Dynamic Underload Host Detection for Performance Enhancement in Cloud Environment

Deepak Kumar Singh Yadav(✉) and Bharati Sinha

Computer Engineering Department, NIT Kurukshetra, Thanesar, India
{deepak_32123201,bharatisinha}@nitkkr.ac.in

Abstract. Cloud computing provides on-demand availability of computing resources, data storage and computing power. Cloud service providers often have functions distributed over multiple locations, each of which is a data center.Cloud computing relies on sharing of resources to achieve coherence and typically uses a pay-as-you-go model. However, cloud computing faces some significant challenges in efficiently managing resources, optimizing performance, and reducing energy consumption. Among the mentioned challenges ensuring optimal energy consumption is the key concern. Further, improvisation in energy efficiency helps minimize carbon emissions and also enhances overall performance. One of the primary reason of energy misuse in computation is host underload i.e., the host is not operating on its optimum capacity. The challenge of host underload detection, can be efficiently managed with the help of linear regression method. Our proposed approach aims to simultaneously reduce consumption of energy, minimize virtual machine (VM) migration and uphold SLA (Service Level Agreement) compliance. Any reduction in the number of VM migrations, results in better resource utilization and also mitigates the impact on performance caused by frequent migrations. This approach seeks to strike a balance among energy efficiency and meeting SLA, without compromising quality of service provided.

Keywords: Cloud Computing · Host Overload · Cloud datacenter · QoS · SLA

1 Introduction

CC has transformed the operational landscape for businesses by offering convenient and immediate access to a diverse array of computational resources for example servers, storage, and applications. This innovation empowers businesses to store and process data remotely, eliminating the requirement for extensive physical infrastructure and dedicated IT personnel [1]. CC offers numerous benefits, including scalability, cost-effectiveness and flexibility. Moreover, the exponential rise of CC has also brought about several challenges, one of which is host underload.

Host underload denotes to a situation in cloud computing where a physical server or virtual machine (VM) is operating with a workload significantly lower than its

K. K. Patel et al. (Eds.): icSoftComp 2023, CCIS 2031, pp. 29–40, 2024.
https://doi.org/10.1007/978-3-031-53728-8_3

capacity. It means that the resources allocated to the host are not fully utilized, resulting in inefficient resource allocation and potential wastage as show in Fig. 1. Underload hosts may consume unnecessary power, leading to increased energy costs and decreased operational efficiency [1, 2]. Addressing host underload involves strategies such as load balancing, dynamic resource allocation, and workload consolidation to optimize resource utilization, improve performance, and achieve cost-efficiency in cloud computing environments.

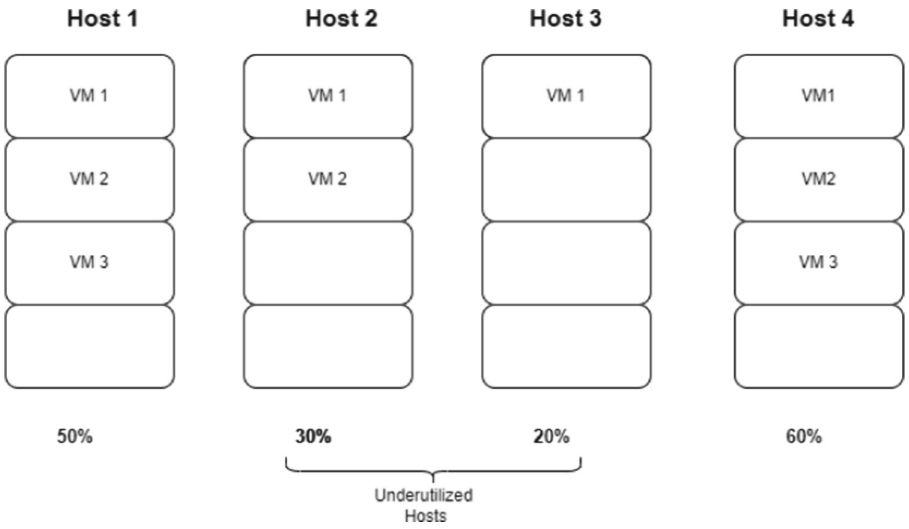

Fig. 1. Underutilized Host.

