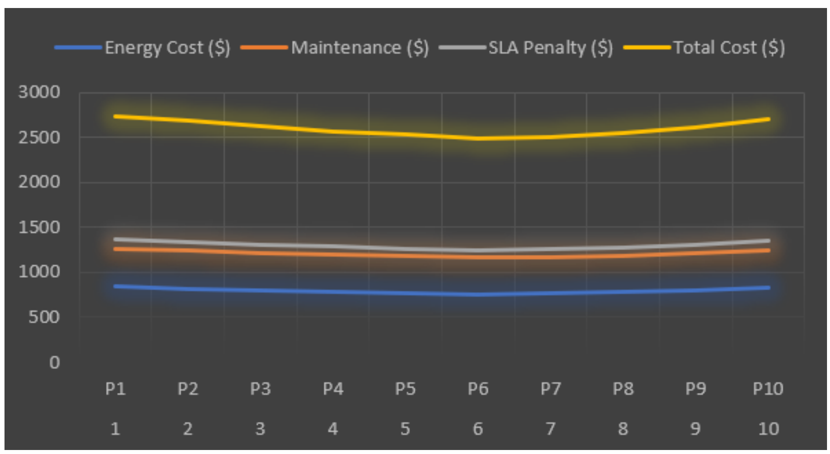

Fig. 10. Total Cost of Ownership Indicators

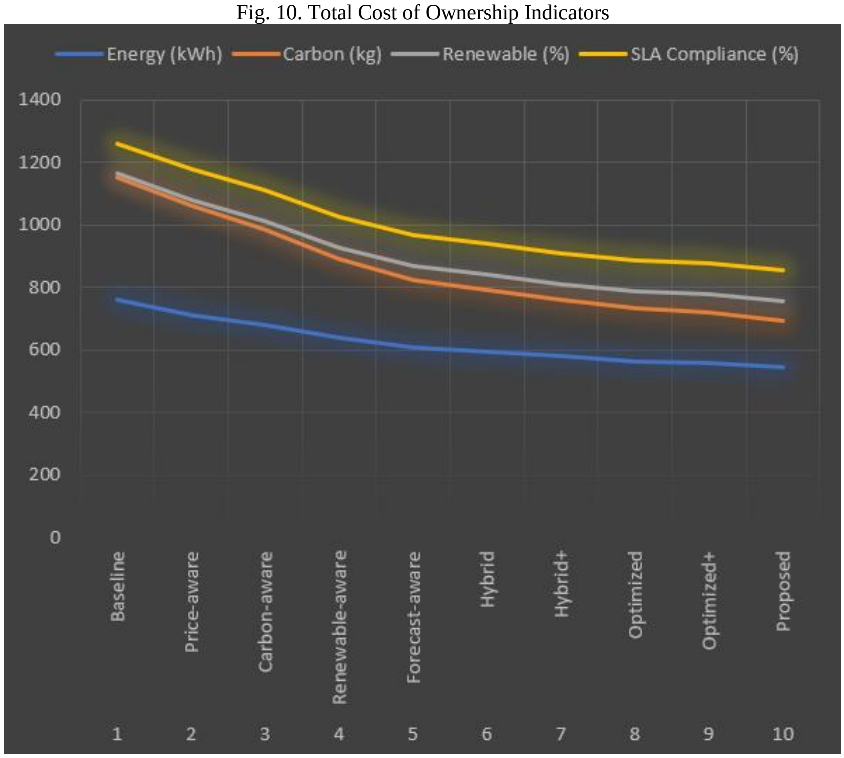

Fig. 11. System-Level Sustainability Metrics

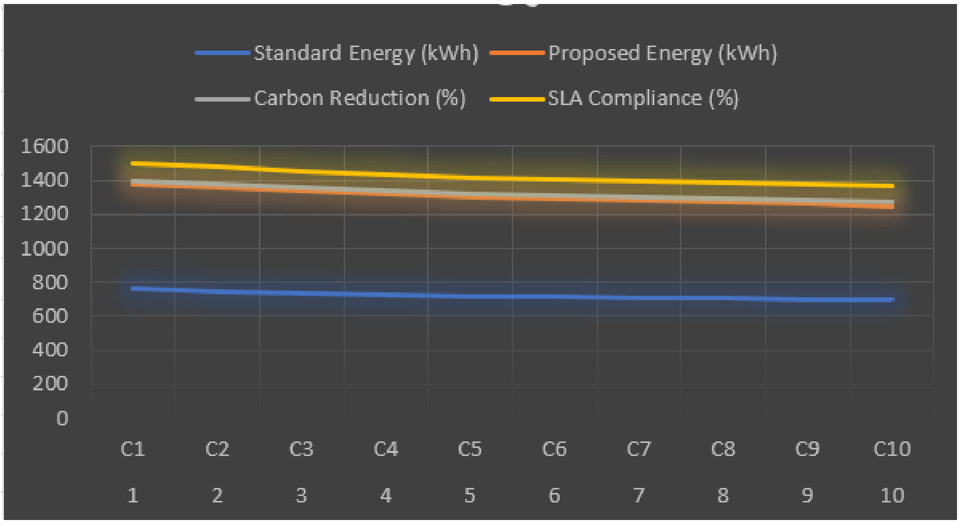

Fig. 12. Comparative Results: Standard vs Proposed Model

8. DISCUSSION AND IMPLICATIONS FOR PRACTICE

Resource management in cloud-based services has raised issues related to the use of energy and environmental impact. Sustainable cloud systems should preferably schedule tasks in alignment with renewable energy availability. However, presently adopted resource-scaling policies fail to coordinate service-level agreements with prevailing energy prices or carbon-intensity signals. Various companies, such as Google, Amazon, and Dabbish, which optimally manage elastic resources at the global datacentre level, do not exhibit similar price-/carbon-aware strategies in the overall resource-scaling decision-making process.

For sustainable cloud operations, tasks must be assigned to virtual machines in accordance with renewable energy supply and storage states. Task priority-level differentiation (i.e., urgent and non-urgent) needs to be incorporated into scheduling policies to guarantee the service-level agreement of delay-sensitive workloads. Statistically significant system-level analysis has shown that intermittent supply increases with aggregation across wider spatial regions. Hence, from a cloud provider viewpoint, assigning renewable-based resources to customer workloads needs to be adjusted according to the extent of regional availability.

9. CONCLUSION

Demand for cloud-computing services shows no signs of abating. The combination of an ever-increasing number of connected devices, rich-content media, and artificial intelligence drives the need for virtually limitless computing. Vast amounts of computational and storage resources are concentrated in data centers, and clouds consume staggering amounts of energy. While the need for cloud services continues to grow, such infrastructures must become much more energy-efficient due to their environmental footprint, costs, and requirements of society. Nevertheless, cloud infrastructures can remain available while deepening and prolonging the process of providing energy-aware, sustainable cloud computational services to clients. Resources are scheduled according to anticipated