Computer Science and Engineering 2020, 10(1): 1-6
DOI: 10.5923/j.computer.20201001.01

Resource Utilization Prediction Model for Efficient Dynamic Virtual Machine Consolidation in Cloud

Muhammad I. Umar[1,*], Kabiru I. Musa[1], Musa Nehemiah[1], Muhammad Aliyu[2], Rumana K. Aminu[3], Nahuru A. Sabongari[1,4], Mohammed K. Dauda[1]

[1]Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria
[2]Department of Computer Science, Federal Polytechnic, Bauchi, Nigeria
[3]Department of Computer Science, F.C.T College of Education Zuba Abuja, Nigeria
[4]Department of Mathematical Sciences/Department of Computer Science Kaduna Polytechnic, Nigeria

Abstract Virtual Machine (VM) consolidation is an optimization approach for VM placement in cloud infrastructure, which is one of the effective ways to efficiently utilize cloud resources in order to optimize number of VM migrations, Service Level Agreement (SLA) violations and energy consumption problems. Static VM consolidation strategies have not been effective in handling variation of workloads in the cloud systems. Dynamic VM strategies have been proposed in the literatures. However, in the dynamic VM consolidation process, most existing algorithms consolidate active servers mainly based on the current resource requirements and forgo the demands of resources in the future during VM allocation. Thus, they result in needless VM migrations and cause high rate of SLA violations in data centers. This research proposed a prediction-based VM consolidation approach. The proposed method utilized a multi-resource utilization to predict the current and future CPU and memory utilization of active servers during allocation stage. The proposed method was compared with the existing one that did not consider future utilization of resources. CloudSim toolkit 3.0 simulator was used to implement and evaluate the performance of the algorithms. The proposed algorithm was found to have improved in terms of number of VM migrations and SLA violations. Using the number of hosts (400, 600, and 800), the proposed method was 28%, 32% and 38% decrease in number of VM migrations and 0.03243%, 0.03247% and 0.01474% decrease in SLA violations compared with the existing method, respectively. Recommendations and future directions were suggested for further improvements.

Keywords Virtual Machine consolidation, Virtual Machine Migration, Service Level Agreement, Cloud Computing

1. Introduction

Cloud computing is an internet-based computing technology that provision shared pool of virtual computing resources transparently, on-demand and on a pay-per-use model. These virtual resources include CPU, operating platforms, storage, memory, network bandwidth etc. (Sharma & Saini, 2016; Patel & Patel, 2017). These resources are created by multiplexing the physical servers in the data centers and are shared to multiple users across the globe transparently and in isolation from each other through virtualization technology (Pietri & Sakellariou, 2016).

For efficient resource management, modern cloud data centers seize the advantage of virtualization to reduce cloud computational cost and energy budgets (Ahmad, et al., 2015). However, poor utilization of these resources can lead to a lot of problems in the cloud infrastructure, such as SLA violations and VM migration problems. Under-utilization of resources is the major source of over consumption of energy in data centers, which result in poor server utilization patterns (Jain, et al., 2018). Virtualization is a powerful technology that facilitates better use of the available data center resources using a technique called VM consolidation which involves packing up of several VMs unto a single or few physical servers to reduce energy consumption and improve resource utilization in cloud data centers (Abdelsamea, et al., 2014; Damodar & Koli, 2015 and Wang & Tianfield, 2018). However, in the VM consolidation, early models, using static threshold values for host selection during the VM consolidation process have proved ineffective to handle the variation of workload in the cloud infrastructure (Sharma & Saini, 2016), therefore, recent works have more focused on the heuristics that can determine the upper and lower threshold for dynamic workloads based on the change in CPU utilization of the host server.

* Corresponding author:
ibrahimmumar40@gmail.com (Muhammad I. Umar)
Published online at http://journal.sapub.org/computer

However, Sajitha and Subhajini (2018) opined that RAM and bandwidth are also critical resources to be considered in computing load as opposed to most existing works that dwelled on CPU usage only. More so, in a dynamic VM consolidation process (Fig. 1) Abdelsamea et al (2014), most literatures based on only the current resource requirements of target host and neglected the future utilization demands during the VM allocation stage. As a result, they produce needless VM migrations (which may lead to more energy consumption in the data center) and increase the rate of SLA violations in data centers (Farahnakian et al., 2016).