To solve this problem we proposed the algorithm uses linear regression technique for underload host detection in cloud computing. Underload host detection can significantly impact the SLA, Energy, and Performance. Linear regression is a widely utilized machine learning methodology i.e. commonly employed to forecast the value of a variable based on another variable [3]. Within this particular framework, the variable that is to be anticipated is commonly referred to as the dependent variable (y), whereas the variable employed for the purpose of prediction is denoted as the independent variable (x).

2 Literature Review

This section presents different strategies to overcome the underload host problem, reduce energy consumption and SLA violation and improve performance.

Youssef Saadi proposed an algorithm that focuses on energy efficiency for the consolidation of VM in cloud data centers. The aim is to minimize energy consumption while considering the utilization of hosts and ensuring the data center operates at optimal throughput. The authors compare their suggested strategy with onset algorithms, namely IQR and LR, and demonstrate its hopped-up [1]. The scheme successfully meets SLA requirements, as indicated by the simulation results.

Minarolli proposed a methodology for local resource allocation that involves modifying the CPU allocation assigned to virtual machines (VMs) in response to the current workload. Additionally, global resource allocation is achieved by VM migration, which aims to balance the distribution of resources among hosts that are either overloaded or underloaded. In order to forecast resource utilization, the authors utilize Gaussian processes as a machine learning methodology for time series prediction, taking into account long-term tendencies [2].

Abbas Horri, examines the trade-off that exists between energy usage and SLA Violation inside cloud systems. The objective of cloud service providers is to enhance their revenue by employing energy-efficient resource management strategies, which include the consolidation of virtual machines (VMs) and converting inactive servers into sleep modes. Nevertheless, inadequate consolidation may lead to the development of Sleep-Related Arousal Variants (SLAV). The author proposed the implementation of consolidation algorithms that aim to optimize energy usage while simultaneously minimizing Service Level Agreement Violations (SLAVs), in order to achieve a harmonious equilibrium between these two aims [3]. The simulation results indicate that the proposed methods effectively reduce the quantity of virtual machine migrations, SLAV (Service Level Agreement Violations), and total transferred data in comparison to existing strategies.

Nimisha introduced a novel algorithm, known as the Host Utilization Aware (HUA) Techniques, which aims to detect underloaded hosts. The algorithm presented in this study aims to estimate the upper limit of hosts that can be made available by taking into account the overall utilization of the data center. The primary emphasis of the Author's work is on the crucial stage of identifying underloaded hosts throughout the process of workload consolidation. The algorithm predicts the maximum number of hosts that may be made available by taking into account the overall utilization of the data center. The experimental results illustrate the effectiveness of the HUA Algorithm in accurately identifying hosts with low workload and thus freeing up a larger number of hosts. This leads to a reduction in energy consumption while still ensuring compliance with Service Level Agreement (SLA) requirements [4].

Weichao Ding revolves around the optimization of resource allocation and utilization in Cloud data centers. The overarching objective is to achieve a reduction in energy consumption while simultaneously upholding a high degree of compliance with service-level agreements (SLAs). The suggested framework by the author introduces a novel approach to dynamic virtual machine (VM) consolidation [5]. This approach is based on the prediction of resource use and the PPR (performance-to-power ratio) of heterogeneous hosts. The framework has four distinct stages, including host overload detection, VM selection for migration, host underload detection, and VM allocation with a modified power-aware best-fit reducing method. The proposed methodology effectively reconciles the trade-off between energy usage and performance. The suggested strategy has been validated for its effectiveness and scalability through experimental evaluations.

Bernardi Pranggono examines the matter of VM consolidation in cloud data centers and presents a novel approach for classifying host load inside an energy-performance VMC framework. The primary aim is to decrease energy usage while simultaneously guaranteeing the fulfillment of quality of service (QoS) criteria. The proposed approach entails the classification of hosts experiencing underload into three distinct states: underloaded, normal, and critical. This classification is achieved by the utilization of an underload detection algorithm. Additionally, the study presents the introduction of overload detection as well as the VM selection strategies. The overload detection policy, referred to as Mean (Mn), utilizes the mean to forecast the upper threshold [6]. On the other hand, the VM selection policy, known as Maximum Requested Bandwidth (MBW), relies on the maximum requested bandwidth.