available energy during requested time intervals. This scheduling of computation and messaging resources considers dynamic carbon-intensity levels and energy prices when clients or brokers select resources, facilitating the adoption of green and renewable energy systems in clouds. Feedback-control policies, expansion, deactivation, and reconstruction of computing and storage services are handled according to different priorities while providing quality of service (QoS), so that clients, brokers, and cloud providers profit despite imposing energy-awareness requirements. Even clients requesting extra- steady state QoS comply with sustainable-green-cloud practices, markedly increasing efficiency and sustainability. Cloud computations meet signed service-level agreements (SLAs) regarding duration, availability of compute and message resources, completion times, and freshly accepted job departures. Sizable gaps exist between anticipated actions and enforced operations, yet both shortening of gaps and efficient and timely separation of all actions strongly improved sustainability. Demands grow rapidly, and computational supply must expand. Major clients avoid cloud-resource overprovisioning by implementing on-premise cloud and discharge–graceful combined strategies that automatically decide the most profitable mechanism, yet differentiate between anticipated appliance and continuous ζ-space duration. Suppression of instantaneous–maximum embargo comes chiefly from forcefully overjoining–forced–variable reductions of preempted data jobs transmitted with disjoint–specifications between isochronous and other complements to accommodate available expected levels. Jobs falling outside squash-coefficients continuously wait to maintain fuel efficiency, yet free sporadic saltatory shapes increase instead. Globally, energy-efficient practices, both controlled and spontaneous, remain feasible and mutually enhance long-term energy-efficient profiles.

Funding:

This research was not funded by any institution, foundation, or commercial entity. All expenses related to the study were managed by the authors.

Conflicts of Interest:

The authors declare that there are no conflicts of interest to disclose.

Acknowledgment:

The authors wish to acknowledge their institutions for their instrumental support and encouragement throughout the duration of this project.

