International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(REVIEW ARTICLE)

Check for updates

Cloud data center performance optimization through machine learning-based workload forecasting and energy efficiency

Aravind Nuthalapati *

Microsoft, USA.

International Journal of Science and Research Archive, 2024, 13(02), 2353–2361

Publication history: Received on 28 October 2024; revised on 04 December 2024; accepted on 07 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2435

Abstract

The accelerating adoption of cloud models has increased the amount of complexity in cloud data centers with particular emphasis on the energy management load efficiency on resources in relation to the workload tris. This paper introduces a new fully-converged architectural framework enhanced by machine learning features that addresses several common issues: workload forecasting, adaptive scheduling and energy optimization for better performance at hyper scale cloud data centers. Using a Gated Recurrent Unit (GRU), the method described in the paper learns and remembers the complicated sequence of nonlinear workloads, enabling it to allocate resources appropriately in advance. Schedule optimization is achieved via a gradient boosting method in which resource managers are proactively chosen based on the expected workload in order to enhance scheduling and reduce task waiting times. Furthermore, virtual machine clustering models based on energy consumption patterns are incorporated into the design to enhance the framework's efficiency in energy usage optimization by controlling the number of virtual machines and their migration. The test case application based on realistic data center traces verified better energy effectiveness and resource utilization levels with service level agreement compliance for this entire integrated method. The study further underscores the opportunity for machine learning models to pinpoint and even combine distinct operational stresses prevalent in cloud data center environments.

Keywords: Cloud Data Centers; Machine Learning; Workload Forecasting; Energy Efficiency; Dynamic Scheduling

1. Introduction

Today, cloud computing is the backbone of modern digital infrastructure, supporting various applications [1], from web services and big data processing to AI workloads [2]. As the demand for services in the domain of cloud grows, it is a requirement for the data centers to rapidly and efficiently scale their physical resources while managing energy consumption and the quality of services [3]. However, it is an extremely challenging balancing act; it is becoming increasingly difficult to operate a data center sustainably and efficiently given the variations in workloads, complex scheduling demands, and energy-intensive operations [4]- [5].

Programs developing Artificial Intelligence (AI) turn out to be promising solutions to these dilemmas, data centers thereby can foresee workload demand, smartly transform the deployments in real-time, and use the energy in the best way [6] - [8]. One of the main themes in the recent works is the role of machine learning in the improvement of various aspects of data center operations such as workload forecasting, scheduling optimization, and energy management [9] - [11]. That is, accurate workload forecasting enables data centers to proactively allocate resources, which takes care of latency issues and improves responsiveness [12] - [14]. In such cases, Gated Recurrent Unit (GRU) networks, with their strength in capturing complex patterns and dependencies, have proven to be so in predicting the fluctuating workloads

∗ Corresponding author: Aravind Nuthalapati.

[15-17]. By determining upcoming and missing requests but sending this type of information data centers can utilize it to manipulate the resources dynamically thus overcoming the problem of over-allocation [18].