Nirmal Kr. Biswas revolves around the resolution of two key issues in Smart Cities: the optimization of energy consumption as well as the mitigation of SLA violation. To tackle these challenges, Biswas proposes the utilization of Cloud of Things (CoT) technology. The growing processing capabilities inside cloud computing have necessitated the identification of an optimal balance between energy consumption and service level agreement violations (SLAV). The proposed methodology put forth by the author involves the integration of a novel New Linear Regression prediction model, host underload/overload detection as well as a VM placement policy. The objective of this technique is to effectively mitigate both energy consumption and service level agreement violations. The new linear regression prediction model has been developed with the purpose of forecasting forthcoming CPU use by employing a linear regression line and a mean point (MNP) [7]. The proposed algorithms are assessed using an expanded version of the CloudSim Simulator.

The suggested techniques are evaluated using CloudSim simulation with different types of random workloads as well as PlanetLab real. MadMCHD algorithm shows significant improvements compared to commonly used algorithms such as thr, mad, iqr, lr, and lrr, in minimizing SLA violation rates and VM migrations. The combination of MadMCHD and MPABFD algorithms further reduces SLA violations overall.

3 Proposed Work

In this section, we first briefly introduce Linear Regression, then show the detail of our suggested Linear Regression algorithm for underload host detection. With the help of this algorithm, we decrease energy consumption and improve the performance matrix.

3.1 Linear Regression

Linear regression is a fundamental technique in the field of machine learning, commonly employed for the purpose of predicting the value of a variable based on the value of another variable. The variable to be predicted is commonly referred to as the dependent variable (y). The variable employed for predicting the value of other variables is commonly referred to as the independent variable (x). The objective of the linear regression algorithm is to determine the optimal values for B0 and B1 in order to identify the

most suitable line of best fit. The best fit line is characterized by its ability to minimize the error between projected values and actual values.

To calculate linear regression best fit line.

$$Yi = \beta 0 + \beta 1 Xi \tag{1}$$

where Xi = Independent variable, Yi = Dependent variable, β1 = Slope/Intercept, β0 = constant/Intercept (Fig. 2).

.

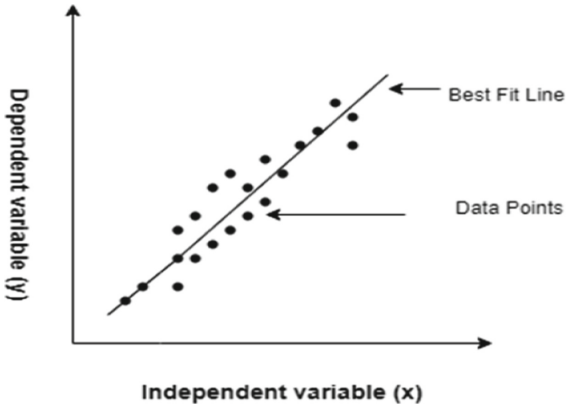

Fig. 2. Linear Regression.

Here's a step-by-step explanation of how to use linear regression for dynamic underload prediction.

- Data Collection
- Data Preprocessing
- Split the Data
- Model Building
- Model Training
- Prediction
- Monitoring and Updating

3.2 Energy Efficient Linear Regression Algorithm for Underload Host Detection

The proposed Energy efficient linear regression algorithm for the host underload detection can be formulated. In the proposed algorithm, we predict the upcoming load on the host on behave of the previous host load history and set the minimum threshold value. The host is considered underutilized whenever the host load value is down from the minimum threshold value [9]. This proposed algorithm is helpful in improving energy efficiency and also improving performance.

3.3 Proposed Algorithm

1. Begin method is Host Under Utilized(host)
2. Cast host to Power Host Utilization History and assign it to _host
3. Get utilization History from _host
4. Set length to 10 (adjustable parameter)
5. If utilization History length is less than length,
6. Return host.get Utilization Of Cpu()
7. Create an array utilization History Reversed with length elements
8. Reverse the utilization History and store it in utilization History Reversed
9. Declare estimates as an array and assign the result of getParameterEstimates(utilizationHistoryReversed)
10. Calculate migrationIntervals as ceil (getMaximumVmMigrationTime(_host) / getSchedulingInterval()).
11. Calculate predictedUtilization as (estimates[0] + estimates[1] * (length + migrationIntervals)) * getSafetyParameter()
12. Add a history entry with host and predictedUtilization
13. Return predictedUtilization
14. End method is HostUnderUtilized.

