A Comprehensive Review on Autonomous Consolidation of Virtual Machine for Energy and Resource Management

Gaurav Bajpai*

Department of Computer Science and Engineering, Amity University Uttar Pradesh, Lucknow Campus, India, gaurav.bajpai@s.amity.edu

Pawan Singh

Department of Computer Science and Engineering, Amity University Uttar Pradesh, Lucknow Campus, India psingh10@lko.amity.edu

Abhay Kumar Agarwal

Department of Computer Science and Engineering Kamla Nehru Institute of Technology, Sultanpur, India abhaykumaragarwal@knit.ac.in

ABSTRACT

Energy efficiency and awareness of power utilization are significant challenges while building up a cloud computing infrastructure and next-generation data centers. Consolidating Virtual Machines (VMs) is frequently used to lessen data centers' negative impact on the environment and their utility bills. When idle or underused hosts enter sleep mode, it is essential to consider workload variations and energy-efficient and optimized resource utilization. Excessive load on an individual server decreases the Quality of Service (QoS). VM consolidation is supported by the procedures of placing and migrating virtual machines. Problems with variability and scalability of hardware facilities, load volatility, and migration complexity all contribute to the difficulty of the VM consolidation process. This study provides an in-depth look at the strategies and problems of virtual machine consolidation in terms of energy efficiency, resource allocation, migration, and placement schemes. This study proposes the reduction of carbon emissions, which directly-indirectly saves the environment.

Additional Keywords and Phrases: Energy Efficiency, Virtual Machines, Quality of Service, Artificial Intelligence, Classification.

* Place the footnote text for the author (if applicable) here.

1 INTRODUCTION

Recent years have seen a significant increase in the complexity, scale, and power consumption of datacenters. The researchers are up against a formidable obstacle in cutting energy use drastically. Therefore, energy management is the most critical concern that needs close attention. Resource estimation has become more potent, affordable, and widely accessible owing to significant advancements in storage and processing time technology and the successful usage of the internet. Organizational teams seek ways to save costs without compromising performance standards. According to NIST, cloud computing is a deployment methodology for providing simple, on-demand access to a common pool of re-configurable computational resources that can be quickly deployed and released with little tasks or managed service involvement. The ability of any device or virtual computer to have precisely the correct quantity of network connections makes cloud computing an innovative and frequently accessible business strategy. When it comes to cloud computing, a large number of cloud centers store the data in an orderly fashion; therefore, it is clear that there is a significant demand for resource usage. It was predicted that by 2020, the datacenters power industry will be worth $17.4 billion. Assuming a CAGR of 8.5% from 2020 to 2027, it is expected to reach $30.9 billion by the end of the forecast period. In 2012, datacenters accounted for around 2% of the world's total power consumption, projected to rise to 3% by 2020. Energy-efficient design encompasses not just computers and other electronic gadgets but also their supporting infrastructure, including air conditioners and backup power generators. By promoting greater usage of the datacenter's available resources, consolidating the virtual machines can bring about significant benefits for cloud computing. Static and dynamic VMC algorithms are crucial for datacenters energy efficiency. Static VMC minimizes energy use whereas dynamic VMC increases resource use. Forced consolidation is the biggest downside of VMC. Static Threshold based dynamic VMC cannot control SLA breaches. Dynamic VMC based on SLA threshold limit breaches by preceding VM migration from overloaded PMs. Dynamic VMC methods based on adaptive thresholds minimize energy usage but increase VM migrations. In virtualized Clouds, live and offline VM movement enabled dynamic VMC to utilize workload variability. VM migration causes performance and energy overheads, necessitating careful inspection and clever strategies for avoiding unproductive migrations. The work in this paper conducts a comprehensive review of the consolidation of the virtual machine with energy and resource management. Section two of the article gives a brief of the motivation and the previous work done in the consolidation of the virtual machine. Section three provides an insight into the artificial intelligence-based consolidation of the virtual machine and autonomous resource management. Section four of the study presents our work using reinforced learning (RL).