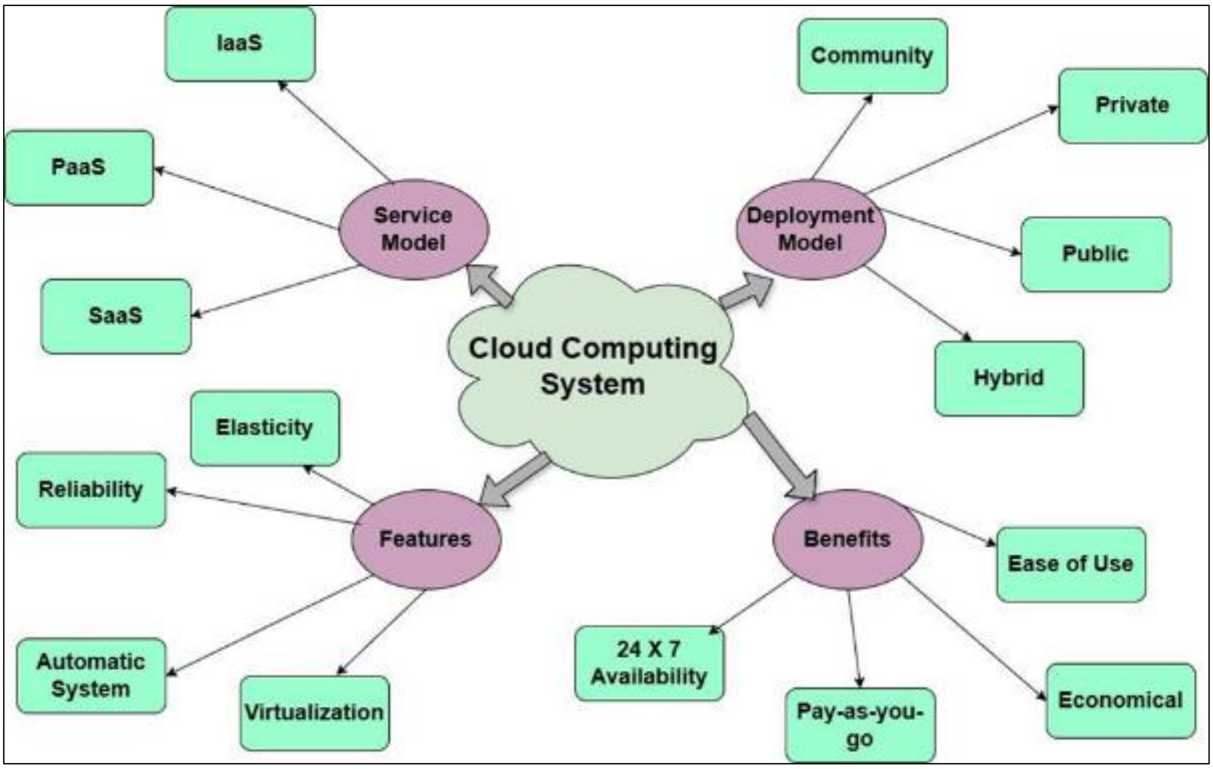

Figure 1 Overview of Cloud Computing Systems

The cloud computing system illustrated in Figure 1 gives a clearer picture of the main components, functionalities, and advantages of a cloud computing system. The system is broadly classified into three main dimensions: Service Models, Deployment Models, and Features. Service models cover Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) that provide flexible infrastructure, application development platforms, and internet based software solutions, respectively. The deployment models such as Private, Public, Hybrid, and Community Clouds provide a varying degree of control, cost effectiveness, and accessibility to match the organisational needs [19].

Cloud computing systems overview in Figure 1 shows the essentials, traits, and conceivable upsides of the cloud computing environment. The structure consists of the three fundamental aspects: Service Models, Deployment Models, and Features. Service models are IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service), which offer the standalone infrastructural model, platform for application development, and the sole web-based program respectively. Private, Public, Hybrid, and Community Cloud solutions are available to organizations that need different degrees of control, cost management, and accessibility from one side to another. [19].

Figure 1 also elaborates on the features of cloud computing, such as Elasticity which is dynamically varying the infrastructure according to the load, and Reliability which is guaranteeing uninterrupted use. Virtualization maximizes the number of users who can share the same resource, while automation eliminates manual labor for the control of resources. Cloud computing has its main advantages, such as ease of use, cost efficiency due to the pay-as-you-go models, and 24/7 access cloud computing has become a vital component of IT operations. This illustration does a great job, capturing the diverse nature of cloud infrastructures and their significance for different enterprises and institutions [20] - [22].

Scheduling optimization is another vital sector in which machine learning can bring change. Traditional scheduling techniques face bot- oul challenges involving rapid and unexpected changes in workload patterns, which are typical in modern advanced data centers [23]-[25]. In a nutshell, Fernández-Cerero et al. [26] proposed a model that uses gradient boosting regression to enhance the scheduling process. Fewer delays and a higher volume of work are the outputs of the methodology which can forecast scheduling durations and dynamically pick a resource management strategy according to workload conditions. Adaptive scheduling thus becomes a vehicle for more efficient resource allocation in data centers and helps shift their operation costs lower, thus improving service quality. Nevertheless, it is energy efficiency that is the biggest difficulty of all.

Data centers are huge energy consumers thanks to their very processes leading not only to high operational costs but also to the negligence of the environment through the excessive carbon emissions. To address this, the researchers have focused on methods to optimize the placement and migration of virtual machines (VMs) to try to achieve high energy savings without violating the service guarantee. The value of clustering techniques, such as TSSAP and Kmeans, which are used to classify the VMs according to their energy usage patterns, was shown by Awade et al. [27]. Having identified underutilized or overworked VMs, data centers can now better schedule the with and migrate VMs or consolidate resources leading to reductions in energy and eco-friendly operations [28].

