Enhancing Cloud Energy Efficiency through Predictive Machine Learning for Inter- and Intra-Data Center VM Consolidation

Santina Capalbo[1,†], Eugenio Cesario[2,*,†], Paolo Lindia[1,†], Federica Lobello[1,†] and Andrea Vinci[3,†]

[1]*DIMES Department, University of Calabria, 87036, Rende, Italy*
[2]*DICES Department, University of Calabria, 87036, Rende, Italy*
[3]*Institute for High-Performance Computing and Networking (ICAR), CNR—National Research Council of Italy, 87036, Rende, Italy*

Abstract

The rapid expansion of Cloud Computing and large-scale data centers has resulted in a substantial increase in energy consumption, primarily due to hardware operation and cooling requirements. This rise in power usage has significantly elevated operational costs, making energy efficiency a critical concern for data center management. Virtual Machine (VM) consolidation is a well-established strategy to address these challenges by reducing the number of active physical servers while ensuring compliance with Service Level Agreements (SLAs). However, the effectiveness of consolidation heavily depends on the accurate prediction of VM resource demands. This paper proposes a consolidation approach—encompassing both intra- and inter-data center strategies—for energy-aware VM allocation across physical servers. The system leverages predictive machine learning models to forecast future computational needs of individual VMs. By anticipating these demands, the framework dynamically allocates VMs across the servers of the considered data centers to optimize server utilization and minimize energy consumption, without compromising performance or SLA compliance. Preliminary experimental results demonstrate that the proposed approach significantly reduces overall power consumption, particularly when guided by machine learning-driven workload forecasting.

Keywords

Cloud energy optimization, machine learning for energy efficiency, green and sustainable computing, Virtual Machines (VMs) consolidation

1. Introduction

In recent years, Cloud Computing has spread rapidly, becoming a strategic resource for companies and organizations. This paradigm avoids the costs associated with installing and maintaining physical infrastructure, allowing companies to focus their resources on innovation and process optimization. An increasing number of companies and scientific communities are transferring their data, software, and services to the cloud, taking advantage of a scalable infrastructure that avoids the costs and complexity associated with directly managing hardware and software. This phenomenon has been facilitated by the adoption of the pay-as-you-go model, which, together with the spread of high-performance data centers and fast networks, has contributed to the rapid spread of on-demand computing in industry and research [1] [2] [3]. Nevertheless, this expansion has been accompanied by a considerable increase in operating costs, mainly due to the enormous energy consumption required by data centers, which need large amounts of power not only to run the hardware but also to keep them cool. With the evolution of cloud technologies, the scientific community has intensified its efforts to develop more energy-efficient

2nd Workshop on Green-Aware Artificial Intelligence, 28th European Conference on Artificial Intelligence (ECAI 2025), October 25–30, 2025, Bologna, Italy
*Corresponding author.
†These authors contributed equally.
santina.capalbo@dimes.unical.it (S. Capalbo); eugenio.cesario@unical.it (E. Cesario); paolo.lindia@dimes.unical.it (P. Lindia); federica.lobello@unical.it (F. Lobello); andrea.vinci@icar.cnr.it (A. Vinci)
0009-0006-3541-1741 (S. Capalbo); 0000-0002-4987-0459 (E. Cesario); 0000-0002-7550-1331 (P. Lindia); 0009-0004-0590-3679 (F. Lobello); 0000-0002-1011-1885 (A. Vinci)

solutions, promoting more sustainable management models. Among the most promising strategies are task scheduling optimization, better resource allocation, and the intelligent use of Virtual Machines consolidation techniques [4] [5]. In this context, Power Usage Effectiveness (PUE) [6] has become a fundamental metric for measuring the energy efficiency of data centers. Introduced by The Green Grid consortium (www.thegreengrid.org/), the PUE is defined as the ratio between the total energy consumption of a data center and the energy used exclusively for IT equipment (servers, storage, switches, etc.). A PUE value of 1 is ideal, indicating that all energy is used only to power IT devices, without waste. In practice, modern data centers aim to maintain values below 1.5, thanks to more effective cooling solutions, virtualization technologies, and advanced monitoring systems.