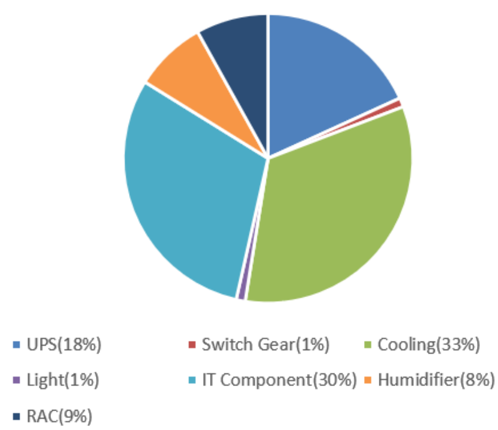

Figure 1: Energy Utilization by cloud data center

2 PREVIOUS WORK AND MOTIVATION

This section focuses on the various works carried out in the field of machine learning for cloud computing. Beloglazov et al. [1] proposed that datacenters' resources to client applications with the suggested energy-aware allocation heuristics increase the datacenter's energy efficiency while still meeting the agreed-upon Quality-of-Service SLA (QoS). The performance metric is the energy a datacenter's hardware utilizes because of the software workloads. They created a virtual datacenter with one hundred physically distinct computers. The findings demonstrate that the penalty produced by 1.1% SLA breaches is reasonable given the 66% improvement in energy savings gained utilizing the MM (Minimization of Migration) policy as opposed to the non-migration aware DVFS (Dynamic Voltage and frequency scaling) policy. VCONF's efficacy is demonstrated using a testbed of cloud environments consisting of XenVMs and typical server loads. The method has strong adaptation and scalability and can locate ideal solutions in smaller networks [12].

Parallelizing Q-learning speeds up convergence to excellent policies. Agents approximate optimum rules and share their learning experiences to increase performance. Policy convergence speeds up when there is a considerable exchange of information by agents. Simultaneous agent learning and our innovative state space formalism enable enhanced uncertainty argumentation using cloud resources. Convergence time is reduced due to parallel learning when the rate parameter is shifted to 100 and 150 (reqs/sec). The Reinforcement Learning based Dynamic Consolidation technique achieves continuous energy and performance optimization without requiring prior knowledge of workload and adjusts dynamically to the situation. On a 250 MIPS CPU with 100% utilization, the program runs for 150,000 MI or 10 minutes of execution and average SLA violation percentages in the actual workload of 15.6%, 33.6%, 44.3%, and 46.8% of the energy consumption may be attained. In the randomized workload, IQR approaches are used accordingly [4].

Authors in [14] shows the increased effectiveness achieved by utilizing a more complex and agile RL approach, with the innate capacity to adapt effectively to a constantly evolving cloud environment. Advanced Reinforcement Learning Consolidation Agent (ARLCA) achieved an overall decrease in power utilization of 25%, with an estimated average energy efficiency of 39.7 kWh.

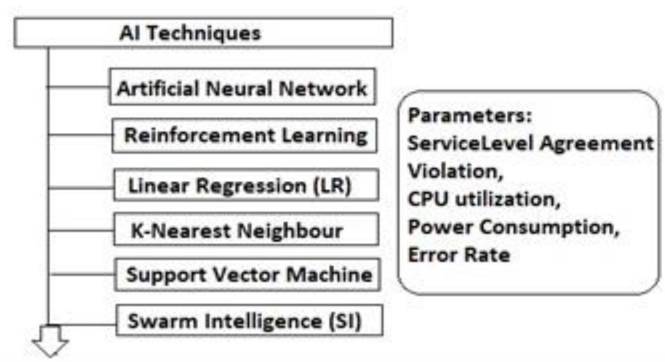

Figure 2: Energy Utilization by cloud data center

The approach suggested by [8] weighs the importance of energy utilization and VM performance loss, which would be ultimately resolved with the greedy method. Despite the RLVMP's success in reducing energy use, issues like SLA breaches remain unresolved. On a cloud platform, you may leverage the process of learning the best scheduling decision approach to make scheduling decisions.