In this algorithm, we take the three-parameter for enhancing performance. The First parameter is SLA violation second parameter is Energy, and the last parameter is No of VM migration.

$$\text{PM} = \text{E} * \text{SLA} * \text{No Of VM Migration} \tag{2}$$

where PM is Performance matrix, E is Energy, SLA is Service level agreement.

This formula is used for calculating overall performance, and after calculating overall performance, we quickly identify which technique gives the best result (Fig. 3).

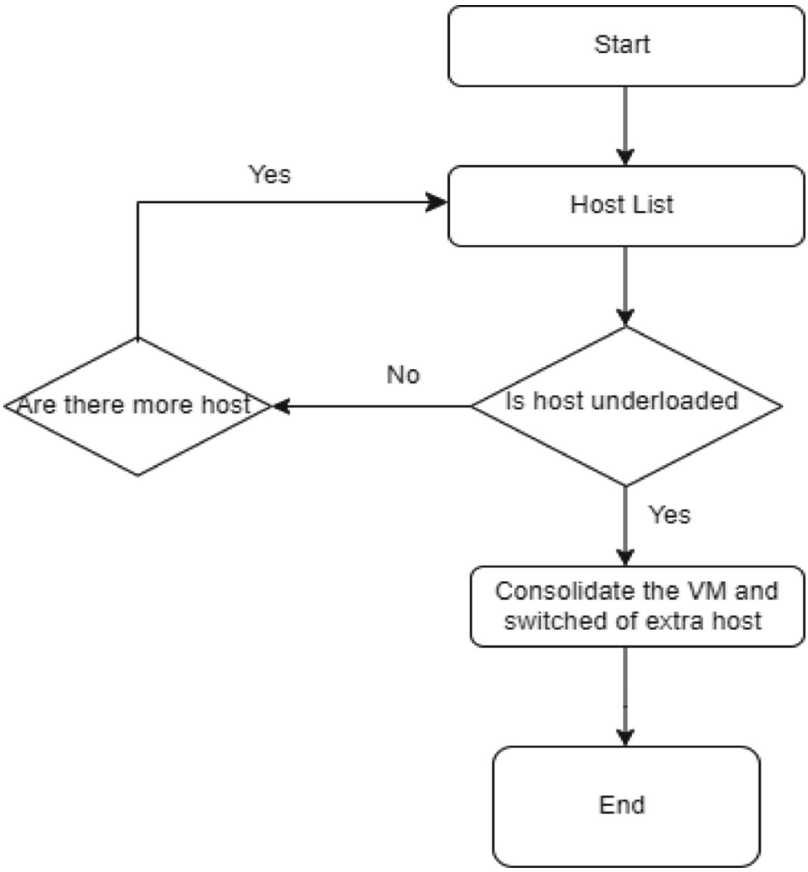

Fig. 3. Workflow Diagram of the Proposed Algorithm.

4 Result

4.1 Simulation Setup

The presented technique is implemented using the CloudSim toolkit 3.0.3 simulator. CloudSim is specifically designed to simulate various components of a cloud system, including virtual machines (VMs) and data centers. It offers support for VM selection and allocation policies, power models, and diverse workload types. The hardware and software requirements for implementing the proposed method are outlined as follows.

Hardware Details. The hardware used in the implementation of the proposed method.

- Processor: Intel(R) Core(TM) i3-6100U CPU @ 2.30 GHz
- RAM: 8 G.B
- Hard disk: 2 TB

Software Details.

- Operating system: Windows 10 Pro 64-bit Version
- IDE: IntelliJ IDEA Edition 2021.3
- Programming Language used: JAVA 17.0.1

4.2 Performance Evaluations

To simulate the proposed method we have used the cloudsim. The performance of proposed technique is measured in the term of the SLA violation, energy consumption, number VM migration. To check the effectiveness and correctness of the presented method we have done comparison with the some existing algorithm like IQRMC, IQRMMT, IQRMU, LRMMT, LRMU, MadMc, MadMMT, MadMu.