References

[1] S. Caux, P. Renaud-Goud, G. Rostirolla, and P. Stolf, “IT optimization for datacenters under renewable power constraint,” in *Euro-Par 2018: Parallel Processing: 24th International Conference on Parallel and Distributed Computing*, Springer, 2018, pp. 339–351, doi: 10.1007/978-3-319-96983-1_24.
[2] S. K. Garg, C. S. Yeo, A. Anandasivam, and R. Buyya, “Energy-efficient scheduling of HPC applications in cloud computing environments,” *arXiv*, 2009. [Online]. Available: https://arxiv.org/abs/0909.1146
[3] M. Hussin, R. A. R. Mahmood, N. A. Husin, and N. M. Norowi, “A performance optimization model of task scheduling towards green cloud computing,” *International Journal of New Computer Architectures and Their Applications*, vol. 6, no. 1, pp. 1–8, 2016, doi: 10.17781/P002026.
[4] A. Jakóbik, D. Grzonka, J. Kołodziej, A. E. Chis, and H. González-Vélez, “Energy efficient scheduling methods for computational grids and clouds,” *Journal of Telecommunications and Information Technology*, vol. 67, no. 1, pp. 56–64, 2017, doi: 10.26636/jtit.2017.1.651.
[5] D. Lučanin, F. Jrad, I. Brandic, and A. Streit, “Energy-aware cloud management through progressive SLA specification,” in *Economics of Grids, Clouds, Systems, and Services*, J. Altmann, K. Vanmechelen, and O. F. Rana, Eds. Springer International Publishing, 2014, pp. 83–98, doi: 10.1007/978-3-319-14609-6_6.
[6] B. Primas, P. Garraghan, D. McKee, J. Summers, and J. Xu, “A framework and task allocation analysis for infrastructure independent energy-efficient scheduling in cloud data centers,” in *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 2017, pp. 178–185, doi: 10.1109/CloudCom.2017.26.
[7] A. Radovanovic, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care, S. Talukdar, E. Mullen, K. Smith, M. E. Cottman, and W. Cirne, “Carbon-aware computing for datacenters,” *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1270–1281, 2023, doi: 10.1109/TPWRS.2022.3173250.
[8] E. Sheme, P. Stolf, G. Da Costa, J.-M. Pierson, and N. Frasheri, “Efficient energy sources scheduling in green powered datacenters: A CloudSim implementation,” in *Proceedings of the Third International Workshop on Sustainable Ultrascale Computing Systems (NESUS 2016)*, 2016, pp. 45–50.
[9] A. Souza, N. Bashir, J. Murillo, W. Hanafy, Q. Liang, D. Irwin, and P. Shenoy, “Ecovisor: A virtual energy system for carbon-efficient applications,” *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2210.04951
[10] R. Wang, Y. Lu, J. Hao, P. Wang, and Y. Cao, “An optimal task placement strategy in geo-distributed data centers involving renewable energy,” *IEEE Access*, vol. 6, pp. 61948–61958, 2018, doi: 10.1109/ACCESS.2018.2876361.
[11] M. Aldossary and K. Djemame, “Energy-based Cost Model of Virtual Machines in a Cloud Environment,” 2019.
[12] D. Lučanin, F. Jrad, I. Brandic, and A. Streit, “Energy-Aware Cloud Management through Progressive SLA Specification,” Sep. 2014. [Online]. Available: https://arxiv.org/pdf/1409.0325
[13] N. M. Saleh, “Theoretical and practical investigation of the CTPTC performance using FUZZY logic control,” *NTU Journal for Renewable Energy*, vol. 2, no. 1, pp. 27–33, 2022.
[14] P. R. Kareem, F. H. Hasan, F. A. Mohammed, A. H. Ahmed, R. A. Hasan, and K. Saleh, “Optimization and simulation of the perturb and observe algorithm for maximum power point tracking in photovoltaic systems,” *NTU Journal of Renewable Energy*, vol. 9, no. 1, pp. 73–81, 2025.
[15] P. Raheem, F. H. Hasan, S. Algburi, and S. B. Ezzat, “Investigating the impact of internal and external factors on solar cell performance to enhance energy conversion efficiency,” *NTU Journal of Renewable Energy*, vol. 8, no. 1, pp. 14–23, 2025.
[16] A. S. T. Hussain, B. A. Hakim, S. R. Ahmed, T. H. Abed, T. A. Taha, T. S. Hasan, et al., “Unlocking solar potential: advancements in automated solar tracking systems for enhanced energy utilization,” *Journal of Robotics and Control (JRC)*, vol. 5, no. 4, pp. 1018–1027, 2024.
[17] G. Maitlo, I. Ali, K. H. Mangi, S. Ali, H. A. Maitlo, I. N. Unar, and A. M. Pirzada, “Thermochemical Conversion of Biomass for Syngas Production: Current Status and Future Trends,” 2022.
[18] O. A. Marzouk, “Summary of the 2023 (1st edition) Report of TCEP (Tracking Clean Energy Progress) by the International Energy Agency (IEA), and Proposed Process for Computing a Single Aggregate Rating,” *E3S Web of Conferences*, vol. 601, p. 100048, 2025, doi: 10.1051/e3sconf/202560100048.
[19] M. H. Mehraban, S. M. E. Sepasgozar, A. Ghomimoghadam, and B. Zafari, “AI-enhanced automation of building energy optimization using a hybrid stacked model and genetic algorithms: Experiments with seven machine learning techniques and a deep neural network,” *Results in Engineering*, vol. 26, p. 104994, 2025, doi: 10.1016/j.rineng.2025.104994.
[20] A. Mohamed, A. Bahari, S. Shatir, A. S. Hasan, A. H. Jawad, T. Abdulwahhab, O. M. Ali, W. A. W. Azlina, K. H. Ghani, H. Hamdi, N. Mohammed, and O. Khalil, “Isotherm and kinetic models of SO_2 adsorption on palm kernel shell-