3 VIRTUAL MACHINE MANAGEMENT WITH AI

This section provides insight into the work that has been done in the consolidation, allocation, and migration of the VMs. Moreover, several studies have focused on the use of AI techniques for autonomously allocating cloud computing resources. To address this dynamic virtual machine (VM) consolidation issue for enhanced energy efficiency and SLA delivery [2], the authors suggested an architecture incorporating energy-aware consolidation algorithms with Machine Learning strategies. They developed a Supervised Machine Learning technique that uses a prediction model to calculate the performance and energy costs associated with specific workloads. Their method dynamically shuts down idle hosts and redistributes workloads as needed. To improve the functionality of real-time applications in the cloud, the study presents a new paradigm based on artificial neural networks. The researcher employs the genetic algorithm for the suggested system's training phase. Reducing task wait time, power usage, and task execution time are just a few ways the system has enhanced performance. The superiority of this model over preexisting ones is also demonstrated through comparison. It improves execution time by 36 %, scheduling time by 77.14 %, and energy efficiency by 13 %. The study employed [5] an artificial neural network to identify host computers as underloaded, normal, and overloaded. The improved best fit reducing approach is also analyzed using an artificial neural network. The modified best fit reducing approach monitors power allocation. It helps update virtual machine allocation and migration minimal power. Results show allocation accuracy of 99.99 %. This model's average gradient and error rate are 54.08 and 0.001478. Buyya et al. [1] suggested a new strategy for energy-efficient VMC that relies on threshold auto upgrades. The generated methods were evaluated using large-scale simulations. The experimental results show that the proposed technique outperformed existing migration-aware rules regarding SLA violations of less than 1 % and VM migrations using less energy. The author in the study [2] analyzed single VM migration and dynamic VMC issues. They identified competitive online deterministic proportions for various problems. Their study concludes that randomized or adaptable algorithms improve algorithm efficiency. Based on exam findings, they designed an adaptable heuristic that investigates resource usage for power and the efficient process of VMC. Kaushar et al. [6] proposed "SLA and Energy-effective Dynamic VMC," which satisfies QoS and SLA standards. This research examines VMC algorithms using different heuristics on a real host. The authors compare existing VMC processes using Cloudsim toolkit. This research aimed to help cloud providers examine the power properties of their innovations and current resources to determine their suitability for moving to energy-efficient cloud architectures. The results appraise present systems and save energy for handling SLA QoS requirements. Using an artificial neural network, the study [11] predicted future data centre availability. The virtual machine's actions are run on forthcoming resources. The work shows that by adopting this strategy, data center computation costs reduce significantly. It also facilitates application responsiveness by improving the VM migration pathways. The author in [13] allocated the virtual machine using the EMBFDA (Enhanced - Modified Best Fit Decreasing). The artificial neural network is used to cross-validate virtual machines utilizing E-MBFD. This analysis helps in finding incorrect allocations. This research assists in VM reallocation, minimizes power usage, and has fewer SLA violations than older methods. Li et al. [7] described a system using basic linear regression. This method predicts energy usage and SLA breaches in cloud data centers. This research is unique since it does not employ native linear regression. Instead, the researchers used eight alternative ways to analyse the suggested model's inaccuracy. This research found that the suggested model reduced SLA violations by 99.16 % and real-world energy usage by 2.436 %.

The author presents a weighted linear regression prediction approach is presented by the author using heuristic information from prior server projections, and the reported method predicted future network bandwidth, RAM, and CPU usage. Live migration can use this strategy to predict host overloading; it also it reduces energy use and SLA breaches.