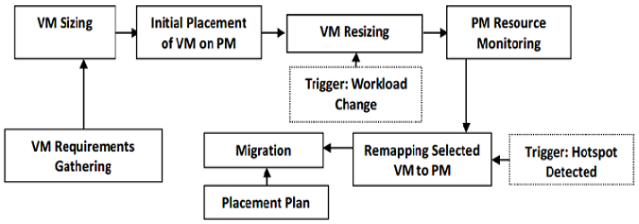

Figure 1. Dynamic Virtual Machine Consolidation Process

In this study, the VM allocation is enhanced by taking into account both the current and future utilization of resources, where a regression-based model is used to approximate the future CPU and memory utilization of hosts before allocation of VMs would take place. This is aimed at improving the rate of SLA violations and VMs migrations which may alleviate energy consumption in the cloud data centers.

The remaining sections of this paper are structured as follows: Section 2 presents VM Consolidation framework, section 3 discusses related works; section 4 presents the proposed method, section 5 presents results and discussion , finally conclusions and future directions are presented in section 6.

2. VM Consolidation Framework

In the business of cloud computing, the Cloud Service Providers (CSPs) have to work hard to keep their customers (cloud users) by ensuring good Quality of Service (QoS) and meeting up with SLA agreements while at the same time improving resource utilization in the cloud infrastructure so as to obtain quick Return On Investment (ROI). VM consolidation is an efficient strategy utilized by CSPs to efficiently manage cloud resources in order to alleviate energy consumption, and maintain an uncompromising SLA agreement. Virtualization technology facilitates in reducing power consumption in data centers by creating multiple VMs onto a server and implementing the process of virtual machine consolidation (VMC) (Alboaneen, Pranggono & Tianfield, 2014; Chang, Gu & Luo, 2017). This is achieved by packing VMs unto fewer servers, ensuring optimum utilization while unused servers are shut down or hibernated to reduce energy consumption. Idle hosts are considered among the most energy consumers in the cloud data centers (Ghobaei-Arani, et al., 2018). VM consolidation is categorized into two variants: Static and Dynamic (Ferdaus & Murshed, 2014). To achieve workload consolidation, hosts are categorized based on their utilization, namely over-loaded hosts, under loaded host and normal or idle hosts (Patel & Patel, 2017; Alboaneen, Pranggono & Tianfield, 2014). The efficient identification of over-loaded and under-loaded hosts is a key step in Dynamic VM consolidation (Motwani, et al., 2016). Dynamic consolidation process basically has the following four steps;

Host Underloaded Detection: This is the problem of deciding which host is Underloaded and requiring migrating VM. Here, all VMs are migrated and the source host are switched off or hibernated. This minimizes energy consumption as the source as the number of active servers will be minimized.

Host Over-loaded Detection: This is the decision of which host is over-loaded and needing VM migration. This will require migrating one or more VMs to other active host or re-activated host. This makes the hosts balanced in order to satisfy QoS such as SLA.

VM Selection: This is the stage when a decision has been established to migrate VM or VMs from cold spot or hot spot. This determines which VMs will be considered candidates for migration.

VM Migration: This is the actual stage of VMs migration. The migration is done considering the service downtime and resource consumption during the migration process.

3. Related Works

Damodar & Koli (2015) proposed SLA-aware algorithm which finds and decides over-loaded host with SLA violation and proposed an efficient algorithm for finding underutilized host. They combined these algorithms to achieve energy performance trade-off. The overload detection finds overload host and get the status of ooverload host whether it result in SLA violation or not. However, the study did not consider when a host is in normal state. Alboaneen, Pranggono and Tianfield (2014) proposed a new scheme of host's load categorization in VM consolidation framework in cloud based data centers to reduce energy consumption while meeting Quality of Service (QoS) requirements. They classified the Underloaded hosts into three further states, i.e., Underloaded, normal and critical by applying underload detection algorithms. Sharma and Saini (2016) proposed a novel method for consolidation of VMs such that it meets SLA and deals with energy-performance trade-off. For the allocation and reallocation of virtual resources depending upon their load, they used a threshold based approach, in which Median method is used to find lower and upper threshold values. An 'Adaptive Load Detection Technique' to detect the over-loaded and under-loaded hosts was proposed by Motwani, et al., (2016), in this work they used a famous statistical method: InterQuartile Deviation (IQD) to adaptively detect load. The proposed method obtained a better optimization in energy efficiency and performance.

