International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Survey on Maximizing Energy Efficiency and Resource Utilization in Virtual Machines Using Prediction Algorithms

Mr. P.Udayasankaran[1], Dr. John Justin Thangaraj[2]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Cloud computing has received a good response widely now a days. The recent growth of cloud shows many users are already adopting it at an unprecedented rate for both personal and professional requirements, because there are naturally high rate of datacenter deployments and implementations worldwide. Cloud datacenters are known to be significant energy consumers and environmental polluters as a result of this growing adoption. There are many problems presently, including: Resource allocation and Energy. It makes systems poor and expensive. Forecasting methodology can improve cloud efficiency and reduce operational costs. The enhanced dynamic of Cloud systems comes with a number of operational and analytical issues. The maximizing energy efficiency and resource utilization of this research work have various behavioral changes and clients are yet not fully understood. Our paper presents an in depth analysis of challenges and research works carried out in maximizing energy efficiency and resource utilization using various prediction algorithms.

Keywords: *Cloud data center, Prediction algorithm, Resource allocation*

1. Introduction

The affordable hosting is a safe and adaptable computing under the management of contemporary data centres and it is made possible by technological advancements. Due to this higher resource utilization and lower costs good results are obtained. Effective approaches for forecasting data centre resource usage can greatly simplify user and provider for self-management and usage optimization. Users can dynamically modify the leased resources to lower hosting costs while keeping the desired performance for their applications and level of service. Moreover, precise resource projections and usage permits the providers to effectively distribute virtual resources workloads using Virtual Machine, and also Plan ahead for resource capacities and understanding the energy needs for anticipated user load and workload. Various hosted jobs' and loads, resource usage estimation for data centres is difficult. While there are a number of estimating techniques for cloud resource consumption employing deep learning or time series learning, they always use a single model that frequently fails to adequately represent the dynamics of the workload. To develop the efficiency and services eventually deploying many data centres at various areas. Each areas have enormous of servers, components, and heat reduction components. Data centres use high level of energy for the user requests. United States declared in the mid of 2014 , all the DC in the United states used approximately 60 to 70 billion kilowatt-hours, which is approximately 1.8% of the total United States energy consumption. United States projected it will increase more usage up to 200 billion kilowatt-hours in the year 2020 [1]. Various established companies like Google pay, Amazon and Microsoft, using more power consumption and increasing their cost for paying high amount for their usage. We offer a novel approach to automatically and adaptively determine the best model to precisely predict the consumption of data centre resources in order to address the issues. Our adaptive multi-methods methodology permits choosing the predictive method that learns the fastest and also taking into account various conditions experienced in a commercial data centre. Our strategy focuses on developing estimation models using many techniques, then choosing the one that would produce the most accurate forecast given the current situation.

2. Literature review

Beloglazov, A. and Buyya, R [2] presented an architecture consisting of a sizable cloud data center with n different types of nodes with performance measured in MIPS. Users send req for m different Virtual Machines MIPS-based. A VM that violates the SLA cannot obtain the quantity of resource requested, which may occur as a result of VM consolidation. The tier-based software system architecture includes a manager, both local and worldwide. The regional director dwell in a Virtual Machine that is present on each physical node. Each physical node's local managers are a part of a Virtual Machine and they are in charge of monitoring the node's thermal status and

[1]*CSE Department , Saveetha School of Engineering , Tamilnadu , Chennai ,India ORCID ID : 0009-0002-0868-0162*
[2] *CSE Department , Saveetha School of Engineering , Tamilnadu , Chennai ,India ORCID ID : 0000-0002-1384-9340*
** Corresponding Author Email: [1]udayasankaran97@gmail.com*
**[2]johnjustinphd@gmail.com*

resource usage. The information regarding resource usage and the VMs selected for migration is sent by the local management to the global managers. Additionally, they send out commands to apply DVFS, resize virtual machines, and turn on and off idle nodes. Each global manager processes data received from its local managers and is connected to a group of nodes. Limitations of this work include it in initial stage with simple version and a decentralized cloud architectural model. It does not adapt allocation of VMs on runtime and also independent of the workload type. Algorithms of all the proposed optimizations are not fully developed.