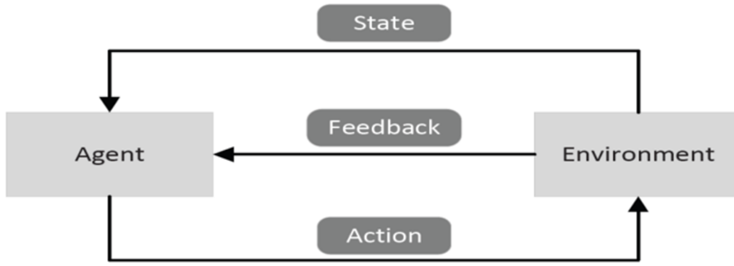

Figure 3: Energy Utilization by cloud data center

The k-nearest neighbour classifier classifies unlabelled observations. It organizes the findings in classes with the most similar-labelled occurrences. Simple classification machine learning algorithm. Simple and efficient categorization approach. Unlabelled and labelled observations' Euclidean distance is determined. Unlabelled observations are categorized by minimum distance and maximum neighbours [20]. The paper [20] proposed an algorithm that examined prior scheduling tasks to schedule virtual machines. The approach is based on classification using machine learning classifiers like Nave Bayes and KNN. In this research, cloud management processes are applied to minimize performance deterioration. The suggested method is compared to SVM, and the results show that it is more efficient. It decreases errors to 0.02 %. The research [16] offered a machine learning-based approach to assign maximal host resources while maintaining SLA. According to this research, creating many virtual machine instances helps cloud providers maintain and use datacentre resources more effectively. The suggested approach is compared to SVM and regression tree. The test results reveal that these models perform worse than the suggested model. The researcher develops a method to assign PMs virtual machines [10]. This method uses a Support Vector Machine (SVM). This virtual machine allocation strategy helped maximize resource usage and minimize virtual machine migrations during system training. Large datacentres also utilize PMs and VMs. The researchers said the algorithm could establish a balanced, resilient, and stable network.

The study [10] provided a methodology for predicting time series using statistical data. Using auto-regression integrated moving average and support vector regression, the researcher suggested a time series model. Classification using SVM. The real dataset is used to test the created model for total pages transferred and migration time. The SVR model predicts dirty pages with 94.6 % accuracy, while ARIMA only achieves 91.7 %. The study also showed that employing a SVM and a support vector machine regression method was accurate and appropriate compared to other machine learning systems. Swarm intelligence works with collaborative algorithms [15]. The study [9] presented the Fruit Fly Hybridized Cuckoo Search Algorithm to optimize energy usage. This strategy helped the researcher reduce cloud datacentre energy and resource use. This approach is compared to particle swarm optimization(PSO), ant colony system, and genetic algorithm. The research concluded that the proposed technique is more energy-efficient than ACS, Genetic Algorithm(GA), and PSO. Comparatively, it utilized 71 % less infra and 67 KwH less power.

Their technique addresses data transport costs and execution time, like Wei et al. [18]. The authors introduced a PSO technique where particles represent workflow tasks linked to cloud resources. Each particle is assessed using a fitness function to find the best task-to-resource mapping.

Table 1: Analysis of machine learning approaches in the consolidation of VM and resource management in the cloud computing environment