In this study, we propose an approach to perform energy-efficient allocations of virtual machines in a Cloud environment through predictive machine learning models. In particular, virtual machine allocation across servers is performed through a two-level consolidation strategy involving both inter-data center and intra-data center optimizations. The approach involves continuous monitoring, predictive modeling, and periodic reallocation to reduce the number of active servers while ensuring performance and compliance with SLAs. In particular, the approach aims to reduce the total energy consumption by working at two levels. First, the proposed strategy prioritizes VM allocations to physical servers hosted by data centers with better PUE values by giving lower priorities to data centers showing lower efficiencies. Second, it leverages predictive machine learning models to anticipate future CPU demands and optimize VM consolidation inside each data center, with the aim of further reducing the whole energy consumption.

The rest of the paper is organized as follows. Section 2 provides an overview of existing approaches for energy-efficient virtual machine allocation found in the literature. Section 3 presents our approach that leverages machine learning models to drive virtual machine allocation considering the energy efficiency of data centers. Section 4 reports an analysis of the experimental results obtained by comparing different scenarios and presents an assessment of the effectiveness and efficiency of the proposed strategic approach. Finally, Section 5 concludes the work by summarizing the results obtained and proposing directions for the continuation of research activities.

2. Related work

Several works have addressed the issue of reducing energy consumption in data centers. An overview of the most relevant proposed approaches is provided below, with a focus on optimizing task scheduling, resource allocation, and virtual machine consolidation, both intra- and inter-data center. In [3] the authors predict resource utilization – CPU, memory, and network usage – in cloud data centers by proposing a combined CNN-LLSTM model. Data are first processed using Vector Autoregression (VAR), followed by CNN (Convolutional Neural Networks) which extracts complex features from the components of VM resource utilization. These features are then fed into the LSTM (Long Short Term Memory) to generate the final predictions. A different strategy for predicting workload and resource utilization – CPU and RAM – time series is presented by [7], which proposes a combined model of Bi-directional LSTM and Grid LSTM (BG-LSTM) to capture bidirectional dependencies and extract and concatenate features from both the time and frequency domains. In [6] the authors predict data centers Power Usage Effectiveness (PUE) by comparing the performance of three deep learning models: Multilayer Perceptron (MLP), Resilient Backpropagation-based Deep Neural Network (DNN), and Attention-based Long Short-Term Memory (LSTM). To this end, they identify the input features that have a strong influence on the variation of the PUE through the Sobol Sensitivity Analysis, and define the relationship between them through the Hinton diagram. To reduce energy consumption by minimizing the number of active servers through consolidation and balanced usage of multidimensional resources – CPU, RAM, and bandwidth –, in [5] the authors propose a hybrid Virtual Machine Placement (VMP) algorithm that integrates an improved permutation-based genetic algorithm (IGA-POP) with a multidimensional resource-aware best-fit allocation strategy. In [8] the authors propose an approach for resource provisioning – CPU, memory, and storage space – in a cloud environment. They combine the Imperialist Competition Algorithm (ICA) with K-means to cluster workloads based on their Quality of Service (QOS) attributes and then use a decision tree algorithm to determine resource provisioning. The authors in [9] propose a technique for VM allocation through the Enhance-Modified Best Fit Decreasing (E-MBFD) algorithm, which first sorts VMs in decreasing order of their CPU utilization and then allocates them to physical machines after verifying that they have sufficient available resources. The resulting allocations are validated through an Artificial Neural Network (ANN). In order to perform dynamic consolidation of VMs, in [10] the authors propose two algorithms to detect overloaded and underloaded hosts, considering energy consumption and the number of migrations when dealing with underloaded hosts detection. To predict short-term CPU utilization, they employ the Gray-Markov model on accumulated host data. The authors in [11] propose the Online Multi Resource Feed-forward Neural Network (OM-FNN) model to simultaneously forecast multiple resource demands from running VMs, combined with the Tri-adaptive Differential Evolution (TaDE) algorithm to optimize its predictor. The system clusters tasks based on their predicted resource requirement to facilitate VM autoscaling and allocation to energy-efficient physical hosts. The Online VM Prediction-based Multi-objective Load Balancing (OP-MLB) framework, proposed in [12], follows a three-phase process. In the first phase, tasks are assigned to VMs according to their resource requirements. In the second phase, forecasts of resource utilization are made by an online predictor based on neural networks. Finally, in the third phase, VMs are allocated and migrated using a multi-objective algorithm which prevents underload and overload of physical machines. In [13] the authors propose a combination of an Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict workload patterns and an Advanced Ant Colony Optimization (AACO) technique to optimize resource allocation. By doing so, resources are dynamically reconfigured in response to real-time feedback. To minimize the consumption of "brown" energy and maximize the use of renewable one, [14] proposes a model for inter-data center VM migration. The approach involves determining VM migration requests, performing routing and spectrum allocation over the elastic optical network (EON) infrastructure, and allocating computing resources. Four heuristic algorithms are introduced to determine which VMs to migrate and to which data center. With the same goal, the authors in [15] propose an algorithm for inter-DC migration over the elastic optical network infrastructure. To reduce the consumption of brown energy, they introduce the Sliding-Window Lower Confidence Bound (SW-LCB) algorithm, based on the multi-armed bandit (MAB) formulation. Additionally, to enhance the cost efficiency of optical network devices and migration, they incorporate optical grooming techniques. In [16], the Green Energy Oriented Virtual-machine Migration Algorithm (GEOVMA) is proposed, aiming to minimize the combination of average response time, brown energy consumption, and carbon emissions. This goal is achieved by dynamically migrating VMs between data centers powered by renewable energy, employing the policy of Minimum Migration Time. The authors in [17] propose two multi-phase algorithms for load balancing and optimizing task scheduling across VMs distributed in multiple data centers. The max select load balancing (MSLB) algorithm does not consider communication delay and bandwidth, whereas the max select load balance with communication delay (MSLBCD) algorithm takes both into account.

