ITC 2/50

Information Technology and Control
Vol. 50 / No. 2 / 2021
pp. 332-341
DOI 10.5755/j01.itc.50.2.28056

Energy-Aware and Proactive Host Load Detection in Virtual Machine Consolidation

Received 2020/11/23 | Accepted after revision 2021/05/07

http://dx.doi.org/10.5755/j01.itc.50.2.28056

HOW TO CITE: Fard, S. Y. Z., Sohrabi, M. K., Ghods, V. (2021). Energy-Aware and Proactive Host Load Detection in Virtual Machine Consolidation. *Information Technology and Control,* 50(2), 332-341. https://doi.org/10.5755/j01.itc.50.2.28056

Energy-Aware and Proactive Host Load Detection in Virtual Machine Consolidation

Seyed Yahya Zahedi Fard

Department of Computer Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran
e-mail: Yahya.zahedi@gmail.com

Mohammad Karim Sohrabi

Department of Computer Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran
e-mail: Amir_sohraby@aut.ac.ir

Vahid Ghods

Department of Electrical Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran
e-mail: v.ghods@semnaniau.ac.ir

Corresponding author: Amir_sohraby@aut.ac.ir

With the expansion and enhancement of cloud data centers in recent years, increasing the energy consumption and the costs of the users have become the major concerns in the cloud research area. Service quality parameters should be guaranteed to meet the demands of the users of the cloud, to support cloud service providers, and to reduce the energy consumption of the data centers. Therefore, the data center's resources must be managed efficiently to improve energy utilization. Using the virtual machine (VM) consolidation technique is an important approach to enhance energy utilization in cloud computing. Since users generally do not use all the power of a VM, the VM consolidation technique on the physical server improves the energy consumption and resource efficiency of the physical server, and thus improves the quality of service (QoS). In this article, a server threshold prediction method is proposed that focuses on the server overload and server underload detection to improve server utilization and to reduce the number of VM migrations, which consequently improves the VM's QoS. Since the VM integration problem is very complex, the exponential smoothing technique is utilized for predicting server utilization. The results of the experiments show that the proposed method goes beyond existing methods in terms of power efficiency and the number of VM migrations.

KEYWORDS: Virtual machine consolidation, energy consumption, resource management.

1. Introduction

Cloud computing has been raised as one of the most crucial exciting technological developments, which is growing rapidly to provide almost unlimited process, storage, and network resources for its users [13]. The growing interest in cloud resources has led to the creation of enormous cloud data centers, which need a lot of energy, and cause a large amount of operational costs. Worldwide data center power utilization information shows a nonlinear increase over the past ten years and a similar trend is predicted for the future [28]. Cloud providers utilize a lot of data centers around the world [27]. Energy is one of the significant factors of the entire cost of managing a data center and its facilities [18]. Several methods have been proposed in the literature to enhance resource utilization in cloud environments to lead to energy efficiency. As some examples, Load balancing and resource utilization were improved in [6] using a combination of reinforcement learning (RL) and coral reefs optimization. Multiple RL-based agents were utilized in [7] for online scheduling of dependent tasks of the cloud's workflows to improve resource utilization. Cooperative RL agents managed the cloud resources in [8] for multiple online scientific workflows. SARSA RL agents and genetic algorithm were used in [9] for cloud resource utilization.

VM consolidation is a common and useful method to enhance resource efficiency and power effectiveness in the data centers, which has been extensively investigated in recent years. For example, server consolidation frameworks for cloud data centers have been reviewed in [2]. Server consolidation techniques in virtualized data centers have been also reviewed in [33]. A joint VM and container consolidation approach was utilized in [17] for energy-aware resource management in cloud data centers. EMC2 was proposed in [20] as an energy-efficient and multi-resource fairness VM consolidation in cloud data centers. Another energy-efficient VM consolidation algorithm in cloud data centers was also proposed in [38]. VM consolidation leverages by using virtualization as the main technique for achieving better performance and improving energy consumption. Virtualization techniques enable the data centers to deploy and run multiple instances of the VMs on a single physical server, and during the VM lifetime, it is possible to move VMs among physical servers without downtime. To reduce the number of active servers, VMs should be regularly relocated among the consolidation process using live migration according to the resource demands, and to reduce the energy consumption, the idle servers should be switched to sleep or power-off mode [11]. Due to the dynamic resource utilization model of the most current applications because of their highly variable workloads, consolidating VMs in the cloud data centers is complex. Unrestricted consolidation of the VMs may cause performance degradation by increasing the requests of applications for resources. If the requested resource or program of an application is not provisioned, its response time is increased and causes degrading quality of service (QoS) of that application. The QoS is defined for the services of the cloud service provider for the end-users.