Kumar, K.S. and Jaisankar, N. [3] introduces a novel framework that offers effective green computing which upgrades cloud computing architecture. The overall system efficiency in a cloud based data centers will be significantly increased with implementation and effective work by using various methodologies and techniques to obtain resource utilization effectively. One can significantly reduce the power consumption by utilizing the various methodologies to reduce the energy consumption by scheduling the jobs assigned to nodes to lower power consumption. The cost of running the servers is the biggest operating expense in a cloud data center. A typical scheduling system will uniformly distributes the load across all servers in the data center. While scheduler could be reasonable in theory, in fact it is incredibly ineffective when power usage rises to its maximum capacity whenever the scheduler assigns Virtual Machine to a processor. In contrast, this study shows that the power consumption is significantly reduced if the scheduler distributes the virtual machines with the intention of completely utilizing all processor cores within each node. Consequently, there is a significant demand for a sophisticated scheduling system. The limitation of their work includes maximizing energy savings and QoS parameters are not concentrated in their proposed work.

Menaka, M. and Kumar, K.S.[4] propose a scheduling algorithm for container resource migration that we call UVPOC, and we set up scheduling mechanism to watch the real time performance of resource utilization and decide whether to implement resource allocation through pre-booting virtual machines (VMs) or through container migration. We create a multi-level scheduler mechanism that primarily serves as a resource monitor and a data source for LSTM prediction. The first-level scheduler acquires in allocating, managing, and monitoring global resources for physical machines. Each node which are the resources is installed and monitored and keep on watching and tracks the running status of the task and also real time processing of the resources utilized. The module which are logically processed and output results are evaluated and put to use by the Resource allocation module, which also delivers the work performed by scheduling with reference to the virtual machine and various resources. Second-level scheduler is in charge and also specialized in gathering the running status of data and task based on the scheduling outcomes of the first-level scheduler's Resource allocation module. To ensure a durable operation, the first-level scheduler is informed for executing result. Limitation of their work is they proposed a 2 level scheduler. There is no specified procedure; it only anticipates the time series with that knowledge, allowing us to improve depending on our intuition.

Yazdanian, P. and Sharifian, S [5] proposed a ConvNets, a LSTM prediction blocks that are used for processing lengthy sequences of data and trace in google to forecast RAM and CPU requests in the future with accuracy and efficiency. LSTM devices with two stages are stacked in order to boost network capacity. As an alternative to over fitting, we employ the dropout strategy in which all layers which will be selected alternatively to avoid accidental occurrences of the data and the layer which is going to be exposed. At the testing period, to build subsequences from the testing inputs and processes them one at a time to the model. It is performed after various training model of entire data. As a result the model can quickly and accurately estimate the testing strategies with the data used for testing. Limitation of this model is it uses historic data which might change and the paper doesn't give a method for efficient resource management but it predicts the resources required in the future.

Zhang, W et al [6] proposed a method for forecasting workload using Recurrent Neural Networks (RNN). This implementation and design is used to identify the RNN and its parameters that have the greatest influence in order to acquire the optimal parameter set. The design approach known as orthogonal experimental design (OED) is used to address issues with various levels and components. This strategy, which is an effective and economical experimental design method, chooses a portion of typical samples from the total samples to run the experiment. In our work, we employ OED to identify the best RNN parameters. OED tables can quickly demonstrate which set of parameters is the best, even though it is challenging to determine whether the RNN network's parameters fit within the local optimal solution. Limitation of this work is, although the RNN is a powerful element for performing and processing sequences, it performs poorly when faced with lengthy sequences and loses its learning ability. It also has complex computations, and the training phase is lengthy. The paper doesn't give a method for efficient resource management but it predicts the resources required in the future

Kumar, J. and Singh, A.K [7] developed and presents a workload prediction model based on black hole algorithm and neural networks. Able to outperform back propagation

in terms of mean squared error up to 134 times. A method for workload prediction in data centers that addresses scalability. The data center has a workload predictor device installed that uses past workload as input. According to the present condition of the data center will have advanced understanding of resource utilization and servers which are used effectively called as resource management. Limitation of their work includes despite greater precision it performs poorly when faced with lengthy sequences of time series and loses its learning ability which has complex computations, and the training phase is also lengthy. The paper doesn't give a method for efficient resource management but it predicts the resources required in the future.