This article brings together these technologies into one unified machine learning model that solves the problems of workload forecasting, adaptive scheduling, and energy-efficient resource management. The aim of this model combining GRU for workload prediction, gradient boosting for adaptive scheduling, and clustering for energy optimization is to make a smarter and more sustainable approach to data center management. The model can predict workloads and thus data centers can prepare for and adjust resources which leads to lower energy, faster response times, and better overall efficiency.

This article is going to describe each of these modules in detail with an explanation of the methods for workload forecasting, scheduling, and energy management, and analysis of the results of experiments which were conducted to test the model's performance. In the proposed machine learning-based data center management framework, a new set of approaches to cloud infrastructure maintenance will facilitate sustainable practices and the ease of scaling in accordance with growing demands, driving us into the future of cloud computing technologies.

2. Related Works

Machine learning has emerged to predict workloads in cloud data centers, specifically aimed at improving resource allocation and energy-efficient management. Although traditional statistical approaches have gained wide usage, they predominantly exhibit shortcomings in portraying the complex and non-linear dynamics of discrete workloads characteristic of cloud environments. Therefore, due to these restrictions, recent investigations have focused on deep learning architectures that particularly shine when modeling complicated data patterns. In the predictive scheduling techniques review, Kashyap and Singh [29]reviews express the effectiveness of deep learning in adapting accurately towards unexpected spikes in workload. On this note, a hybrid model is proposed by Fang et al. [30] which combines linear regression with wavelet neural networks for enhanced accuracy in short-term predictions, thus benefiting resource efficiency as well as energy conservation issues markedly.

Scheduling optimization is another significant area of remarkable potential for ML, particularly in the context of fluctuating workload demands. Meanwhile, efficient scheduling is key to optimizing resource utilization while maintaining service quality. A multi-agent reinforcement learning framework was developed by Zhang et al. [31] to enhance workload scheduling among geographically dispersed data centers, subsequently attaining improved GPU utilization along with lowered operational costs. In task scheduling within cloud environments, deep reinforcement learning was applied, extending the work of Song et al. [32]. Although resources are better utilized, energy consumption has substantially diminished as compared to previous studies. All these works mark the scope for ML-based scheduling under changing workload conditions wherein data centers can remain efficient at their best without degradation in performance.

Energy management in data centers is crucial because of the energy overheads when executing large-scale data processing tasks. Machine learning models have increasingly improved the prediction and management of energy consumption. Ilager et al. [33] constructed a gradient boosting model to predict temperatures in data centers and, thus, facilitate better energy management by reducing peak temperatures. The approach applied for thermal prediction saved not only energy but also reduced rather significantly the overheating risks in high-density server environments. Likewise, Perdigão [34] used federated learning to replicate energy consumption behaviors within smart grids, thereby offering an uninterrupted method for anomaly detection and classification regarding energy flow that would ultimately boost considerable energy efficiency. All these works demonstrate ML's ability to enhance energy management and rather support sustainable practices in data centers.

Integrated forecasting, scheduling, and energy management for workloads have been introduced as a comprehensive solution for the optimization of cloud data centers thus it has been gaining popularity. Qi et al. [35] has come up with a hybrid evolutionary learning framework named SHIELD that is designed to take into account such factors as carbon emissions, wastewater, and energy consumption during the operation of a data center. Resource utilization coupled with environmental impact reduction was proven by this framework. Besides, Netflix's predictive auto scaling engine, Scryer, is a perfect example of the use of predictive auto scaling in the real world. Drawing from historical data as well as recent usage trends, Scryer dynamically adjusts resources in advance to prove the case for a unified ML-driven approach to workload prediction, resource management, and energy efficiency.

Though ML-based models demonstrate a great deal of prospective benefit, challenges still exist with the successful implementation of these models in cloud-based data centers. Effectively obtaining high-quality, real-time data for the training of accurate ML models is one of the main challenges. In addition, the interpretability of ML models is one of the major issues that have to be taken into consideration as this is the means of building trust with data center operators, whose modification of the models requires understanding the model's decisions. Adaptability is yet another issue as ML models must be prepared for the increasing size and complexity of contemporary data centers [36]. Thus, the next phase of research should seek to overcome these problems by creating interpretable, scalable, and data-efficient ML models that can effortlessly incorporate existing cloud infrastructures, thus providing a greener and more efficient data center operations.

