BIUST Teaching, Research, & Innovation Symposium 2023 (TRDAIS 2023)
Botswana International University of Science and Technology
Palapye, Botswana, 18 - 19 September 2023
ISSN: 2521-2293

Energy Efficiency for Cloud Data Centers Using Machine Learning in Botswana

Nomsa Puso
Dept. of Computer Science and Information Systems
Botswana International University of Science & Technology
Palapye, Botswana
pn22100155@studentmail.biust.ac.bw

Tshiamo Sigwele
Dept. of Computer Science and Information Systems
Botswana International University of Science & Technology
Palapye, Botswana
sigwelet@biust.ac.bw

***Abstract*—Botswana is adopting cloud computing technology, and in the future, it will be dominated by more cloud data centers that will require more power supply from the grid. Botswana Fibre Networks (BoFinet) has planned to build the biggest cloud data center in the capital city with at least 400 racks, requiring more than 8MW from the power grid. Botswana government services will be hosted in this data center as virtual machines. Currently, Botswana's power supply is less than the demand, leading to power blackouts that have disrupted the subscribers like industrial and healthcare. These power blackouts have negatively impacted the economy of the country. More cloud data centers in Botswana will draw more electricity from the grid, which will cause more power blackouts unless sustainable sources like solar power are used. However, solar power adoption is shallow despite Botswana's high ultraviolet (UV) index of 9, indicating sufficient sunlight. There is a need for sustainable energy-efficient methods in cloud data centers. This paper proposes the most suitable machine learning approach to minimize energy consumption in cloud data centers which is applicable to Botswana. The proposed framework involves virtual machine placement optimization and shutting down low utilization data center servers to save energy while maintaining the quality of service (QoS). Machine Learning is a cutting-edge Industry 5.0 technology that can be applied to optimization for more accurate outcome predictions without being explicitly programmed to do so. The proposed framework will significantly reduce energy consumption and greenhouse gas emissions.**

Keywords*—Energy Efficiency; cloud computing; cloud data centers; virtual machine; quality of service (QoS), machine learning

I. Introduction

Cloud computing is a fast-growing technology that combines information technology (IT) efficiency and business agility major trends. High energy consumption poses as one of the major issues arising globally on cloud data centers. Hence, this paper is concerned with the development of a framework that minimizes or reduces energy consumption while maintaining the energy costs and QoS on the cloud using machine learning. The project basis lies on the utilization of deep reinforcement learning (DRL) as a machine learning algorithm for saving energy in cloud computing, it integrates the ability of feature learning and complex non-linear-function approximation deep learning (DL) ownership with the ability of decision-making reinforcement learning (RL) ownership, allowing the agent to perceive information in high-dimensional space, train models and make decisions based on the received information [1]. Machine learning (ML) has been on the forefront of other technologies when it comes to energy efficiency on cloud data centers.

The main aim of the study is to propose a framework that minimizes energy consumption while maintaining the energy costs and quality of service (QoS) on cloud data centers using machine learning. The overall goal of this study is achieved through the following contributions:

- Provide a literature review on the use of machine learning algorithms for saving energy in the cloud.
- Compare the performance results of each machine learning algorithm when solving energy efficiency problem.
- Develop a reinforcement learning model that would be able to minimize high energy consumption while maintaining energy costs and QoS on the cloud.
- Propose a prediction of underload or overload servers.
- Compare and test the framework with other models.
- Evaluate the proposed framework and publish its results or outcomes.

II. Literature Review

This section outlines state-of-art machine learning techniques and research related to green cloud computing, which reduces a significant portion of energy consumption in various aspects of a cloud computing system. Many researchers highlighted various innovations for energy efficiency in cloud computing. The author in [2] presented work, explored reinforcement learning algorithm for the virtual machine (VM) consolidation problem demonstrating their capacity to optimize the distribution of virtual machine across the data center for improved resource management. Author in [3] applied the random forest (RF) and multilayer perceptron (MLP)-based model to observe the low overhead and less energy consumption without significantly affecting the time to complete the tasks on the cloud. The author in [4] made use of machine learning technique called cooperative reinforcement learning (Q learning) agents for reducing user costs, reducing energy consumption, load balancing of resource, enhancing utilization of resources and improving availability and security.