3. Proposed approach

This section presents an architecture that leverages machine learning models to enable energy-efficient allocation and migration of virtual machines (VMs) across (inter-) and within (intra-) data centers. The proposed system, illustrated in Figure 1, consists of multiple geographically distributed data centers interconnected via an Elastic Optical Network (EON) infrastructure [14]. Each data center is characterized by a distinct energy efficiency level, quantified using the Power Usage Effectiveness (PUE) metric. PUE is defined as $PUE = \frac{\text{Total DC Power Consumption}}{\text{IT Facilities Power Consumption}}$, providing a standardized measure of energy efficiency (further details can be found in [6]). Notably, PUE values may vary over time due to factors such as seasonal changes or fluctuations in the availability of locally generated (e.g., renewable) energy.

Each data center comprises different key components, such as Physical and Virtual Machines, VM

Monitor, VM Consumption Modeler, and VM Migration Manager. The Physical and Virtual Machines are, respectively, the servers and the VMs that execute client tasks. VMs can be migrated between servers within the same data center or across external ones by the VM Migration Manager which, in managing server loads, prioritizes data centers with a better efficiency indicator and switches off inactive servers to save energy. The VM Monitor is a module that continuously records the CPU usage of VMs over time, supplying crucial data for modeling and analysis. The VM Consumption Modeler employs machine learning algorithms to build a predictive model for each VM to forecast its future CPU usage, by analyzing resource usage patterns. Different types of algorithms can be used, such as classification models, regression models, or neural networks. As stated above, the VM Migration Manager is responsible for periodic energy-efficient VMs consolidation by using the predictions from the VM Consumption Modeler.