Authors	Algorithm used	Parameters/Resources	Tools used	Pros/Cons	Findings and Gaps
Berral et al. [2]	Linear Regression, greedy approach	Power, CPU utilization, SLA timing	OMNeT++	significant energy and performance enhancements	SLA drops: 1%, Energy consumption improved by 10%, RL-based approach may be used in optimizing search space for scheduling
Kaushar et al. [6]	ANN	SLA, QoS	CloudSim	literature survey on SLA, Dynamic VM consolidation	Need of scalable and reliable datacentres
Radhakrishnan et al. [11]	ANN, VMMDA, RDFA	Processing time, response time	MATLAB, Cloud Analyst	reduced response time,	other cloud services may be incorporated for better results
Patel et al. [10]	SVM	Dirty page, time for migration, downtime	Xen	predicting time series using statistical data	Dirty page prediction:94.6%, Different regression models can further be used
Kakkar et al. [5]	ANN, MBFD (Modified Best Fit Decreasing)	CPU utilization, SLA timing future	CloudSim	allocation accuracy of 99.99 %	average gradient and error rate are 54.08 and 0.001478
Li et al. [7]	Linear Regression	future CPU consumption	CloudSim	energy reduction by 25.5%	SLA violation:99.2%
Balaji Naik et al. [9]	Swarm intelligence	Migration, execution time, CPU, convergence	CloudSim	utilized 71% less infra and 67 KwH less power	network overhead and SLA violations need to be evaluated
Shalu et al. [13]	ANN, E-MBFD	Energy Utilization, SLA	CloudSim	Reduced SLA violation, energy-efficient than conventional	false allocations can further be optimized
Talwani et al. [16]	KNN	CPU utilization, bandwidth, memory	CloudSim	enhanced SLA	VM migration:100%, FNR: 82%, Multi-class ML can further be used
Kavya et al. [20]	KNN, NB	CPU utilization, bandwidth, memory	28 physical machines	reduced number of tasks,	The rate of error is reduced by 0.025%,and RF and DT methods can be evaluated

4 VIRTUAL MACHINE MANAGEMENT WITH REINFORCEMENT LEARNING

The fields of Machine Learning and Artificial Intelligence have taken a keen interest in RL in past years. For the purposes of both learning and optimization, RL is the more appealing option. In its most basic form, it is predicated on the idea of a self-reliant agent having the capacity for autonomous learning through experience and observation of its surroundings. For reinforcement learning, the focus is on how an agent should respond in a changing environment to maximize the rewards it receives over the long run based on some abstract objective. Markov Decision Processes (MDPs) are a natural fit for modelling RL control issues because they give a framework for sequential decision making in the presence of unfavourable ambiguity. In order to properly distribute RB and regulate interference in 5G networks, a resource allocation strategy that incorporates an online learning algorithm has been developed. Resource allocation is controlled by high-abstraction functions included in the system's architecture. the method demonstrates a higher level of success in terms of throughput (20%), spectral efficiency (SE), fairness (F), and outage ratio (OR) for transmissions at varying tiers. The study [19] suggested scheduling VM migration alongside upgrading any real computer to save upgrade time. Deep reinforcement learning was used to construct a scheduler. RAVEN is an experience driven scheduler. By interacting with the environment, the system may determine the minimal migration time. The results show that the simulated scheduling method is more successful than other methods. The work in the paper [17] suggested optimum allocation strategies utilizing reinforcement learning and fuzzy logic. This system improved datacentre energy usage and SLA violations. Cloudsim is used to simulate this proposed solution. During the allocation process, the datacentre's infrastructure efficiency, power use effectiveness, and CPU utilization are reviewed and graphically shown. The created framework improves energy efficiency by considering these criteria, according to the results. The author in [19] proposed a dynamic RL-based approach for optimal VM selection and migration. That result reveals that as compared to the traditional technique, a learning agent reduces energy usage and migrations.

Table 2: Analysis of reinforcement learning approaches in the consolidation of VM and resource management in the cloud computing environment

Authors	Algorithm used	Parameters/Resources	Tools used	Pros/Cons	Findings and Gaps
Thein et al. [17]	RL, Fuzzy	PUE, SLA violation, CPU utilization	CloudSim, PlanetLab	Reduced processing time	The limitation is that it considers only energy source
Duggan et al. [3]	RL (Lr-RL)	CPU utilization, migration	CloudSim	reduced co2 emissions	95% confidenc, pvalue<<0.00016
Thien et al. [17]	RL	CPU utilization, migration	CloudSim	enhanced energy levels	SLA violation: 12.2%, DC: 8.4%
Ying et al. [19]	DRL (Raven)	CPU, memory, convergence	CloudSim	Reduced time for datacenter upgrade	Reduced migration time than Min-DIFF, sparse_reward_proble exists
Abbas et al. [21]	Deep RL	energy conservations, SLA violations	PlanetLab on cloud sim	maximizing energy conservation minimizing the SLA violations	Optimized prediction can be implemented