ISSN: 2521-2293

Also, author in [5] used deep reinforcement learning to propose a novel hierarchical state space formulation coupled with a hybrid actor-critic technique for energy-efficient resource scheduling in edge-cloud environment. The author in [6] developed the ultra-low-power implementation of the DRL framework using stochastic computing technique, which has the potential of significantly enhancing the computation speed and reducing hardware footprint and therefore the power/energy consumption. Table 1 depicts different approaches of machine learning techniques in the field of energy efficiency and the metrics they have considered for the performance evaluation.

TABLE I. SUMMARY OF ENERGY EFFICIENCY-RELATED WORK IN CLOUD ENVIRONMENT

Authors	*Learning Model*	*Objective (Energy saving method)*	*Metrics*	*Limitations*	*Dataset*
Thein, et al., (2018) [7]	**Reinforcement learning and fuzzy logic**	Presented work provides the effective management of physical resources hosted by the infrastructure using dynamic resource demand patterns, Service Level Agreement, and resource utilization.	Accuracy (Power Usage Effectiveness in an efficient range from 1.79 to 1.96, resource utilization above 50%)	It considers only energy sources and the energy consumption for CPU and data centers. For a very large number of infrastructure resources, the scheduling process may become slow.	PlanetLab Virtualized Research datasets
McGough, et al., (2018) [3]	**Random Forest (RF) and MultiLayer Perceptron (MLP)**	Presented the work to observe low overhead and less energy consumption using ML techniques.	Accuracy (45.6-51.4% of the energy can be saved without significantly affecting the time to complete tasks)	The proposed approach uses real trace-logs allowing for complex situations to occur in the presented platform.	2010 exemplar datasets used with High Throughput Computing (HTC)-Sim
Rajalaksh-mi, et al., (2019) [8]	**Reinforcement Learning**	In the presented work, the learning agent improves the quality of the VM consolidation algorithm for energy consumption.	Correlation (the results shows that reinforcement learning (RLVC) algorithm gives minimum SLA violation compared to others by 8.5%)	The number of hosts can be increased to simulate the check the behavior of the proposed work.	The Cloudsim 3.0 use the PLANET LAB workload
Zhang et al., (2018) [9]	**Linear and Logistic Regression**	In the presented work, they use the classification of machine learning to model and analyze the multi-dimensional cloud resource allocation problem.	Response time (98% prediction accuracy)	A discussion on whether the resource allocation algorithm based on machine learning satisfies the strategy proof of the auction mechanism.	DAS-2 [ASCI (2017)] dataset from Grid Workloads Archive
Shaw et al., (2022) [2]	**Reinforcement Learning**	Presented work explores RL algorithm for the VM consolidation problem demonstrating their capacity to optimize the distribution of virtual machine across the data center for improved resource management.	Energy Consumption (energy efficiency is improved by 25% while also reducing service violations by 63% over the popular Power-Aware heuristic algorithm)	One of the fundamental challenges faced by model-free learning agents is tradeoff between exploration and exploitation.	Real workload data from PlanetLab
Madhusud han et al., (2021) [10]	**Genetic Algorithm (GA) and Random Forest (supervised machine learning technique)**	The aim of the work is to minimize power consumption while maintaining better load balance among available resources and maximizing resource utilization.	Energy Consumption, Execution Time, Resource Utilization, Average Start Time and Finish Time (GA-RF model improves energy consumption, execution time, and resource utilization of the data center and hosts as compared to the existing models).	The model was not tested with various machine learning and deep learning approaches for the better solutions and performance study.	Real workload traces from PlanetLab