Virtual machine allocations across servers is performed through a two-level consolidation strategy involving both inter-data center and intra-data center optimizations, as outlined below. First, each virtual machine vm is assigned to the most energy-efficient data center, determined in terms of PUE values, available to host vm. Once a data center is selected, an intra-data center consolidation is executed to allocate all VMs onto the minimal number of servers required to satisfy performance constraints. Idle servers are then switched off to reduce overall energy consumption. In particular, the intra-data center VM consolidation follows the methodology described in [4]. The process involves continuous monitoring and logging of resource utilization. At regular intervals, this data is analyzed to build predictive models of VM resource demands. These forecasts are then used to plan VM migrations that minimize the number of active servers. Unused servers are transitioned into low-power modes, thereby improving energy efficiency while maintaining compliance with Service Level Agreements (SLAs). To prevent SLA violations caused by unforeseen demand spikes, a safety margin is enforced by limiting server utilization at a threshold $\delta < 1.0$, leaving a buffer $1 - \delta$ for unexpected demands.

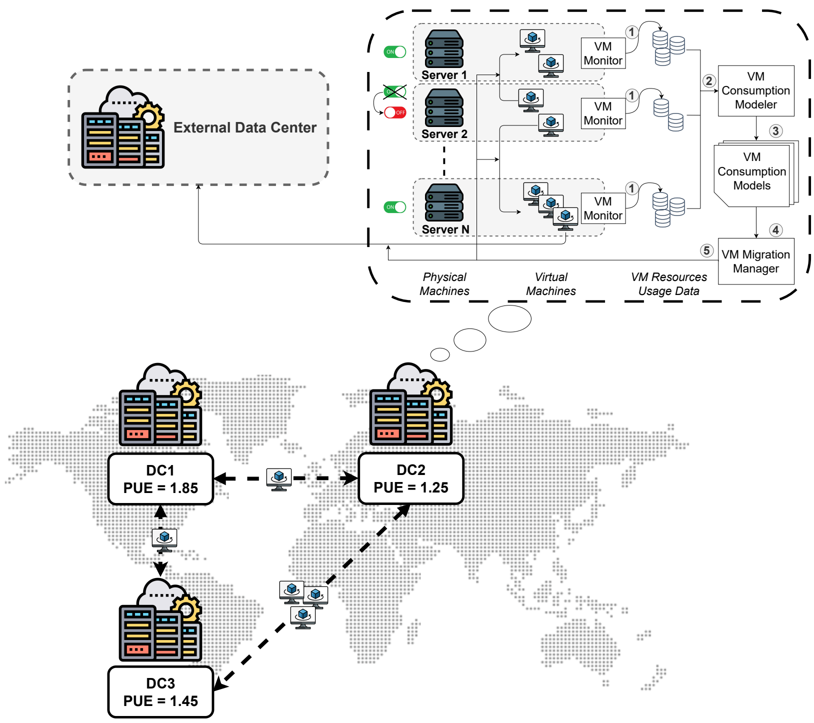

Figure 1: The energy-aware cloud architecture.

4. Experimental results

In this section, we present some results obtained from the proposed energy-aware approach tested on synthetic data, and we describe the ad-hoc data generator exploited to produce these data. The experiments are aimed at evaluating the potential energy savings derived from the migration of virtual machines among servers hosted by data centers with low-efficiency indicators to data centers with higher efficiencies. The objective is to evaluate whether the migration process guided by machine learning models can reduce the energy consumption of a Cloud system compared to a non-energy-aware scenario (*random approach*), while still meeting performance requirements and ensuring reliable service quality for users. The effectiveness of the proposed approach in terms of energy savings is demonstrated through experiments carried out in a simulated cloud environment. In the following, we present preliminary results achieved so far.

4.1. Synthetic data generation and experimental settings

An ad-hoc data generator was developed to simulate resource-usage traces for virtual machines, in order to analyze synthetic workloads. Generative usage patterns have been built through underlying probability distributions. In particular, four normal distributions f_i, for $i = 1, \ldots, 4$, have been defined. Each distribution has distinct means μ_i and standard deviations σ_i, which generate data samples from distinct, non-overlapping data ranges. Each virtual machine has a reference time period which is segmented into four temporal windows, with a unique distribution applied to each temporal window. Then, by a random assignment process, one distribution is associated to each virtual machine, generating data for each specific time window. In this way, four distinct usage patterns are created and subsequently assigned to each virtual machine. This allows to capture variability in CPU demand over time.