Chang, Gu and Luo (2017) proposed a novel VM selection policy and a resource aware utility model to guide the VM migration process. Kaur, Diwakar and Vashisht (2017) presented some alternative robust techniques to overload detection for deciding an adaptive threshold of CPU utilization and compare them with existing techniques in CloudSim simulator. The authors tried to optimize host overload detection. However, they use only CPU utilization, neglecting other critical resources parameters such as memory and network bandwidth. Shaw, Kumar and Singh (2017) proposed a novel approach of adding a constraint to the existing VM consolidation technique to avoid unnecessary VM migration. They also proposed heuristics for VM selection algorithm. A dynamic Algorithm, which considers multiple factors such as CPU, memory and bandwidth utilization of the node for empowering VM consolidation by using regression analysis model, was proposed by Sajitha and Subhajini (2018). The authors presented Energy Conscious Dynamic VM Consolidation with auto adjustment of three threshold values such as upper threshold, middle (prone-to-upper) and lower threshold. However, Network resource utilization and traffic of data center was not considered for VM placement and future resource utilization was not considered during allocation stage. Wang and Tianfield (2018) proposed a new VM placement policy, namely Space Aware Best Fit Decreasing (SABFD) and a new migration VM selection policy, namely high CPU utilization based migration VM selection (called HS). However, SABFD used only the current CPU utilization to perform consolidation.

Most existing literatures, utilizing threshold-based VM consolidation strategy are mainly focused on single CPU utilization. However, power consumption by a server in data center is connected with its processor, RAM, hard disk, and bandwidth (Sajitha & Subhajini, 2018). Furthermore, most literatures considered only the current resource requirements of destination host and neglected the future utilization during the VM allocation. As a result, they generate unnecessary VM migrations (which can lead to more energy consumption in the data center) and increase the rate of SLA violations in data centers.

4. The Proposed Method

To enhance the existing work, the proposed work considered the future utilization of both CPU and memory of host before allocating VMs. This will lead to reduced number of VM migrations and SLA violations, subsequently reducing the energy consumption in the data center. When a VM is considered for migration, it is pertinent to locate an appropriate host based on VM characteristics and overall policy pursued in the data center (Sajitha & Subhajini, 2018). The proposed approach takes into account both the current and future utilization of resources, where a regression-based model is used to approximate the future CPU and memory utilization of VMs and hosts.

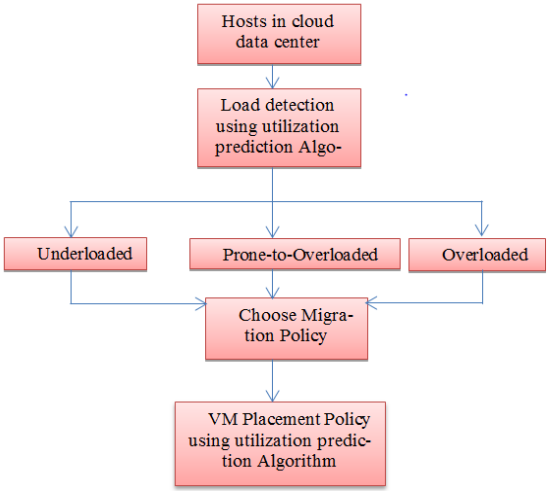

Figure 2. The Proposed Resource Prediction based DVMC Flowchart

Figure 2 depicts the proposed resource prediction based Dynamic Virtual Machine Consolidation (DVMC) proposed in this study. This proposed approach, however, add another detection of hosts called prone-to-over-loaded. This is included to ensure hosts prone to over-loaded are assigned loads (VMs) only when they will not become over-loaded. A migration could take place only if the destination host has enough CPU and memory resources to accommodate the candidate VM at the moment and in the future time. In order to predict the resource utilization, a regression-based prediction model was used, K-Nearest Neighbor Regression (K-NNR). The prediction model estimates the resource utilization of VMs and hosts based on CPU and memory.

The proposed_DVMC algorithm focuses on migrating VMs from the over-loaded and predicted over-loaded hosts. The algorithm checks periodically for the total capacity of each resource (i.e. CPU or memory). The algorithm considered a 2-dimentional capacity vector of CPU and Memory only. If any exceeds the total capacity required, the hosted physical machine is considered over-loaded or predicted over-loaded host. For example, the used CPU capacity of a host is estimated as the sum of the CPU utilization of the five VMs if five VMs are hosted by the same host (Physical machine). The proposed_DVMC moves some VMs from the over-loaded or predicted over-loaded hosts repeatedly until there are no over-loaded hosts left. The overall goal of this is to move some VMs from over-loaded and predicted over-loaded hosts with the aim to reduce the number of VM migrations and minimize SLA violations. These SLAs are critical to guarantee QoS satisfaction to the cloud users.