Analysis of SLA Voilation. SLA violation occurs when a service provider fails to meet the agreed-upon performance levels or standards specified in the SLA. Analyzing SLA violations is essential to identify the root causes, assess the impact on stakeholders, and take appropriate measures to prevent future occurrences [10, 11]. The Fig. 4 demonstrate the value of SLA violation and finding clearly shows that our proposed method outperformed the existing techniques IQRMC, IQRMMT, IQRMU, LRMMT, LRMU, MadMc, MadMMT, MadMu.

Analysis of Energy Consumption. It refers to the amount of electrical power consumed by data centers and associated infrastructure to support the operation of cloud services. Cloud computing relies on large-scale data centers that house numerous servers, storage systems, networking equipment, and cooling systems, all of which consume significant amounts of energy [12]. Analyzing energy consumption in cloud computing is crucial for improving energy efficiency, reducing environmental impact, and optimizing resource utilization.

Energy consumption of data center is determine with the given formula.

$$(\text{EC}) = E_1 + E_1 + -E_n \tag{3}$$

where E_i= energy consumption of host. E_i Can be calculate using linear interpolation method based on given utilization.

Figure 5 demonstrate the results of the energy consumption of presented approach compare with the other existing techniques and results clearly indicates that presented technique perform better in terms of energy saving.

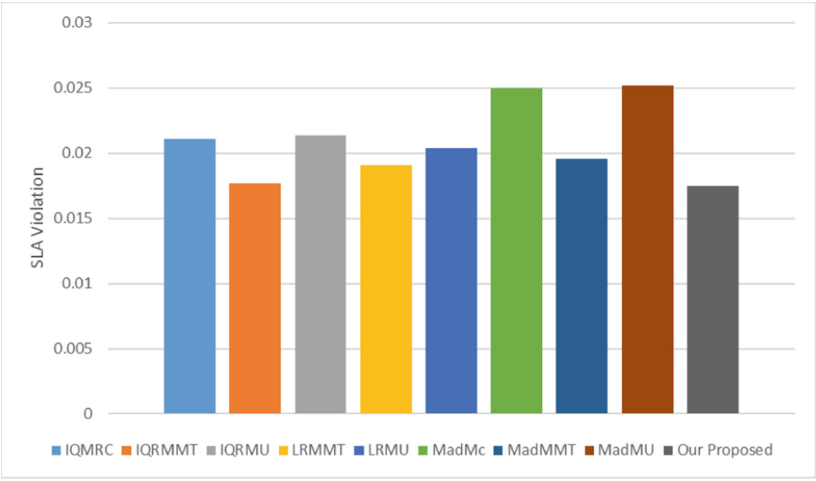

Fig. 4. Analysis SLA Violation.

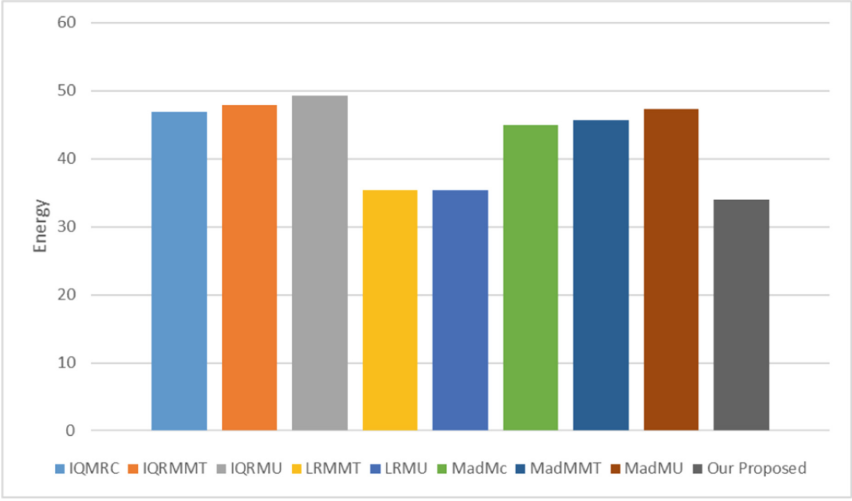

Fig. 5. Analysis of Energy Consumption.