3. Comparative Analysis of Techniques

The content of this section expresses the detail about how accurately machine learning techniques can be applied in forecasting workloads, scheduling optimization, and carrying out energy management in the cloud data centers. We outline the methods' pros and cons, and thus we show how different models solve certain issues in cloud data center management and their capability to availability.

3.1. Workload Forecasting Techniques

3.1.1. Overview of Forecasting Techniques

Volumetric flow rate forecasting is the name for resource allocation in cloud data centers. It is through a number of machine learning models that workload patterns can be predicted in a successful manner, each tailored for complex, fluctuating workloads.

Gated Recurrent Units (GRU): According to [37], GRU’s, because of their ability to capture non-linear patterns and handle long-term dependencies, provide a reliable accuracy in workload forecasting with the cost of computation reduced in comparison to LSTM models.

Long Short-Term Memory (LSTM): LSTMs are the leaders in time-series forecasting and are experts in dealing with long-term dependencies, therefore, they are widely used in such applications. On the negative side, real-time applications pose a challenge due to the relatively higher computational demands of LSTMs, which GRUs often tackle better.

Linear Models: Linear models, on the other hand, such as traditional models Linear Regression (LR) and Ridge Regression (RR), investigated by [37], perform excellently in linearly recognized trends but can hardly cope with the dynamically moving workloads. Hybrid techniques that utilize linear models together with the neural networks approach are sometimes used for a balance between simplicity and adaptability.

3.1.2. Performance Comparison

Accuracy: Models based on GRU usually guarantee high precision in the predictors thanks to the ability to model non-linear workload fluctuations. Thus, they still offer reliability in cloud data centers that have unpredictable workloads.

Computational Efficiency: GRUs impose less computation than LSTMs real-time large-scale forecasting, making them an excellent candidate for this type of forecasting. On the other hand, linear models are lightweight but rigid, so they can’t deal with non-linear patterns, thus reducing the diversity of complex cloud environments.

Scalability: GRUs and LSTMs not only show better scalability for large datasets but also simpler linear models can’t make it hard to learn large datasets.

3.2. Scheduling Optimization Techniques

3.2.1. Scheduling Techniques in Cloud Data Centers

Efficient scheduling helps to realize the full efficiency of the utilized resources while reducing the waiting time. The rise of cloud computing has seen the advent of scheduling models based on machine learning to automatically replace traditional rule-based systems in dynamic cloud environments.

Gradient Boosting Models: Fernández-Cerero et al. [38] implemented a resource management system that uses gradient boosting to predict workload demands and select resource managers accordingly. The improved scheduling approach showed higher efficiency in scheduling and lower task delay than the conventional static methods.

Reinforcement Learning (RL): RL models are suitable for dynamic scheduling as they discover optimal policies over time. RL-based models learn by self-correcting according to real-time feedback, thus they have higher throughput and work well in variable environments [39].

Deep Reinforcement Learning (DRL): RL has been combined with deep learning to produce models such as Deep Q-Networks that can make more complicated scheduling decisions. According to Yi et al. [40], DRL can deal with complex task scheduling in hyper-scaling data centers by adapting to rapidly changing workload demands.

3.2.2. Comparison of Scheduling Techniques

Adaptability: The RL-based models have the upper hand in adaptability owing to their unique design that enables them to adjust dynamically in real-time to the live feedback, in contrast to the traditional boosting models based on the pre-trained data.

Throughput and Delay Reduction: DRL models do this by maximizing the completion of tasks and minimizing the time in delays by choosing different schedules that are out of the box, though this may be accomplished only with the use of high-end computational devices.

Efficiency in Stable vs. Dynamic Environments: Models using gradient boosting are suitable in predictable environments where there are workload patterns and most of the time, one could expect that to happen, but in situations with high variability, there are limitations when comparing with RL models.

3.3. Energy Optimization Techniques

3.3.1. Approaches to Energy Optimization

Machine learning for energy optimization in data centers that rely on high energy is mainly about regulating resource consumption and optimizing the allocation and consolidation of virtual machines.