5. Results and Discussion

5.1. Experimental Setup

This section presents the simulation setup, VM migration policy and experimental results and discusses the results. CloudSim toolkit 3.0 was used to simulate the different

algorithms, using NetBeans IDE 8.2. The toolkit has been developed by the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, University of Melbourne. It supports both system and behavior modeling of cloud system components such as data centers, Cloud Information Service (CIS), broker, VMs and resource provisioning policies (Damodar & Koli, 2015).

5.2. Simulation Setup

Table 1 and Table 2 contain the hosts parameters and VM parameters respectively. The experiment is conducted in three scenarios, using 400, 600 and 800 number of hosts each time in other to enable the algorithms to be analyzed and evaluated based on *rate SLA violations* and *number of migration* performance metrics.

Table 1. Hosts Parameters

Parameter	Value
Number of Host	400, 600, 800
Number of Cores	2
MIPS	2000
RAM	4096(MB)
Bandwidth	1GB
Storage	1GB

Table 2. VMs Parameters

Parameter	Value
Number of cores	1
MIPS	2500
RAM	1740 (MB)
Bandwidth	1(MB)
Storage	2.5GB

5.3. VM Selection Algorithm

To evaluate the performance and robustness of the proposed_DVMC and Existing_DVMC algorithms, a single VM selection policy was used across the algorithms. Minimum Migration Time (MMT) was selected as the migration policy. MMT involves selecting a VM with the minimum migration time. The length of a VM migration takes as long as it needs to migrate the memory assigned to the VM over the network bandwidth link between source and destination PMs.

5.4. Performance Metrics

The proposed approach aimed at guaranteeing that SLAs are not violated; and minimize the number of migrations. Thus, the performance of proposed algorithm is assessed through the following metrics:

- **SLA Violations:** This measured the SLA violations due to over-utilization of resources. This indicates the percentage of time, during which active PMs have experienced the CPU or memory utilization of 100%.
- **Number of VM Migration:** This involves the task of moving a VM from one host environment to another.

Table 3 shows the results of the experiment after several rerun simulations and the average of the results are computed and tabulated. In each of the scenarios of number hosts, the proposed_DVMC shows improvement over the existing_DVMC. The number of migration of an algorithm is improved when the value is less than the value of the existing algorithm. Likewise, the percentages of SLA violations also have a noticeable improvement in the proposed_DVMC compared with the existing approach. The lower the percentage of SLA violations values, the better the algorithm.

Table 3. The Results of the Executed Experiments

Algorithm	No.VM migration	SLA (%)
	400 Hosts	
Proposed_DVMC	3650	0.05832%
Existing _DVMC	6384	0.09075%
600 Hosts		
Proposed_DVMC	4003	0.06778%
Existing_DVMC	7615	0.10025%
800 Hosts		
Proposed_DVMC	4401	0.06990%
Existing_DVMC	9611	0.08464%

Figure 3 shows the VM migration vs. number of hosts. In each of the scenarios, the proposed_DVMC produce less VM migrations. This improves the performance of the system. The generation of unnecessary VM migrations can lead to more energy consumption in the data center and degradation of the whole performance of the system. The proposed method improves the unnecessary VM migration by predicting the future CPU and Memory utilization of destination host and thus migrate VMs only when it is necessary. The migration does not occur when it is known the destination hosts will soon be over-loaded, leading to another VM migration to be triggered. The proposed method using the prediction model is able to predict the future hosts CPU and memory utilization and thereby allocates VMs to only appropriate hosts.

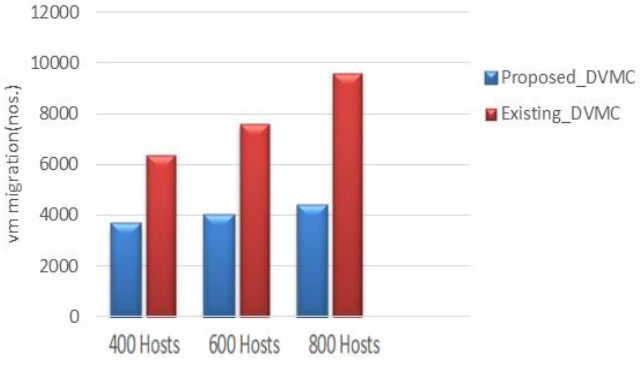

Figure 3. VM Migration Vs. No of Hosts

Figure 4 shows the graph of SLA (%) vs. number of hosts for the three scenarios. The proposed_DVMC algorithm generates less SLA violations which imply a significant improvement over the existing_DVMC algorithm.