Analysis of VM Migration. VM migration is a process of moving a running or idle virtual machine instance from one physical host or data center to another. This is done for various reasons, including load balancing, resource optimization, hardware maintenance, and disaster recovery. The number of VM metrics can be calculated by utilizing the below formula:

$$\text{VM Mirgration(F,t1,t2)} = \sum_{i=1}^{N} \int_{t1}^{t2} Mig_i(F) \tag{3}$$

where, N is count of VM, $Mig_i(F)$ is migration count of host i in the time interval t1 to t2.

The Fig. 6, clearly shows that the proposed method outperformed the existing techniques.

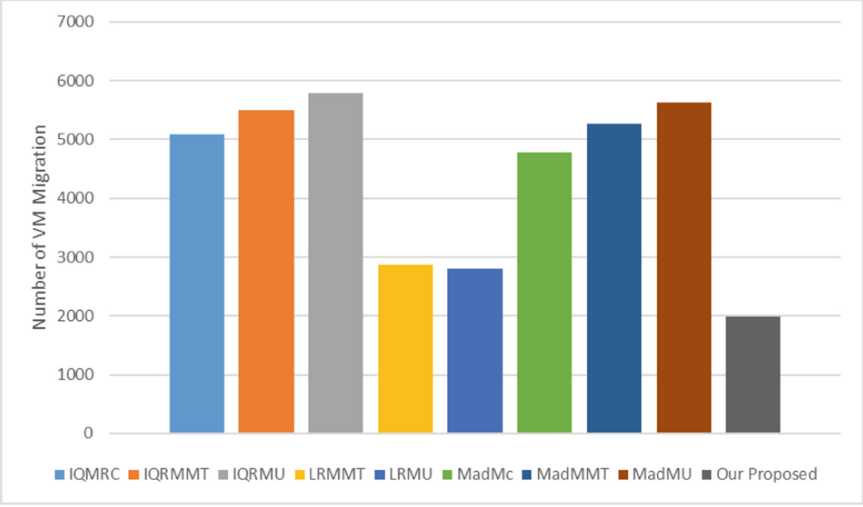

Fig. 6. Analysis of VM Migrations.

5 Conclusion

A dynamic underload host detection approach has been proposed using linear regression. The objective of the presented approach is to minimize consumption of energy, SLA violation and the number of VM migrations. With the increasing demand of cloud computing the energy consumption is also rising sharply. Hence, energy saving becomes prime concern in cloud because it also helps in reduction of cost, reduction in carbon emission and it also increase the performance of the cloud. The efficiency and accuracy of the presented technique is evaluated in terms of SLA violations, consumption of energy, and the number of migrations as performance metrics. The experiments are conducted over cloudsim tool. The performance of the presented technique is compared with

the standard existing techniques like IQMRC, IQRMMT, IQRMU, LRMMT, LRMU, MADMC, MADMMT, MADMU. The results shows that the presented approach outperforms existing techniques in terms of consumption of energy, SLA violation as well as number of VM migrations. In the future, we can leverage real-time data for host underload prediction to improve the accuracy and efficacy of the proposed algorithm in real-time scenarios.