The work done [17] in Reinforcement learning increases energy conservation and throughput in cloud datacentres by consolidating virtual machines dynamically. Virtual machine consolidation utilizing reinforcement learning saved energy and improved performance, according to experiments. The author [3] used a placement mechanism using optimization techniques. Although it significantly impacts the security of the datacentres. A multiobjective methodology using deep reinforce learning is proposed for efficient use of power and performance. The study in the paper [22] suggested a performance model of convergence speed up using reinforcement learning. Error reduction in Qfunction is also done for faster initialization. The study provides significant work on deep reinforce learning (DRL) for the VMs. It aims at maximizing energy conservation ns minimizing SLA violations.

5 CONCLUSION

One of the major challenges in cloud computing is the efficient migration and distribution of virtual machines inside the cloud datacenters. Researchers have worked hard to come up with novel strategies and algorithms to address this optimization issue. Reinforcement learning is a potential way to autonomically modify cloud application resource allocations. Reinforcement learning methods require attention and experience to handle the primary criteria of self-adapting cloud infrastructures: appropriate allocation policies from the start, rapid convergence to the best policy, and capacity to handle application performance model change. The detailed analysis of the review finds that while considering RL with greedy approach power utilization improved by 10 % and with the adaptive threshold approach the SLA violation reduced to less than 1 %. With the use of ANN Modified Best Fit Decreasing (MBFD) the allocation accuracy reached up to 99.9 % and with KNN the VM migration achieves 100 % by Talwani et al. The use of DRL raises some security concerns as well. Summarizing to the review the study finds that there is a scope and need in the field of RL to further consider different resource optimization parameters to enhance the consolidation and allocation of the virtual machine in the cloud computing environment.

This survey helps to conclude that cloud services can assist many requests by remote users with the help of immense computing power, storage capacity, and high-end networking resources. Today, one can provide diverse and enriching services in a Cloud environment using the multi core architecture of VM and high-speed internet connectivity. Also, various related articles that we have considered in our work show that through cloud computing, users can solve multi-structure complex tasks with the help of proper resource utilization. One crucial aspect is that the best resource allocation strategy has to be applied to make cloud resources work efficiently because user-defined jobs need to be scheduled on various machines. Hence, virtualization helps us achieve better resource utilization by creating different VMs over physical servers, improving resource utilization and abstraction.

REFERENCES

[1] Anton Beloglazov, Jemal Abawajy, and Rajkumar Buyya. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future generation computer systems, 28(5):755–768, 2012.

[2] Josep Ll Berral, Íñigo Goiri, Ramón Nou, Ferran Julià, Jordi Guitart, Ricard Gavaldà, and Jordi Torres. Towards energy-aware scheduling in data centers using machine learning. In Proceedings of the 1st International Conference on energy-Efficient Computing and Networking, pages 215–224, 2010.

[3] Martin Duggan, Kieran Flesk, Jim Duggan, Enda Howley, and Enda Barrett. A reinforcement learning approach for dynamic selection of virtual machines in cloud data centres. In 2016 Sixth International Conference on Innovative Computing Technology (INTECH), pages 92–97. IEEE, 2016.

[4] Fahimeh Farahnakian, Pasi Liljeberg, and Juha Plosila. Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning. In 2014 22nd Euromicro International Conference on Parallel, Distributed, and Network- Based Processing, pages 500–507. IEEE, 2014.

[5] Devika Kakkar and GS Young. Heuristic of vm allocation to reduce migration complexity at cloud server. In Proceedings of the International Conference on Scientific Computing (CSC), pages 60–66. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2018

[6] Heena Kaushar, Pankaj Ricchariya, and Anand Motwani. Comparison of sla based energy efficient dynamic virtual machine consolidation algorithms. International Journal of Computer Applications, 102(16):31–36, 2014.