Wang, X., et al.[8] proposes a Double Multi-Attribute Auction process, where first, various factors are considered to create the Quality Index , helps to thoroughly assess how well customers and suppliers performed in the transactions. In order to anticipate the price, a methodology of Vector technical Machine is used. The optimum allocation plan is then obtained by solving the Mean-Variance Optimization method. The first step is for AO to get CRP and CRC uses Neural technology Method to convert the numerous qualities to QI, and then they employ some methods to forecast the cost and send the tenders to AO. AO starts planning the auction after gathering the bids. Every potential transaction's transaction price is determined by AO, and the MVO algorithm is used to select the auction winners. Limitation of this works on the auction mechanism which depends upon the present demand of cloud services needed, it also predicts the further resources but doesn't give a fixed solution for Efficient Architecture in the Cloud.

Kumar, K.S et al., [9], carried out various preliminary tests to evaluate the impact of various features on prediction accuracy. In particular, we assessed whether a repository's programming language by itself could deliver a high enough degree of prediction performance. In this way, the machine learning model already in use might be used to evaluate new repositories. According to our preliminary investigation, the prediction accuracy is significantly decreased when pooling builds for each language, that is, from numerous sources. The machine learning strategy had an error that was 800 times larger than the baseline approach, as indicated by the median error ratio of about 800. We consequently come to the conclusion that it is not practical to use per language machine learning models. Therefore, we divided the dataset by repository for our methodology. In other words, rather of developing a machine learning model for every language or even the entire world, we train one for each repository. Given that our preliminary studies have demonstrated that various repositories have drastically varied build processes, this seems logical. Limitation of their work includes using this model we can able to find a 20% error rate.

Li, Y., Tang [10] have proposed the following modifications to the basic bin packing algorithms. They refer to the act of receiving an item into an empty bin as opening the bin. We say a bin has been closed when it is once more empty after being opened. They have changed this so that once a bin is closed, it stays closed forever and we never put anything else in it. In the Dynamic Bin Packing problem, where the objective is to reduce the number of bins utilized, such a choice would be ineffective. But for our applications, it makes sense that once a bin (server) is inactive, we stop paying for it and it blends in with all the other idle servers. Additionally, it produces algorithms that are simpler to understand. Limitation of their work is the gap between the current upper and lower bounds on the competitive ratios of First Fit and Hybrid First Fit should be reduced.

Viswanathan, R. and Devi, C.B [11] proposed a concept of utilizing resource manager to keep track of resources for any cloud service provider. The resource manager is responsible for finding the resource, determining its state, etc. The distribution of resources to clients will take a long time under the current paradigm. This out-of-the-ordinary wait time increases computation time and network traffic, both of which have an impact on subsequent processes and result in high customer billing costs. Therefore, the resource manager have to deliver any service on demand, an effective dynamic resource management algorithm is required. The Greedy method, which is possibly the most simple design strategy used for effective Dynamic Resource

Management by Resource Manager, it is also used in the suggested Job Sequencing algorithm. The Greedy technique suggests that the algorithm can be designed to operate in stages. A choice is made at each stage as to whether the chosen option is the best one, or in our situation, whether it is cost-effective. The limitation of this approach is it's hard to achieve dynamic resource management and similarly as a way to prevent global warming, managing the resources at runtime is vital.

Gunasekaran, J.R., [12] proposed an approach to provide effective ways of managing resource allocations for their applications in order to maximize performance and reduce costs for both cloud providers and tenants. The plan includes three interconnected tasks as a means of achieving this. First, we begin from the standpoint of the tenant, with the first two activities intended to look at the main causes of performance-cost inefficiencies. The third assignment also looks into the main causes of performance-energy inefficiency in datacenters from the provider's point of view. The performance and cost effectiveness of new applications on next-generation cloud platforms can both be enhanced by the three tasks taken together. The

downside of this strategy is that consumers may encounter technical issues like reboots, network outages, and downtime due to IT setup configuration.