References

1. Saadi, Y., El Kafhali, S.: Energy-efficient strategy for virtual machine consolidation in cloud environment. Soft. Comput. **24**(19), 14845–14859 (2020)
2. Minarolli, D., Mazrekaj, A., Freisleben, B.: Tackling uncertainty in long-term predictions for host overload and underload detection in cloud computing. J. Cloud Comput. **6**, 1–18 (2017)
3. Horri, A., Mozafari, M.S., Dastghaibyfard, G.: Novel resource allocation algorithms to performance and energy efficiency in cloud computing. J. Supercomput. **69**, 1445–1461 (2014)
4. Patel, N., Patel, H.: Energy efficient strategy for placement of virtual machines selected from underloaded servers in compute cloud. J. King Saud Univ.-Comput. Inf. Sci. **32**(6), 700–708 (2020)
5. Ding, W., Luo, F., Han, L., Gu, C., Lu, H., Fuentes, J.: Adaptive virtual machine consolidation framework based on performance-to-power ratio in cloud data centers. Futur. Gener. Comput. Syst. **111**, 254–270 (2020)
6. Alboaneen, D.A., Pranggono, B., Tianfield, H.: Energy-aware virtual machine consolidation for cloud data centers. In: 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, pp. 1010–1015. IEEE, December 2014
7. Biswas, N.K., Banerjee, S., Biswas, U., Ghosh, U.: An approach towards development of new linear regression prediction model for reduced energy consumption and SLA violation in the domain of green cloud computing. Sustain. Energy Technol. Assess. **45**, 101087 (2021)
8. Hsieh, S.Y., Liu, C.S., Buyya, R., Zomaya, A.Y.: Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers. J. Parallel Distrib. Comput. **139**, 99–109 (2020)
9. Li, L., Dong, J., Zuo, D., Wu, J.: SLA-aware and energy-efficient VM consolidation in cloud data centers using robust linear regression prediction model. IEEE Access **7**, 9490–9500 (2019)
10. Wang, J., Gu, H., Yu, J., Song, Y., He, X., Song, Y.: Research on virtual machine consolidation strategy based on combined prediction and energy-aware in cloud computing platform. J. Cloud Comput. **11**(1), 1–18 (2022)
11. Melhem, S.B., Agarwal, A., Goel, N., Zaman, M.: Markov prediction model for host load detection and VM placement in live migration. IEEE Access **6**, 7190–7205 (2017)
12. Kulshrestha, S., Patel, S.: An efficient host overload detection algorithm for cloud data center based on exponential weighted moving average. Int. J. Commun. Syst. **34**(4) (2021)
13. Daraghmeh, M., Melhem, S.B., Agarwal, A., Goel, N., Zaman, M.: Linear and logistic regression based monitoring for resource management in cloud networks. In: 2018 IEEE 6th International Conference on Future Internet of things and Cloud (FiCloud), pp. 259–266. IEEE, August 2018
14. Abdelsamea, A., El-Moursy, A.A., Hemayed, E.E., Eldeeb, H.: Virtual machine consolidation enhancement using hybrid regression algorithms. Egyptian Inf. J. **18**(3), 161–170 (2017)
15. Alhammadi, A.S.A., Vasanthi, V.: Multiple regression particle swarm optimization for host overload and under-load detection. TEST Eng. Manag. **17**(2), 1109 (2020)

16. Nehra, P., Nagaraju, A.: Host utilization prediction using hybrid kernel based support vector regression in cloud data centers. J. King Saud Univ.-Comput. Inf. Sci. **34**(8), 6481–6490 (2022)
17. A El-Moursy, A., Abdelsamea, A., Kamran, R., Saad, M.: Multi-dimensional regression host utilization algorithm (MDRHU) for host overload detection in cloud computing. J. Cloud Comput. **8**(1), 1–17 (2019)
18. Jararweh, Y., Issa, M.B., Daraghmeh, M., Al-Ayyoub, M., Alsmirat, M.A.: Energy efficient dynamic resource management in cloud computing based on logistic regression model and median absolute deviation. Sustain. Comput. Inf. Syst. **19**, 262–274 (2018)
19. Yadav, R., Zhang, W., Li, K., Liu, C., Shafiq, M., Karn, N.K.: An adaptive heuristic for managing energy consumption and overloaded hosts in a cloud data center. Wireless Netw. **26**(3), 1905–1919 (2020)
20. Sissodia, R., Rauthan, M.S., Barthwal, V.: A multi-objective adaptive upper threshold approach for overloaded host detection in cloud computing. Int. J. Cloud Appl. Comput. (IJCAC) **12**(1), 1–14 (2022)
21. Hema, M., Raja, S.: An efficient framework for utilizing underloaded servers in compute cloud. Comput. Syst. Sci. Eng. **44**(1), 143–156 (2023)
22. Mao, L., Chen, R., Cheng, H., Lin, W., Liu, B., Wang, J.Z.: A resource scheduling method for cloud data centers based on thermal management. J. Cloud Comput. **12**(1), 1–18 (2023)