Service level agreement is a legal document that both the cloud providers and cloud customers signed specifying the agreed level of QoS services to be provided by the cloud CSPs. Any violations to these may cost the CSPs a great deal of lost. When a host experiences 100% utilization, it will not be able to allocate enough CPU to the VMs on it, so it will generate a number of SLA violations.

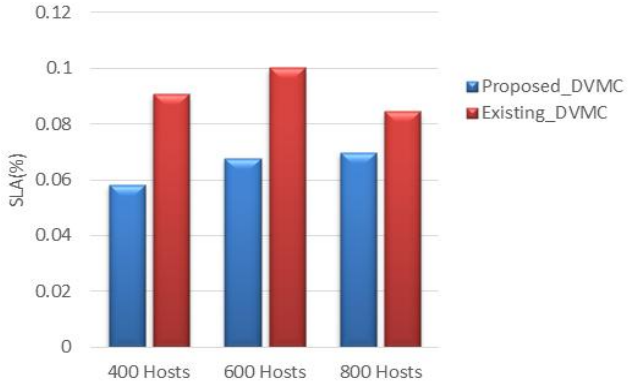

Figure 4. SLA violations vs. Number of Hosts

In scenario one (400 hosts), the number of VM migration of the proposed_DVMC algorithm is 3650 while the existing_DVMC algorithm is 6384, giving 28% decrease in number of VM migration. In scenario two (600 hosts), the number of VM migration for the proposed_DVMC algorithm is 4003 and that of the existing_DVMC algorithm is 7615, leading to 32% decrease in number of VM migration. Lastly, the third scenario (800 hosts) decreased the number of VM migration by 38% where the numbers of migrations of the proposed and existing_DVMC algorithms are 4401 and 9611 respectively.

It can also be deduced that the performance of the proposed algorithm increases over the existing algorithm as the number of hosts increases. The decrease of the number of VM migration could be as a result of predicting the future resource utilization of the destination hosts which helped in determining the appropriate hosts to move VMs unto that may not cause another migration in the host in the near future. This reduces the frequent migrations in the system.

For the SLA violations, the result for the first scenario for both the proposed and existing_DVMC algorithms are 0.05832% and 0.09075% respectively which gives 0.03243% decrease. Also, the SLA violations in the second scenario are 0.06778 and 0.10025% for the proposed and existing_DVMC algorithms respectively which results to 0.03247% decrease. In the third scenario, the SLA violation in the proposed_DVMC algorithm is 0.06990% while the existing_DVMC algorithm is 0.08464, giving a decrease of 0.01474%. From these results, it can be deduced that the proposed_DVMC algorithm, generates less number of SLA violations. This especially helps the CSPs to reduce the cost of dealing with SLA violations issues.

These results could be explained as a result of the efficient prediction of host resource utilization by the prediction model in the proposed algorithm which helps in avoiding 100% resource utilization of the hosts.

6. Conclusions and Future Directions

Cloud computing is a computing paradigm that comes with benefits such as low maintenance of infrastructure, up-front costs and ease of scaling for the users. However, it also comes with issues such as resource utilization, energy consumption, VM migration and service level agreement (SLA) violations among others. VM consolidation (VMC) has been utilized to address these issues. However, most literatures, utilizing threshold-based VMC strategy are mainly focused on single resource (CPU) utilization. Furthermore, most literatures considered only the current resource requirements of destination host and neglected the future utilization during the VM allocation stage. As a result, they generate needless VM migrations (which can lead to more energy consumption in the data center) and increase the rate of SLA violations in data centers. This study proposed a new method that utilized CPU and memory as well as the future utilization of these resources on the hosts during the VM placement. The proposed method utilized prediction model based on two regression-based prediction models: Linear Regression (LR) and K-Nearest Neighbor Regression (K-NNR). This proposed method helped in detecting both the current and future resource utilization on the hosts before placing VMs unto them. Experiments were carried out; the proposed method produced better results in terms of number of VM migrations and SLA violations compared with the existing method without future prediction consideration. Although, reducing number of VM migrations would lead to reduction in energy consumption in the data center, there is still need to measure energy consumption in terms of idle hosts and other factors.

In the future, the researchers intend to explore more number of resources such as hard disk and bandwidth for the prediction model. Also, network utilization and traffic will be considered in order to reduce the migration cost and scalability of the proposed model.