Clustering Algorithms: Bozhgani et al. [41] examined clustering algorithms such as TSSAP and Kmeans which group the VMs by energy usage patterns in order to help strategically VM consolidation, reducing the energy consumption in total.

Predictive Energy Models: Predictive models like Ridge Regression and ElasticNet have been proven to work in estimating energy use on the VM level such that historical data is used for proactive resource adjustments.

Federated Learning: Federated learning provides an energy prediction decentralized approach which is very beneficial for multi-tenant data centers with privacy issues. The method does not require centralized data collection, hence it saves bandwidth while preserving data privacy.

3.3.2. Comparison of Energy Optimization Techniques

Energy Savings: Clustering algorithms cut down energy consumption significantly through the detection of possible VM consolidations while predictive models allow proactive adjustments to resource distribution based on the prognosis of energy demand.

Scalability and Data Efficiency: Predictive models are scalable and can be utilized over extensive data centers. Federated learning is specifically designed for multi-tenant environments, as it does not necessitate data centralization which can minimize both data privacy concerns and bandwidth usage.

Privacy and Bandwidth Efficiency: Federated learning's distributed framework guarantees data privacy by keeping the location of training data hidden. This is an advantageous situation in areas without centralized data storage.

The machine learning methodologies are possibly out of the concept proof that sometimes the real world can be proved by existing some limitations, while their success in cloud data centers in load forecasting, scheduling, and energy efficiency. This subsection brings out the main obstacles and trends that need to be solved in order to reach to the next level of cloud data centers optimization and more effectiveness of ML applications.

This comparative analysis shows that a single machine learning model is not sufficient to solve all the problems of cloud data centers comprehensively. GRU models are capable of accurate and efficient workload forecasting for non-linear patterns, hence they are suitable for resource management. In scheduling, reinforcement learning and its variants are the best, owing to their flexibility, while deep reinforcement learning gives sophisticated decision-making, at the expense of higher computational cost. Concerning energy consumption, clustering algorithms and federated learning appear to be the two most effective approaches, thus, either energy reduction is achieved through distributed systems or privacy and efficiency are still retained for each system.

The integration of these machine learning approaches certainly is a potential path towards an optimal data center. The combination of GRU-based forecasting and reinforcement learning for scheduling and clustering for energy optimization in a hybrid framework could utilize the strengths of each approach. Such an integrated model would increase adaptability, lower operational costs, and facilitate the sustainability of data center operations. Nevertheless, in the meantime, some problems such as real-time scalability, data privacy, and model interpretability still lie ahead. These issues in future research high on the list of priorities should be tackled to ensure the further growth of machine learning applications in cloud data centers.

4. Challenges and Open Issues

Machine learning (ML) has successfully proven to be the solution for workload forecasting, scheduling, and energy efficiency in a cloud data center, but there are still several challenges that can restrict the full potential of these techniques. Here, the discussion deals with the major barriers and matters that are still to be solved in the journey of the comprehensively cloud-based environment that will also render ML applications as effective and scalable.

4.1. Scalability Issues

One of the most significant issues with the use of machine learning models in cloud data centers is scalability. Since cloud environments are growing, models will need to cope with increasing data amounts and intricacies. Deep learning models, for instance, are extremely computationally heavy and have high memory resource requirements which can be a limiting factor in the scalability of hyper-scale data centers. Long Short-Term Memory (LSTM) and Deep Reinforcement Learning (DRL) are very accurate but very expensive to compute which could lead to either long processing times or high energy consumption. While GRU-type models that are lightweight have been brought forth as replacements, striking the balance between scalability and predictive accuracy is still a raging issue. It is imperative that large cloud infrastructures be able to operate models that scale without sacrificing performance.

4.2. Data Quality and Availability

The quality of data and its availability very much determine the efficacy of ML models. For example, in the case of real-time optimization, cloud data centers need high-quality, continuous data streams in order to accurately predict and make decisions. Unfortunately, data centers are often hampered by the incompleteness of the data, adding latency during data acquisition, and the inconsistent data quality due to the complexity of the infrastructure. On top of that, some centers may keep different types of logs or measure data in different ways, which negatively affect the performance of distributed systems. Data that is either low-quality or old can result in inaccurate forecasts, wrong scheduling decisions, and less energy-efficient solutions. Building sound data collection protocols and enhancing data handling in real time should be the main actions to keep this problem at bay.