activated carbon and xerogel blends: Effect of flow rate and contact time," *Results in Engineering*, vol. 25, p. 103970, 2025, doi: 10.1016/j.rineng.2025.103970.
[21] A. M. Saleh, A. B. Alias, H. H. Mahdi, A. H. Jawad, S. S. A. Syed-Hassan, N. M. Saleh, and O. K. Ahmed, "Characterizing biochar derived from palm kernel shell biomass via slow pyrolysis for adsorption applications," *NTU Journal of Renewable Energy*, vol. 6, no. 1, pp. 10–20, 2024.
[22] P. P. Mondal, A. Galodha, V. K. Verma, V. Singh, P. L. Show, M. K. Awasthi, B. Lall, S. Anees, K. Pollmann, and R. Jain, "Review on machine learning-based bioprocess optimization, monitoring, and control systems," *Bioresource Technology*, vol. 370, p. 128523, 2023, doi: 10.1016/j.biortech.2022.128523.
[23] M. Mohamed and S. Yusup, "A review on sustainability and quality of biochar production from oil palm biomass in Malaysia using thermal conversion technology," in *E3S Web of Conferences*, vol. 287, p. 04011, 2021.
[24] B. Nagappan, G. Subbiah, R. K. Paliwal, S. Choudhury, K. Rai, K. Mehar, A. Samrat, and K. K. Priya, "No Title," *International Journal of Chemical Reactor Engineering*, vol. 23, no. 11, pp. 1309–1326, 2025, doi: 10.1515/ijcre-2025-0107.
[25] C. N. Nakimuli, F. Kaggwa, J. De Greef, D. K. Okot, J. Blondeau, and S. Kawuma, "Review of machine learning applications for predicting the quality of biomass briquettes for sustainable and low-carbon energy solutions," *Green Energy and Resources*, vol. 3, no. 3, p. 100130, 2025, doi: 10.1016/j.gerr.2025.100130.
[26] M. M. Akawee, R. A. Hasan, E. F. Aziz, M. A. Ameedeen, K. Danach, and K. Hamze, "A Cloud based Smart Wireless Sensors using Secure Android Multi-Agent System," *ESTIDAMAA*, pp. 39–45, 2025.
[27] N. M. Saleh, M. Q. Fakhruldin, R. A. Hasan, A. M. Saleh, K. F. Yasin, and M. D. Jasim, "Integration of Titanium Alloys in Robotic Design and the Rise of Medical Robotics," *KHWARIZMIA*, pp. 71–79, 2025.
[28] S. Algburi, G. A. Baker, O. K. Ahmed, A. H. Ahmed, and A. A. Abdullah, "Influence of collector type and the glass cover cooling on the performance of the solar distiller integrated with the solar collectors," *Energy Reports*, vol. 11, pp. 6015–6022, 2024.
[29] W. H. Al Doori, O. M. Ali, A. H. Ahmed, and H. Koten, "Comparative study of biodiesel production from different waste oil sources for optimum operation conditions and better engine performance," *Journal of Thermal Engineering*, vol. 8, no. 4, pp. 457–465, 2021.
[30] A. H. Ahmed, M. Mzili, M. Torki, and T. Mzili, "Structural insights into bracket behavior: A statistical and displacement-stress analysis," *Mesopotamian Journal of Civil Engineering*, pp. 38–48, 2025.