Van Ewijk, S., [14] presented a resource efficiency based approach paying attention on the importance of the extraction, conversion, use, and disposal of material resources from an economic and environmental standpoint. Businesses, households, or local and national government may be the subject of attention. The authors identify six essential characteristics and illustrate how various interpretations of the two notions are comparable. More information about resource efficiency, the circular economy, and the possible economic benefits is covered in their work. The difficult problem with resource efficiency and the circular economy is the trade-off between the effects on the environment and the economy. Concerning trade-offs between the environment and the economy, there are primarily two difficulties. First, reducing the use of resources could potentially lower environmental quality. For instance, when a more efficient technology generates more hazardous waste despite requiring fewer material inputs. Second, increased resource efficiency in production could lead to a rebound since reduced costs translate into higher outputs and lower prices, which in turn increase demand. Although this benefits the economy, it partially or completely balances out the original material savings.

Deiab, M. et al.,[15] classified the energy saving techniques in computing environment called Dynamic Power Management (DPM) and Static Power Management (SPM), respectively. Since SPM techniques are related to hardware level efficiency, low power consumption circuit designing is an example of an SPM technique. SPM and DPM are categorically distinct from one another because SPM are more energy efficient at the single system level.

Aryan Gupta et al.[18] discussed on multiple technique that are used for minimizing energy consumption in cloud. They discussed about DVFS, VM Consolidation, scheduling and Green cloud computing. Dynamic Voltage and Frequency Scaling (DVFS):DVFS can tackle other problems including temperature, dependability, and variability. It is used to scale voltage and frequency for energy and power savings. Virtual machine: A web browser or a remote access tool is frequently used to manage VMs. Consolidating virtual machines (VMs) enhances data centers' resource and energy usage. Additionally, it aids in limiting the expansion of data centers and server load increase. Scheduling: Scheduling determines which resources should be distributed to the newly admitted VMs given a group of VMs and the available resources. Green cloud computing aims to lessen trash dumped into the environment and energy consumption. Cost-savings are another benefit of being green. The authors did an analysis on various algorithms including green computing technology and discussed about which algorithm give the less environment impact and which one consumes more energy.

Akhter, N. and Othman, M.[19] presented a two fold approach for VM allocation. The initial steps clarify by placing VMs on hosts and also new request will be addressed and it will be admitted effectively. The next step concentrate mostly about the existing virtual machine and the allocation where it has to be allocated and also specifies a common algorithm for virtual machine allocation. The steps which have been taken for the host which is overloaded will be able to find a position which is new for the host for migration. The next fore coming steps is to find the placement of virtual machine and a host which is under will load effectively with the listed hosts. The complexity in methodology and algorithm makes the work load allotment improperly. The proposed wok will be able to increase the efficiency during the load in a effective manner. Occasionally we will get annual power consumption cost as higher when compared to infrastructure installation costs.

Wang, W., et al., [20] discusses about how providers allocate the resources efficiently to the machine which are in physical state. Their objective is to minimizing the energy consumption. They used Multi-Agent based technique in centralized order and decentralized multi-agent based Virtual machine which is properly allocated. The method which is proposed for different techniques will assist the physical machines for properly managing and utilizing resources. The new methods and different allocation strategy helps the physical machine in different stages. Auction-based techniques and methodology will be able to decide the allocated agents and the physical machines which is submitted. Limitations of their work includes the authors main focus is on allocating the Virtual machine to Physical Machines but our main work is to concentrate how to reduce and minimize the energy and to be cost effective to the users.

Garg, N et al.,[21] compares various task Scheduling algorithm and perform an analysis on which one provides better understanding about the various methodologies which will reduce the energy effectively and also help to decrease the waiting time for the user and also to reduce the time for tasks. Max-Min and Min-Min algorithm used for security enhancing using some methods of encryption. MAX-MIN will compute all the tasks and it will take and compute the maximum task and it will be paired with the task have minimum time along with this more task can be performed simultaneously. Limitation of their work includes that it will not not give more accuracy of the result.