4.3. Model Interpretability

Interpretability is a vital issue in ML models, especially in workload forecasting, scheduling, and energy management. Although many data center operators are afraid to trust in the so-called "black-box" models such as DRL or deep neural networks, which might be very accurate and not transparent. The fact that they are unable to make out how the decisions are made is a serious problem for the operators who are due to this fact unwilling to trust the models, particularly, when they are deployed in mission-critical environments. Furthermore, for troubleshooting and improving model performance, interpretability is the most important factor. Some methods like Explainable AI (XAI) are introduced to solve this gap, but further research is needed to develop interpretable models for cloud data centers.

4.4. Energy and Computational Costs of ML Models

On the contrary, the use of energy-efficient ML models could be energy-consuming due to the fact that those requiring heavyweight computation like LSTMs and DRL are utilized. The models that are live need considerable computing power, which ultimately alights the center's energy consumption along with the still possible savings from energy efficient practices like optimized scheduling and VM consolidation. Besides that, things may not be different in edge data centers or micro data centers, which are low-capacity resource-constrained environments. Therefore, there is a demand for the construction of lightweight ML models that give precise forecasts with low computational needs, hence, maintaining energy savings and smooth operation in data centers.

4.5. 4.5 Data Privacy and Security Concerns

In the multi-tenant cloud environments, the data's privacy and security are the priority when the clients' data are used to train ML models. Federated learning has been introduced as a possible solution that allows models to learn from decentralized data without centralizing. Nevertheless, the implementation of federated learning will introduce other challenges like ensuring the data consistency through nodes and the security threats handling, like data poisoning attacks. Furthermore, the bandwidth limitations are tied up with the federated learning, as the model updates have to be sent back and forth regularly, which may cause the network performance to drop. Further studies are required to design strong, privacy-protected frameworks that assure client data safety while still allowing for efficient model training and deployment.

4.6. Adaptability to Real-Time Requirements

Cloud data center optimization refers to the many dynamic workloads and resource demands, and as such, it must be done in real-time. Nonetheless, many ML models still encounter problems with real-time responsiveness. It is possible that the models that were trained using historical data will not be able to adjust to sudden changes in workload patterns resulting in the inefficient use of resources and scheduling. Even though reinforcement learning models are adaptable, they often need a lot of time to be trained, which is not practical in an environment that is always changing. Models that combine real-time adaptability with predictive accuracy are still undeveloped, and further innovation is needed to find ways to make models responsive to real-time changes while still being as accurate as possible.

5. Conclusion

In the paper, the disruptive influence of machine learning on the optimization of cloud data centers has been explored in depth - with the whole showcased through three critical areas: workload forecasting, scheduling optimization, and energy management. ML methods including Gated Recurrent Units (GRU) for forecasting, gradient boosting and reinforcement learning for scheduling, as well as clustering algorithms for energy optimization have shown their capability in solving problems associated with dynamic and resource-rich cloud environments. These methods move together to make a more efficient operation, increasingly complex data centers, that consume less energy and preserve service quality.

Even with the technology getting better, issues like scalability, model interpretability, data privacy, and real-time adaptability still exist. However, closing this gap is a source of opportunity for future research. The hybrid machine learning model that combines forecasting, scheduling, and energy management is the way to go towards complete optimization. Besides that, the techniques in Explainable AI (XAI), federated learning, and lightweight model design will be crucial in developing solutions for data centers that are scalable, secure, and sustainable. Cloud computing demand will surge, thus, the employment of cloud data centers powered by intelligent machine learning algorithms will be indispensable for both operational efficiency and environmental sustainability.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Luo, F., Zhao, J., Dong, Z. Y., Chen, Y., Xu, Y., Zhang, X., & Wong, K. P. (2015). Cloud-based information infrastructure for next-generation power grid: Conception, architecture, and applications. *IEEE Transactions on Smart Grid*, *7*(4), 1896-1912.

[2] Shahane, V. (2021). Harnessing Serverless Computing for Efficient and Scalable Big Data Analytics Workloads. *Journal of Artificial Intelligence Research*, *1*(1), 40-65.

[3] Khan, A. A., & Zakarya, M. (2021). Energy, performance and cost efficient cloud data centres: A survey. *Computer Science Review*, *40*, 100390.