ACKNOWLEDGEMENTS

Authors are grateful to the reviewers of this manuscript for their expert advice. Authors are thankful to Department of Mathematical Sciences at Abubakar Tafawa Balewa University, Bauchi, Nigeria for support.

REFERENCES

[1] Sharma, O. & Saini, H. (2016). VM Consolidation for Cloud Data CentRe using Median based Threshold Approach. Procedia Computer Science 89 (2016) 27-33. DOI: 10.1016/j.procs.2016.06.005.

[2] Patel, N. & Patel, H. (2017). Energy Efficient Strategy for Placement of Virtual Machines Selected from Underloaded Servers in Computer Cloud. *Journal of King Saud University-Computer and Information Sciences (2017)*, DOI: https://doi.org/10.1016/j.jksuci.2017.11.003.

[3] Pietri, L. & Sakellariou, R. (2016). Mapping Virtual Machines onto Physical Machines in Cloud Computing: A Survey. *ACM Computing Surveys*, 49(3). DOI: http://dx.doi.org/10.1145/2983575.

[4] Ahmad, R.W., Gani, A., Ab Hamid, S.H., Shiiraz, M., Yousafzai, A. & Xia, F. (2015). A Survey on Virtual Machine Migration and Server Consolidation Techniques for Cloud Data Centers. *Journal of Network and Computer Applications*, 1-15. DOI: http://dx.doi.org/10.1016/j.jnca.2015.02.002.

[5] Jain, N. & Joshi, K.K. (2015). A Survey of VM Allocation and Migration Algorithms for Energy Efficient Data Center. *International Journal of Innovative Research in Technology*, 2(6), 431-435.

[6] Abdelsamea, A., Hemayed, E.E., Eldeeb, H. & Elazhary, H. (2014). Virtual Machine Consolidation Challenges: A Review. International Journal of Innovation and Applied Studies, 8(4), 1504-1516.

[7] Damodar & Koli, 2015. Performance analysis of an energy efficient Virtual machine consolidation algorithm in Cloud computing. International Journal of Computer Engineering and Technology (IJCET), 6(5), 24-35.

[8] Wang, H. & Tianfield, H. (2018). Energy-aware Dynamic Virtual Machine Consolidation for Cloud Datacenters. IEEE.

[9] Sajitha, A.V., & Subhajini, A.C. (2018). Dynamic VM Consolidation Enhancement for Designing and Evaluation of Energy Efficiency in Green Data Centers Using Regression Analysis. International Journal of Engineering & Technology, 7 (3.6) (2018) 179-186.

[10] Farahnakian, F., Pahikkala, T., Liljeberg, P., Plosila, J. Hieu, N.T. & Tenhunen, H. (2016). Energy-aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model. *IEEE Transaction on Cloud Computing. DOI 10.1109/TCC.2016.2617374*.

[11] Alboaneen, D.A., Pranggono, B. & Tianfield, H. (2014). Energy-aware Virtual Machine Consolidation for Cloud Data Centers. 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. IEEE.

[12] Chang, Y., Gu.C. & Luo, F. (2017). Energy Efficient Virtual Machine Consolidation in Cloud Datacenters. *The 2017 4th International Conference on Systems and Informatics (ICSAI 2017)*. IEEE, 401-406.

[13] Ghobaei-Arani, M., Rahmanian, A.A., Shamsi, M. & Rasouli-Kenari (2018). A learning-based approach for virtual machine placement in cloud data centers. Int J Commun Syst. 2018, 1-18. DOI: https://doi.org/10.1002/dac.3537.

[14] Ferdaus, M. H. & Murshed, M. (2014). Chapter 8 Energy-Aware Virtual Machine Consolidation in IaaS Cloud Computing, 179-208.

[15] Motwani, A., Rafique, S.Q., Singh, P.N. & Sondhi, J. (2016). Adaptive Load Detection Technique for Effective Virtual Machine Consolidation. *International Journal of Electrical, Electronics and Computer Engineering*, 5(1).

[16] Kaur, A., Diwakar, A. & Vashisht, R. (2017). Alternatives to VM consolidation techniques for energy aware cloud computing. IEEE.

[17] Shaw, S.B., Kumar, J.P. Singh, A.K. (2017). Energy-Performance Trade-off through Restricted Virtual Machine Consolidation in Cloud Data Center. *2017 International Conference on Intelligent Computingand Control (I2C2).*