The experimental environment set for our tests is composed of three data centers, each one hosting 40 physical servers (a total of 120 servers, distributed on three data centers). Moreover, 600 virtual machines are running to deal with users' requests, distributed among the servers of the three data centers. The efficiency indicator of each data center is expressed in terms of Power Usage Effectiveness (PUE). PUE values have been taken from publicly available data provided by Google Data Centers [18]. Specifically, the *2024 PUE Yearly Report* [18] is divided by quarters (three months), with each data center having different PUE values for each quarter. In our experiments, Data Center 1 and 3 have an average better PUE than the Data Center 2. Resource usage was sampled every six hours over each day for an entire year, resulting in a dataset composed of 878,400 instances. The values of δ have been fixed at $\delta = 1$ for the Oracle scenario and at $\delta = 0.75$ for the ML-based one.

4.2. Energy consumption results

This subsection analyzes the energy performance for three different VM migration scenarios, exploited as benchmarks to highlight the effectiveness and drawbacks of various migration strategies in terms of energy efficiency. A detailed description of each scenario is provided below:

- *Oracle*. This scenario describes a perfect (ideal) case in which the VM Migration Manager queries an "oracle" to foresee the upcoming CPU requirements of each virtual machine. Having this information in advance allows the system to plan and execute optimal VM migrations ahead of time, to reduce energy usage. When allocating virtual machines to physical servers, priority is given to those located in data centers with higher energy efficiency, with the aim of further improving the optimization of total consumption. While this is not achievable in real-world settings, the Oracle scenario acts as a theoretical benchmark, illustrating the best (theoretically) achievable result if future demands could be predicted without errors.
- *ML-based*. In this scenario, virtual machine migrations are driven by predictions provided by the VM Migration Manager, based on the expected usage of the virtual machines. To estimate CPU demands, we adopted a machine learning model, specifically a *Multi-layer Perceptron* [19] [20], which allows us to accurately predict future loads. This predictive approach simulates a

more realistic environment than the Oracle-driven scenario, enabling us to assess how much predictive models can come to optimal energy performance even without perfect knowledge of the future. Furthermore, in order to further optimize overall energy consumption, the VM Migration Manager exploits these forecasts to allocate the virtual machines on the physical servers, by giving priority to physical machines hosted by the most efficient data centers.

- *Random.* This is a baseline scenario, in which virtual machines are randomly allocated among available servers, not taking into account the efficiency of data centers. This is an uninformed approach, in which relocations are carried out without considering either the expected demands of the virtual machines or the energy efficiency of the data centers, thus making both the assignment to physical servers and the choice of destination data center completely random. The Random scenario serves as a benchmark for evaluating the benefits of forecast-based strategies in terms of energy efficiency, highlighting how much informed decisions can improve the results compared to a purely stochastic approach.

The energy efficiency of the proposed approach is illustrated in Figure 2, which shows the total energy consumption of the three data centers during the entire test period, comparing the three different scenarios. The total energy consumption is computed as the cumulative energy needed by the system to perform computational and VM migration tasks. The figure shows that, as expected, the Oracle scenario is the best case in terms of energy savings. In fact, since Data Center 1 and Data Center 3 have a higher efficiency indicator than Data Center 2, the consolidation of virtual machines is performed by prioritizing their allocations to the most efficient data centers, leaving some servers in Data Center 2 turned off and, as a result, allowing for a reduced utilization of the least efficient data center. Similarly, for the ML-based strategy, it is evident that by applying the logic in which virtual machine allocation prioritizes servers located in more energy-efficient data centers, Data Center 2 (i.e., the least efficient) ends up consuming a lower amount of energy. This is because it contains powered-off servers, which contributes to a lower overall energy consumption. On the other hand, the Random scenario shows that the total energy consumption is higher than those computed in the other two scenarios. In fact, since the consolidation of virtual machines occurs randomly and uniformly within the servers and the choice of the destination data center is also random, all three data centers have approximately the same energy consumption.