[4] Dittakavi, R. S. (2023). Achieving the Delicate Balance: Resource Optimization and Cost Efficiency in Kubernetes. *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal*, *12*(2), 125-131.

[5] Uddin, M., Talha, M., Rahman, A. A., Shah, A., Ahmed, J., & Memon, J. (2012). Green Information Technology (IT) framework for energy efficient data centers using virtualization. *International Journal of Physical Sciences*, *7*(13), 2052-2065.

[6] Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022). Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *Journal of Network and Computer Applications*, *204*, 103405.

[7] AR, B., RS, V. K., & SS, K. (2023). LCD-capsule network for the detection and classification of lung cancer on computed tomography images. *Multimedia Tools and Applications*, *82*(24), 37573-37592.

[8] Sathupadi, K. (2023). Ai-driven energy optimization in sdn-based cloud computing for balancing cost, energy efficiency, and network performance. *International Journal of Applied Machine Learning and Computational Intelligence*, *13*(7), 11-37.

[9] Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022). Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *Journal of Network and Computer Applications*, *204*, 103405.

[10] Subeh, P., & Bushara, A. R. (2024). Cloud data centers and networks: Applications and optimization techniques*.International Journal of Science and Research Archive, 2024, 13(02),* 218–226.

[11] MirhoseiniNejad, S., Badawy, G., & Down, D. G. (2021). Holistic thermal-aware workload management and infrastructure control for heterogeneous data centers using machine learning. *Future Generation Computer Systems*, *118*, 208-218.

[12] Kumar, J., & Singh, A. K. (2020). Adaptive Learning based Prediction Framework for Cloud Datacenter Networks' Workload Anticipation. *Journal of Information Science & Engineering*, *36*(5).

[13] Nuthalapati, A. (2022). Optimizing Lending Risk Analysis & Management with Machine Learning, Big Data, and Cloud Computing. Remittances Review, 7(2), 172-184

[14] Zhao, M., Wang, X., & Mo, J. (2023). Workload and energy management of geo-distributed datacenters considering demand response programs. *Sustainable Energy Technologies and Assessments*, *55*, 102851.

[15] Chandra, N., Ahuja, L., Khatri, S. K., & Monga, H. (2021). Utilizing gated recurrent units to retain long term dependencies with recurrent neural network in text classification. *J. Inf. Syst. Telecommun*, *2*, 89.

[16] babu Nuthalapati, S., & Nuthalapati, A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning. International Journal of Science and Research Archive, 12(2), 408-422.

[17] Rodriguez, S. (2023). Gated Recurrent Units-Enhancements and Applications: Studying Enhancements to Gated Recurrent Unit (GRU) Architectures and Their Applications in Sequential Modeling Tasks. *Advances in Deep Learning Techniques*, *3*(1), 16-30.

[18] Huang, H., Wang, Y., Cai, Y., & Wang, H. (2024). A novel approach for energy consumption management in cloud centers based on adaptive fuzzy neural systems. *Cluster Computing*, *27*(10), 14515-14538.

[19] Goyal, S. (2013). Software as a service, platform as a service, infrastructure as a service− a review. *International journal of Computer Science & Network Solutions*, *1*(3), 53-67.

[20] Alharthi, S., Alshamsi, A., Alseiari, A., & Alwarafy, A. (2024). Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions. *Sensors*, *24*(17), 5551.

[21] Bushara, A. R., RS, V. K., & Kumar, S. S. (2024). The Implications of Varying Batch-Size in the Classification of Patch-Based Lung Nodules Using Convolutional Neural Network Architecture on Computed Tomography Images. *Journal of Biomedical Photonics & Engineering*, *10*(1), 39-47.

[22] Zhao, S., Miao, J., Zhao, J., & Naghshbandi, N. (2023). A comprehensive and systematic review of the banking systems based on pay-as-you-go payment fashion and cloud computing in the pandemic era. *Information Systems and e-Business Management*, 1-29.

[23] Cunha, B., Madureira, A., Fonseca, B., & Matos, J. (2021). Intelligent scheduling with reinforcement learning. *Applied Sciences*, *11*(8), 3710.