[7] Lianpeng Li, Jian Dong, Decheng Zuo, and Jin Wu. Sla-aware and energy-efficient vm consolidation in cloud data centers using robust linear regression prediction model. IEEE Access, 7:9490–9500, 2019.

[8] Saiqin Long, Zhetao Li, Yun Xing, Shujuan Tian, Dongsheng Li, and Rong Yu. A reinforcement learning-based virtual machine placement strategy in cloud data centers. In 2020 IEEE 22nd International Conference on High Performance Comput- ing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pages 223–230. IEEE, 2020.Prokop, Emily. 2018. The Story Behind. Mango Publishing Group. Florida, USA.

[9] Banavath Balaji Naik, Dhananjay Singh, and Arun B Samaddar. Fhcs: Hybridised optimisation for virtual machine migration and task scheduling in cloud data center. IET Commun., 14(12):1942–1948, 2020.

[10] Minal Patel, Sanjay Chaudhary, and Sanjay Garg. Machine learning based statistical prediction model for improving performance of live virtual machine migration. Journal of Engineering, 2016, 2016.

[11] A Radhakrishnan and V Kavitha. Energy conservation in cloud data centers by minimizing virtual machines migration through artificial neural network. Computing, 98:1185–1202, 2016

[12] Jia Rao, Xiangping Bu, Cheng-Zhong Xu, Leyi Wang, and George Yin. Vconf: a reinforcement learning approach to virtual machines auto-configuration. In Proceedings of the 6th international conference on Autonomic computing, pages 137–146, 2009.

[13] Shalu and Dinesh Singh. Artificial neural network-based virtual machine allocation in cloud computing. Journal of Discrete Mathematical Sciences and Cryptography, 24(6):1739–1750, 2021.

[14] Rachael Shaw, Enda Howley, and Enda Barrett. An advanced reinforcement learning approach for energy-aware virtual machine consolidation in cloud data centers. In 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), pages 61–66. IEEE, 2017.

[15] Stelios Sotiriadis, Nik Bessis, and Rajkumar Buyya. Self-managed virtual machine scheduling in cloud systems. Information Sciences, 433:381–400, 2018.

[16] Suruchi Talwani, Jimmy Singla, Gauri Mathur, Navneet Malik, NZ Jhanjhi, Mehedi Masud, and Sultan Aljahdali. Machine-learning-based approach for virtual ma- chine allocation and migration. Electronics, 11(19):3249, 2022.

[17] Thandar Thein, Myint Myat Myo, Sazia Parvin, and Amjad Gawanmeh. Rein- forcement learning based methodology for energy-efficient resource allocation in cloud data centers. Journal of King Saud University-Computer and Information Sciences, 32(10):1127–1139, 2020.

[18] Guiyi Wei, Athanasios V Vasilakos, Yao Zheng, and Naixue Xiong. A game- theoretic method of fair resource allocation for cloud computing services. The journal of supercomputing, 54:252–269, 2010.

[19] Chen Ying, Baochun Li, Xiaodi Ke, and Lei Guo. Raven: Scheduling virtual machine migration during datacenter upgrades with reinforcement learning. Mobile Networks and Applications, 27(1):303–314, 2022.

[20] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. An-nals of translational medicine, 4(11), 2016.

[21] Khizar Abbas, Jibum Hong, Nguyen Van Tu, Jae-Hyoung Yoo, and James Won-Ki Hong. Autonomous drl-based energy efficient vm consolidation for cloud data centers. Physical Communication, 55:101925, 2022

[22] Luca Caviglione, Mauro Gaggero, Massimo Paolucci, and Roberto Ronco. Deep reinforcement learning for multi-objective placement of virtual machines in cloud datacenters. Soft Computing, 25(19):12569–12588, 2021.