ISSN: 2521-2293

Authors	*Learning Model*	*Objective (Energy saving method)*	*Metrics*	*Limitations*	*Dataset*
Yan et al., (2021) [11]	**Deep Q-learning (DQN)**	Using machine learning algorithm to achieve multiple optimization goals such as reducing power consumption, ensuring resource load balance, and improving user service quality.	Failure Rate, Average Reward, Power Consumption and SLA Violation (the proposed algorithm outperformed native DQN in terms of convergence speed, Q value estimation accuracy and stability. It has shown the flexibility to achieve multiple optimization goals including power consumption reduction, resource load balancing and SLA quality.)	Advanced DQN was compared with native DQN and simple heuristic algorithms which are easy to implement, but do not have specific optimization goals.	Used CloudSim 4.0 to simulate a cloud data center containing 32 heterogeneous PMs and dynamic arrival of VM requests for one hour
Jayanetti et al., (2022) [5]	**Deep reinforcement learning**	Proposed a novel hierarchical state space formulation coupled with a hybrid actor-critic technique for energy-efficient resource scheduling in edge-cloud environment.	Energy consumption, execution time, percentage of deadline hits and percentage of jobs have been used as evaluation metrics (Proposed DRL technique performed 56% better with respect to energy and 46% with respect to execution time compared to time and energy optimized baselines, respectively.)	The proposed reinforcement learning framework is designed to only operate in a centralized manner.	CloudSim simulation toolkit, dataset was created based on synthetic workflow structures provided by the popular Peagasus workflow framework.
Asghari et al., (2020) [4]	**Cooperative reinforcement learning (Q learning) agents**	Using machine learning technique to reduce user costs, reduce energy consumption, load balancing of resources, enhancing utilization of resources and improving availability and security.	Scheduling time, makespan, resource utilization, cost, power, and energy (The results of the experiments showed that the proposed model is more efficient in comparison with other methods in terms of makespan, resource utilization, cost, and energy consumption.)	Loss of accuracy due to discretization of state space.	Cloudsim, Workflowsim To evaluate the proposed method using real datasets, four classic and standard datasets of this area, namely Montage, Cybershake, SIPHT, and Inspiral have been used. The Pegasus Toolkit has been used also as an open-source toolkit to generate scientific workflows
Li et al., (2018) [6]	**Deep reinforcement learning**	In the presented work, they developed the ultra-low-power implementation of the DRL framework using stochastic computing technique, which has the potential of significantly enhancing the computation speed and reducing hardware footprint and therefore the power/energy consumption.	Power usage, average job latency, power usage and average job latency – achieve up to 54.1% energy saving compared with baselines (All tested cases can achieve at least 47.8% power consumption saving with only a slight increase in job latency. These results prove that weights of the reward function can take an effective control of the trade-off between power, latency, and resiliency)	DRL framework requires a relatively low-dimensional action space due to the fact that at each decision epoch the DRL agent needs to enumerate all possible actions under current state and perform inference using DNN to derive the optimal Q(s, a) value estimate, which implies that the action space in the general DRL framework needs to be reduced.	Real data center workload traces extracted from Google cluster usage traces over month-long period in May 2011. TensorFlow 0.10 is adopted for DNN construction.

ISSN: 2521-2293

A. *Summary of related works*

Deep reinforcement learning, support vector machine regression, k-means, random forest, linear and logistic regression were among the machine learning models used in the literature review, 60% of the authors used DRL technique to solve energy efficiency problems on the cloud. With 40%, CloudSim simulation toolkit is the most frequently adopted for data generation whilst other datasets are taken from different sources and databases. For evaluation purposes, energy consumption 13%, accuracy 9%, power consumption 4%, execution time 4%, resource utilization 4%, average reward 4%, makespan 4%, power usage 4%, average latency 4% and others were used as performance metrics by the authors.

The significant research has been made in recent years to reduce energy consumption in cloud computing. The author in [7] addressed the issue of energy-efficient resource allocation in cloud data centers. Nevertheless, the reinforcement learning, and fuzzy logic-based model had limitations, it only considered energy sources and energy consumption for central processing unit (CPU) and data center, which can be valid for the analysis, however other resources may also influence the scheduling method's outcome. One of the fundamental challenges faced by model-free learning agents is tradeoff between exploration and exploitation [2]. Author in [10] aimed at minimizing power while maintaining better load balance among available resources and maximizing resource utilization, but the model was not tested with various machine learning and deep learning approaches for the better performance study and solutions.

There was a loss of accuracy due to discretization of state space [4]. The author in [6] used a DRL framework which required a relatively low-dimensional action space due to the fact that at each decision epoch the deep reinforcement learning agent needs to enumerate all possible actions under current state and perform inference using deep neural network (DNN) to derive the optimal value estimate, which implies that action space in the general DRL framework needs to be reduced for more accurate results. Compared to the other work, this study proposes a dynamic virtual machine optimization and shutting down low utilization data center servers through the deep reinforcement learning model for saving the energy while maintaining the quality of service.