Katal, A et al., analysis different techniques for energy

efficiency for the data centers and also will compare these approaches with other existing works. They compare various energy conservation techniques using DC,DVFS. Data centers needs these steps: how to identify the effective cooling system and the green technology for reducing the wastage of resources in a effective manner. It has been identified in the cloud computing architecture some components will minimize energy and act as a energy saver.

Kumar, J., et al., [23] presents a novel load-balancing framework with reducing the working expenditure of the data center through advanced utilization methodology of using the resources. The advanced technology opted will use genetic approach of a framed algorithm for allocating the virtual machine over the identified physical machine with optimization. This technique which is utilized will be helpful for advancing and developing the allocation to obtain the solutions. The Algorithm which shows some of the implementations of pseudo code of the specified approach, where there is a population (ϵ) of the n random with the solutions is first properly initialized. A part of the solution put the i^{th} VM at randomly which have been selected j^{th} PM that is now active and satisfies the constraints which are listed. According to that the constraints are helped to summarize that no VM is been allocated to a server if there are sufficient resources are unavailable.

Bermejo, B., et al.,[24] proposed a methods which have two action oriented levels which is used to optimize the way of which it deals the system with energy efficiency. Another level the Physical machine also called the lowest level in which the machine can be able to host a numbers of VM with each having a different level of customer. The VM will not share retrieved information which is totally independent. The next level Controller the higher level will be acting as a global controller. This way of work will be in charge for accessing all the requests which it receives from the machine and also it will order more number of actions. Limitations of this work show that the proposed methodology will be higher than utilization-based technique. The result will not show the controller level output. it only show the working of local level.

Goyal, S., et al., [25] proposed an framework for energy conserve and also for allocating resources in a cloud with optimization based on the algorithm Whale. The various problems is identified to find the Optimized approach and the Framework for the Energy-Resource Allocation in a Cloud. Energy used in a data center also produces more threat to the environmental issues. So to overcome the problem occurred in the environment they used the algorithm optimization and algorithm PSO. This algorithm which is identified will be applied mainly in two phases of which, the first phase deals by limiting the violations of the SLA and the other phase relocates the VMs and making them to merge them in PMs which are active and also used load balancing technique. Limitations of their work are they have not done comparison analysis of the algorithm they have just shown how the algorithm works.

3 Observations

The algorithm what we use for predicting the energy efficiency and resource utilization shouldn't be done for a simplified version of cloud environment. Some algorithms do not adapt allocation of VMs on runtime and are not independent of the workload type. Similarly QoS shouldn't be compromised for energy efficiency. Better algorithms than LSTM, exist so we need to check with every algorithms. Our objective is to predict the time series with that insight we shouldn't improve based on our intuition there should be a defined process. The RNN is a powerful model for processing the sequences, it performs poorly when faced with lengthy sequences and loses its learning ability so a better algorithm than that should be identified. It requires complex computations, and the training phase is also lengthy as well.

4. Conclusion

In this article, we explored multiple algorithms. for predicting energy efficiency and efficient resource utilization and came to a conclusion that the need for efficient prediction algorithms for energy efficiency and resource utilization plays an important part in the goal of using resources as efficiently as possible. To get the intended results, we carefully considered different algorithms and how they operated in a cloud computing context. These algorithms have been successfully applied in cloud environments and have been enhanced in a better way to be suitable for application in existing cloud environments. Numerous studies have been conducted in an effort to find the optimum algorithm for estimating energy efficiency and resource usage, and further research is required.

References

[1] Koronen, C., Åhman, M. and Nilsson, L.J., 2020. Data centres in future European energy systems—energy efficiency, integration and policy. Energy Efficiency, 13(1), pp.129-144.

[2] Beloglazov, A. and Buyya, R., 2010, May. Energy efficient resource management in virtualized cloud data centers. In 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (pp. 826-831). IEEE.