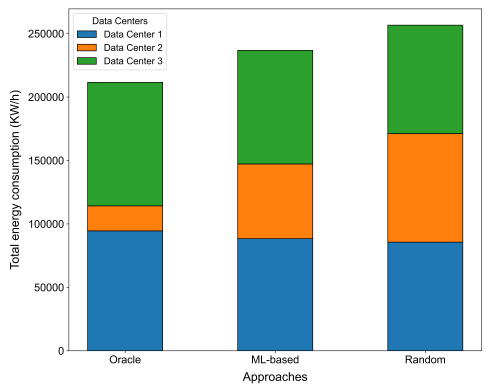

Figure 2: Total energy consumed by the different approaches, split in individual data center consumptions.

Figure 3 shows the total number of virtual machine migrations between servers, i.e., the number of times virtual machines are moved from a source server to a destination server located in different data centers, for the three different study scenarios during the entire simulation period considered. From the figure, we can observe that the number of migrations in the Oracle scenario is lower than those computed in the other two scenarios. In fact, in the Oracle case, virtual machines are primarily distributed across Data Center 1 and Data Center 3, which are the most energy-efficient. In contrast, the machine learning-based approach results in a higher number of migrations. This is mainly due to the lower server load threshold (δ = 0.75), which leads to a more conservative consolidation strategy. As a result, Data Center 2 is utilized more frequently than in the Oracle scenario, increasing the total number of migrations. Finally, in the Random scenario, the number of migrations is higher than in Oracle and lower than in ML-based. This could be due to the fact that, since not only the choice of the destination server is random, but also the data center to which it belongs, some virtual machines tend to migrate more often between servers in the same data center and less between servers in different data centers.

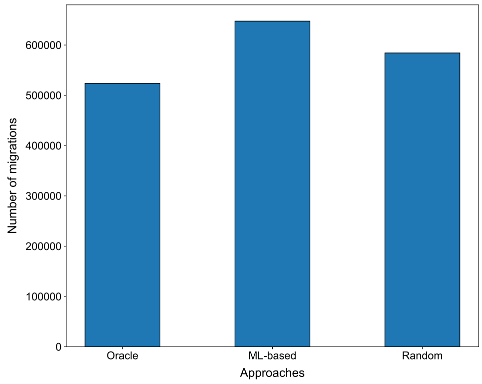

Figure 3: Number of migrations for each approach.

5. Conclusion

This work presented a predictive machine learning-based approach for enhancing cloud energy efficiency through VM consolidation, both within and across data centers. By leveraging predictive models to estimate future CPU usage, the proposed system enables efficient VM allocations that reduce the number of active physical servers and prioritize those hosted in energy-efficient data centers, i.e., the data centers characterized by lower PUE. Experimental results, based on synthetic workloads, demonstrate that the proposed strategy significantly lowers energy consumption compared to uninformed (random) allocation strategies. The incorporation of Power Usage Effectiveness (PUE) metrics enhanced the energy-aware allocation policy, contributing to a more sustainable cloud infrastructure. Future work will focus on extending the approach by considering a policy further optimizing the number of VM migrations, especially those that occur between different data-center. The energy-aware model will also take into account more complex and real-world network topologies. Finally, the approach will be extended by integrating additional resource dimensions, e.g., memory and GPUs.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

Acknowledgments

This work was supported by the "PNRR MUR project PE0000013-FAIR", the "ICSC National Centre for HPC, Big Data and Quantum Computing" (CN00000013) within the NextGenerationEU program, by European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: “SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics” - Prot. IR0000013 - n. 3264, 28/12/2021, and by the research project “INSIDER: INtelligent ServIce Deployment for advanced cloud-Edge integRation” granted by the Italian Ministry of University and Research (MUR) within the PRIN 2022 program and European Union - Next Generation EU (grant n. 2022WWSCRR, CUP H53D23003670006).