III. METHODOLOGY

Deep Reinforcement Learning (DRL) which is a combination of RL and deep learning, has overcome many issues through the function approximation, thereby eliminating the need for agents to visit all states during the training process and for storing state transition data in space-consuming tabular formats [5]. The inherent characteristics of the reinforcement learning paradigm, such as learning through experience combined with the use of neural networks for function approximation, make DRL an ideal candidate for dealing with the unpredictable dynamicity associated with cloud computing environments. A DRL technique is widely used for managing complex computing and networking infrastructures because it overcomes the curse of dimensionality by using neural networks as function approximators [12]. Deep reinforcement learning is very popular today because of its ability to solve complex consequential decision-making problems. The DRL model is pre-trained and used in real time for obtaining the scheduling decisions.

Caviglione et al., 2021, states that DRL belongs to a family of reinforcement learning methods that is well suited to dealing with sequential decision making problems. In more detail, they are based on an agent interacting with an environment in discrete steps via actions taken in response to an observation of the environment's state. The environment returns a reward as a result of an action, which is a scalar value measuring the effectiveness of the action. The goal is to maximize the total reward, which is calculated as the sum of the rewards obtained at each iteration. In contrast to supervised learning, rewards do not assign a label to a correct or incorrect answer [12].

The deep reinforcement learning model will be adopted to save energy in the cloud with reasons that it is the most used machine learning model from the literature. It is well known for cost reduction, minimizing energy consumption with high accuracy, easy to scale up and finally it uses deep neural networks function approximators solve energy efficiency issues on cloud data centers.

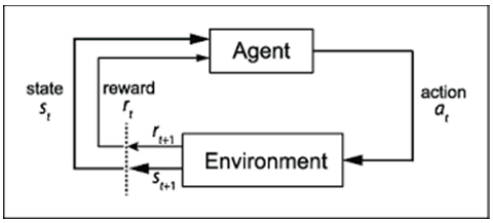

Fig. 1. Illustration of a simple flow of the agent-environment interaction in a Markov decision process

A. *DRL model's five key elements:*

- Agent: The model's algorithm/function that performs the requested task.
- Environment: The environment in which the agent operates. It uses the agent's current states and actions as input, rewards, and the agent's next states as output.
- State: It refers to the agent's position in the environment. There are two states: current and future/next.
- Action: The agent chooses and executes the moves in order to gain rewards.
- Reward: The agent's desired behaviors are referred to as rewards. Rewards, also known as feedback for the agent's actions in a given state, are described in the model as results, outputs, or prizes.

ISSN: 2521-2293

B. *Proposed Methodology*

Fig. 2. Steps of the proposed methodology

1. Gathering data: As the first step, data is collected from various sources, and the type of data collected is determined by the desired project.

2. Preparing that data: Data preparation is the process of cleaning raw data because it may contain missing values, inconsistent values, and duplicate instances, it cannot be used directly to build a model.

3. Choosing a model: The best performing learning algorithms are being researched, and their effectiveness is determined by the type of problem that needs to be solved.

4. Training: This is whereby the model is trained to improve its ability and the training dataset is fed to the learning algorithm.

5. Evaluation: In this stage, the model is evaluated to see if it is any good. To assess performance, metrics such as accuracy, precision, recall, and others are used.

6. Hyperparameter Tuning: Since this is an experimental process stage, increasing the number of testing cycles may result in more precise results.

7. Prediction: Finally, the created system or model is now used to solve real world problems, and this is where the true value of machine learning is realized.

IV. Evaluation

The key cloud performance metrics that are going to be used to evaluate the effectiveness of the proposed algorithm are energy/power consumption, accuracy, makespan, resource utilization, average reward, power usage effectiveness, execution time and average job latency. DRL-based model will be developed as a result of this study and it will involve the virtual machine placement optimization and shutting down low utilization for minimizing high energy consumption, excessive carbon emissions while maintaining the costs and QoS in cloud data centers. The proposed model will surpass the prediction accuracy ranging from 16.20-98% that was taken from the literature review.