[24] Jishamol, T. R., & Bushara, A. R (2016). Enhancement of Uplink Achievable Rate and Power Allocation in LTEAdvanced Network System. International Journal of Science Technology and Engineering (IJSTE). 211-217.

[25] Velayutham, A. (2019). Ai-driven storage optimization for sustainable cloud data centers: Reducing energy consumption through predictive analytics, dynamic storage scaling, and proactive resource allocation. *Sage Science Review of Applied Machine Learning*, *2*(2), 57-71.

[26] Fernández-Cerero, D., Troyano, J. A., Jakóbik, A., & Fernández-Montes, A. (2022). Machine learning regression to boost scheduling performance in hyper-scale cloud-computing data centres. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3191-3203.

[27] Awad, M., Leivadeas, A., & Awad, A. (2023). Multi-resource predictive workload consolidation approach in virtualized environments. *Computer Networks*, *237*, 110088.

[28] Yadav, M., & Mishra, A. (2024, May). Efficient Workload Distribution for Sustainable Server Utilization in Cloud Data Centers. In *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)* (pp. 1-6). IEEE.

[29] Kashyap, S., & Singh, A. (2023). Prediction-based scheduling techniques for cloud data center's workload: a systematic review. *Cluster Computing*, *26*(5), 3209-3235.

[30] Fang, L., & He, B. (2023). A deep learning framework using multi-feature fusion recurrent neural networks for energy consumption forecasting. *Applied Energy*, *348*, 121563.

[31] Zhang, S., Xu, M., Lim, W. Y. B., & Niyato, D. (2023, December). Sustainable AIGC workload scheduling of geo-Distributed data centers: A multi-agent reinforcement learning approach. In *GLOBECOM 2023-2023 IEEE Global Communications Conference* (pp. 3500-3505). IEEE.

[32] Song, P., Chi, C., Ji, K., Liu, Z., Zhang, F., Zhang, S., ... & Wan, X. (2021, July). A deep reinforcement learning-based task scheduling algorithm for energy efficiency in data centers. In *2021 International Conference on Computer Communications and Networks (ICCCN)* (pp. 1-9). IEEE.

[33] Ilager, S., Ramamohanarao, K., & Buyya, R. (2020). Thermal prediction for efficient energy management of clouds using machine learning. *IEEE Transactions on Parallel and Distributed Systems*, *32*(5), 1044-1056.

[34] Meira, J., Matos, G., Perdigão, A., Cação, J., Resende, C., Moreira, W., ... & Aguiar, R. L. (2023). Industrial Internet of things over 5G: A practical implementation. *Sensors*, *23*(11), 5199.

[35] Qi, S., Milojicic, D., Bash, C., & Pasricha, S. (2023, October). SHIELD: Sustainable hybrid evolutionary learning framework for carbon, wastewater, and energy-aware data center management. In *Proceedings of the 14th International Green and Sustainable Computing Conference* (pp. 56-62).

[36] Muhammed Kunju, A. K., Baskar, S., Zafar, S., & AR, B. (2024). A transformer based real-time photo captioning framework for visually impaired people with visual attention. *Multimedia Tools and Applications*, 1-20.

[37] Khan, T., Tian, W., Ilager, S., & Buyya, R. (2022). Workload forecasting and energy state estimation in cloud data centres: ML-centric approach. *Future Generation Computer Systems*, *128*, 320-332.

[38] Juiz, C., Bermejo, B., Fernández-Montes, A., & Fernández-Cerero, D. (2024). Towards a General Metric for Energy Efficiency in Cloud Computing Data Centres: A Proposal for Extending of the ISO/IEC 30134-4. In *CLOSER* (pp. 239-247).

[39] Nuthalapati, A. (2024). Architecting data lake-houses in the cloud: Best practices and future directions. Int. J. Sci. Res. Arch, 12(2), 1902-1909.

[40] Yi, D., Zhou, X., Wen, Y., & Tan, R. (2020). Efficient compute-intensive job allocation in data centers via deep reinforcement learning. *IEEE Transactions on Parallel and Distributed Systems*, *31*(6), 1474-1485.

[41] Bozhgani, M. S. A., & Moghadam, M. H. Y. (2024, May). Analyzing Power Usage in Datacenter Workloads Using Random Forest and LSTM Models. In *2024 8th International Conference on Smart Cities, Internet of Things and Applications (SCIoT)* (pp. 43-48). IEEE.