[3] Kumar, K.S. and Jaisankar, N., 2017. Towards data centre resource scheduling via hybrid cuckoo search algorithm in multi-cloud environment. International

Journal of Intelligent Enterprise, 4(1-2), pp.21-35.

[4] Menaka, M. and Kumar, K.S., 2022. Workflow scheduling in cloud environment–Challenges, tools, limitations & methodologies: A review. Measurement: Sensors, p.100436.

[5] Yazdanian, P. and Sharifian, S., 2018, December. Cloud Workload Prediction Using ConvNet And Stacked LSTM. In 2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) (pp. 83-87). IEEE.

[6] Zhang, W., Li, B., Zhao, D., Gong, F. and Lu, Q., 2016, October. Workload prediction for cloud cluster using a recurrent neural network. In 2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI) (pp. 104-109). IEEE.

[7] Kumar, J. and Singh, A.K., 2016, December. Dynamic resource scaling in cloud using neural network and black hole algorithm. In 2016 Fifth International Conference on Eco-friendly Computing and Communication Systems (ICECCS) (pp. 63-67). IEEE.

[8] Wang, X., Wang, X., Wang, C.L., Li, K. and Huang, M., 2014, December. Resource allocation in cloud environment: a model based on double multi-attribute auction mechanism. In 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (pp. 599-604). IEEE.

[9] Kumar, K.S., Anbarasi, M., Shanmugam, G.S. and Shankar, A., 2020, January. Efficient predictive model for utilization of computing resources using machine learning techniques. In 2020 10th International conference on cloud computing, data science & engineering (confluence) (pp. 351-357). IEEE.

[10] Li, Y., Tang, X. and Cai, W., 2015. Dynamic bin packing for on-demand cloud resource allocation. IEEE Transactions on Parallel and Distributed Systems, 27(1), pp.157-170.

[11] Viswanathan, R. and Devi, C.B., 2017, June. A framework using job sequencing algorithm for efficient dynamic resource management in cloud. In 2017 International Conference on Computational Intelligence in Data Science (ICCIDS) (pp. 1-4). IEEE.

[12] Gunasekaran, J.R., 2020, December. Minimizing Cost and Maximizing Performance for Cloud Platforms. In Proceedings of the 21st International Middleware Conference Doctoral Symposium (pp. 29-34).

[13] Van Ewijk, S., 2018. Resource efficiency and the circular economy: concepts, economic benefits, barriers, and policies. UCL Institute for Sustainable Resources: London, UK.

[14] Deiab, M., El-Menshawy, D., El-Abd, S., Mostafa, A. and Abou El-Seoud, M.S., 2019. Energy Efficiency in Cloud Computing. International Journal of Machine Learning and Computing, 9(1), pp.98-102.

[15] Akhter, N. and Othman, M., 2016. Energy aware resource allocation of cloud data center: review and open issues. Cluster computing, 19(3), pp.1163-1182.

[16] Wang, W., Jiang, Y. and Wu, W., 2016. Multiagent-based resource allocation for energy minimization in cloud computing systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47(2), pp.205-220.

[17] Garg, N., Singh, D. and Goraya, M.S., 2021. Energy and resource efficient workflow scheduling in a virtualized cloud environment. Cluster Computing, 24(2), pp.767-797.

[18] Katal, A., Dahiya, S. and Choudhury, T., 2022. Energy efficiency in cloud computing data centers: a survey on software technologies. Cluster Computing, pp.1-31.

[19] Kumar, J., Singh, A.K. and Mohan, A., 2021. Resource-efficient load-balancing framework for cloud data center networks. ETRI Journal, 43(1), pp.53-63.

[20] Bermejo, B., Guerrero, C., Lera, I. and Juiz, C., 2016. Cloud resource management to improve energy efficiency based on local nodes optimizations. Procedia Computer Science, 83, pp.878-885.

[21] Goyal, S., Bhushan, S., Kumar, Y., Rana, A.U.H.S., Bhutta, M.R., Ijaz, M.F. and Son, Y., 2021. An optimized framework for energy-resource allocation in a cloud environment based on the whale optimization algorithm. Sensors, 21(5), p.1583.