V. Conclutions

Due to the ever expanding size of cloud computing facilities and ever-increasing number of users, high energy consumption has become a growing concern in the operation of complex cloud data centers. It does not only result in high costs, but it also produces excessive carbon emissions, which often leads to system unreliability and performance degradation. Energy efficiency is one of the main critical issues in the current cloud computing but since it lowers costs and adheres to green computing principles, ensuring energy efficiency is therefore a significant goal in Botswana. In this paper, DRL model was proposed as the most suitable machine learning technique that can be used for minimizing energy consumption in cloud data centers while maintaining the costs and quality of service. DRL has overcome many energy efficiency problems through the neural networks function approximation method that eliminates the need for the agents to visit all states during the training process. Energy/power consumption, accuracy, makespan, resource utilization, average reward, power usage effectiveness, execution time and average job latency will be used as performance metrics for evaluation. CloudSim simulation as used in the literature, will be adopted in this study to generate the data The proposed model will be better to surpass the prediction accuracy ranging from 16.20-98% that is taken from the literature.

Acknowledgment

I would like to express my deepest gratitude to my supervisor who is always there to guide me in every step of the way. This work is supported by the department of Computer Science and Information Systems, Botswana International University of Science & Technology.

References

[1] H. Fan, L. Zhu, C. Yao, J. Guo, and X. Lu, “Deep reinforcement learning for energy efficiency optimization in wireless networks,” in 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019, 2019. doi: 10.1109/ICCCBDA.2019.8725683.

[2] R. Shaw, E. Howley, and E. Barrett, “Applying Reinforcement Learning towards automating energy efficient virtual machine consolidation in cloud data centers,” Inf Syst, vol. 107, p. 101722, Jul. 2022, doi: 10.1016/j.is.2021.101722.

[3] A. S. McGough, M. Forshaw, J. Brennan, N. al Moubayed, and S. Bonner, “Using Machine Learning to reduce the energy wasted in Volunteer Computing Environments,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.08675

[4] A. Asghari, M. K. Sohrabi, and F. Yaghmaee, “A cloud resource management framework for multiple online scientific workflows using cooperative reinforcement learning agents,” Computer Networks, vol. 179, p. 107340, Oct. 2020, doi: 10.1016/j.comnet.2020.107340.

[5] A. Jayanetti, S. Halgamuge, and R. Buyya, “Deep reinforcement learning for energy and time optimized scheduling of precedence-constrained tasks in edge–cloud computing environments,” Future Generation Computer Systems, vol. 137, pp. 14–30, Dec. 2022, doi: 10.1016/j.future.2022.06.012.

[6] H. Li, R. Cai, N. Liu, X. Lin, and Y. Wang, “Deep reinforcement learning: Algorithm, applications, and ultra-low-power implementation,” Nano Commun Netw, vol. 16, pp. 81–90, Jun. 2018, doi: 10.1016/j.nancom.2018.02.003.

ISSN: 2521-2293

[7] T. Thein, M. M. Myo, S. Parvin, and A. Gawanmeh, "Reinforcement learning based methodology for energy-efficient resource allocation in cloud data centers," Journal of King Saud University - Computer and Information Sciences, vol. 32, no. 10, 2020, doi: 10.1016/j.jksuci.2018.11.005.

[8] N. R. Rajalakshmi, G. Arulkumaran, and J. Santhosh, "Virtual machine consolidation for performance and energy efficient cloud data center using reinforcement learning," Int J Eng Adv Technol, vol. 8, no. 3 Special Issue, 2019.

[9] J. Zhang, N. Xie, X. Zhang, K. Yue, W. Li, and D. Kumar, "Machine learning based resource allocation of cloud computing in auction," Computers, Materials and Continua, vol. 56, no. 1, 2018, doi: 10.3970/cmc.2018.03728.

[10] M. H. S, S. Kumar T, S. M. F. D. S. Mustapha, P. Gupta, and R. P. Tripathi, "Hybrid Approach for Resource Allocation in Cloud Infrastructure Using Random Forest and Genetic Algorithm," Sci Program, vol. 2021, pp. 1–10, Oct. 2021, doi: 10.1155/2021/4924708.

[11] J. Yan, J. Xiao, and X. Hong, "Dueling-DDQN Based Virtual Machine Placement Algorithm for Cloud Computing Systems," in 2021 IEEE/CIC International Conference on Communications in China (ICCC), Jul. 2021, pp. 294–299. doi: 10.1109/ICCC52777.2021.9580393.

[12] L. Caviglione, M. Gaggero, M. Paolucci, and R. Ronco, "Deep reinforcement learning for multi-objective placement of virtual machines in cloud datacenters," Soft comput, vol. 25, no. 19, pp. 12569–12588, Oct. 2021, doi: 10.1007/s00500-020-05462-x.