References

[1] S. Bharany, S. Sharma, O. I. Khalaf, G. M. Abdulsahib, A. S. Al Humaimeedy, T. H. Aldhyani, M. Maashi, H. Alkahtani, A systematic survey on energy-efficient techniques in sustainable cloud computing, Sustainability 14 (2022) 6256.

[2] A. Katal, S. Dahiya, T. Choudhury, Energy efficiency in cloud computing data centers: a survey on software technologies, Cluster Computing 26 (2023) 1845–1875.

[3] S. Ouhame, Y. Hadi, A. Ullah, An efficient forecasting approach for resource utilization in cloud data center using cnn-lstm model, Neural Computing and Applications 33 (2021) 10043–10055.

[4] E. Cesario, P. Lindia, F. Lobello, A. Vinci, S. Zarin, S. Capalbo, Improving cloud energy efficiency through machine learning models, in: 2025 33rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), 2025, pp. 247–251. doi:10.1109/PDP66500.2025.00041.

[5] A. S. Abohamama, E. Hamouda, A hybrid energy–aware virtual machine placement algorithm for cloud environments, Expert Systems with Applications 150 (2020) 113306.

[6] H.-A. Ounifi, A. Gherbi, N. Kara, Deep machine learning-based power usage effectiveness prediction for sustainable cloud infrastructures, Sustainable Energy Technologies and Assessments 52 (2022) 101967.

[7] J. Bi, S. Li, H. Yuan, M. Zhou, Integrated deep learning method for workload and resource prediction in cloud systems, Neurocomputing 424 (2021) 35–48.

[8] A. Shahidinejad, M. Ghobaei-Arani, M. Masdari, Resource provisioning using workload clustering in cloud computing environment: a hybrid approach, Cluster Computing 24 (2021) 319–342.

[9] Shalu, D. Singh, Artificial neural network-based virtual machine allocation in cloud computing, Journal of Discrete Mathematical Sciences and Cryptography 24 (2021) 1739–1750.

[10] S.-Y. Hsieh, C.-S. Liu, R. Buyya, A. Y. Zomaya, Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers, Journal of Parallel and Distributed Computing 139 (2020) 99–109.

[11] D. Saxena, A. K. Singh, A proactive autoscaling and energy-efficient vm allocation framework using online multi-resource neural network for cloud data center, Neurocomputing 426 (2021) 248–264.

[12] D. Saxena, A. K. Singh, R. Buyya, Op-mlb: an online vm prediction-based multi-objective load balancing framework for resource management at cloud data center, IEEE Transactions on Cloud Computing 10 (2021) 2804–2816.

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

[14] L. Zhang, T. Han, N. Ansari, Energy-aware virtual machine management in inter-datacenter networks over elastic optical infrastructure, IEEE Transactions on Green Communications and Networking 2 (2018) 305–315.

[15] F. S. Amri, Z. Huang, K. Liu, J. Pan, Energy-aware inter-data center vm migration over elastic optical networks, in: GLOBECOM 2023 - 2023 IEEE Global Communications Conference, 2023, pp. 5421–5426.

[16] S. Vatsal, S. B. Verma, Virtual machine migration based algorithmic approach for safeguarding environmental sustainability by renewable energy usage maximization in cloud data centres, International Journal of Information Technology (2023) 1–16.

[17] S. C. M. Sharma, A. K. Rath, B. R. Parida, Efficient load balancing techniques for multi-datacenter cloud milieu, International Journal of Information Technology 14 (2022) 979–989.

[18] Google, Google data center pue performance, 2025. URL: https://datacenters.google/efficiency/, last accessed 11 June 2025.

[19] N. AL-Rousan, N. A. Mat Isa, M. K. Mat Desa, H. AL-Najjar, Integration of logistic regression and multilayer perceptron for intelligent single and dual axis solar tracking systems, International Journal of Intelligent Systems 36 (2021) 5605–5669.

[20] E. Agirre-Basurko, G. Ibarra-Berastegi, I. Madariaga, Regression and multilayer perceptron-based

models to forecast hourly o3 and no2 levels in the bilbao area, Environmental Modelling & Software 21 (2006) 430–446.