The main contributions of this paper are summarized as follows:

- Proposing a server threshold prediction method that focuses on detecting overloaded and underloaded servers.
- Improving server utilization, reducing the number of VM migrations, and enhancing the QoS.
- Utilizing an exponential smoothing technique for predicting server utilization to overcome the complexity of the VM integration problem.

The remainder of the paper is composed as follows: Section 2 discusses some recent related works. Section 3 explains the system model and defines the problem. The proposed method of this paper to predict server utilization is also described in Section 3. Section 4 presents the results of simulation and finally, Section 5 concludes the work.

2. Related Works

There are 4 sub-problems in the VM consolidation procedure which need to be solved [25]. The first sub-problem is to decide when a host is overloaded. The next sub-problem is to select and identify the proper VMs for migrating from an overloaded host. VM placement is the third sub-problem, in which, a proper host for the selected VM would be picked. The last sub-problem is to identify the underloaded servers to consolidate the VMs of these servers on a smaller number of active servers [36]. Depending on the specific problems of the VM consolidation and the perspective of the research, a variety of strategies have been suggested in this area of the literature. A systematic review of challenges and open isuues for server consolidation in virtualized data centers has been proposed in [1].

The data center's workload changes continuously, which affects resource efficiency and energy utilization of the physical servers [5]. Most existing studies determine the server's upper threshold for overloaded hosts according to the host resource utilization. By applying VM consolidation on the overloaded and underloaded hosts, the energy consumption is reduced, and the QoS is improved [19, 35, 39].

VM placement was carried out under optimal conditions in cloud data centers in [16]. A VM placement method, called MBFD, has been proposed in [12] that reduced the number of running hosts based on a bin-packing heuristic algorithm, called BFD (best-fit decreasing) [29]. Another VM placement method was also proposed in [22] for OpenStack Neat consolidation. A fuzzy threshold-based adaptive algorithm has been proposed in [31] for identifying the overloaded and underloaded hosts. The fuzzy inference engine utilized the information of the host (resource utilization) to estimate the numerical threshold of CPU utilization. Witanto et al. [34] focused on accomplishing a method to tradeoff between the QoS and the power consumption based on the system preferences. The concept of dynamic selection was also proposed in [34] by searching into a set of consolidation methods and comparing various impacts on service level agreement (SLA) and energy to accommodate the data center with the desired preferences.

The issue of cooperation among memory and processor was focused in [4], and a multi-input and multi-output controller was introduced for VM consolidation from resource capacity perspectives, while respecting SLA constraints. Dynamic modification of the capacity of the resources of the collection of VMs was considered for serving a multi-tier application. While the proposed method had a higher level of resource consolidation, its computational overhead was also high. The K-means clustering-based adaptive three-threshold framework of [37] divided the hosts into four categories, according to the amount of their loads. The number of VM migrations and the effects of these migrations on the QoS were decreased in the proposed method of [22] by consolidating VMs based on evaluating the possibility of change in the conditions of the resources into the overload state.

A comprehensive research on VM consolidation has been conducted in [10]. Two VM selection policies, namely MMT (minimum migration time) policy and MC (maximum correlation) policy, were proposed in [10]. The MMT policy selected the VM for migration that has the shortest required migration time, and the MC policy selected the VM with the highest correlation of CPU usage with the other VMs for relocation to decrease the probability of overloading the host. Ant colony and extreme learning were utilized in [23] to VM consolidation in cloud data centers. Pearson correlation was used for dynamic VM consolidation in [24]. Another energy-efficient method for dynamic VM consolidation was also proposed in [21]. Some applications of VM consolidation in cloud computing were provided in [40].

The proposed method of this paper exploited a server threshold prediction method to detect overloaded and underloaded servers. This method attempts to overcome the complexity of the VM integration problem by utilizing an exponential smoothing technique for predicting server utilization. Server utilization is improved, and the number of VM migrations is reduced using this method, and thus the QoS will be enhanced.

3. The Proposed Method

Predicting the appropriate times of relocating the VMs is a key factor for improving the utilization of the VM consolidation, which leads to consolidate the VMs on fewer hosts. Various techniques are used in statistical and computational methods for this prediction. Due to the nature of the VM consolidation problem, we leverage the host extended exponential smoothing method [14-15]. Exponential smoothing is a general guideline procedure that uses the exponential window function for smoothing time series data. Although the effects of distant periods are not significant in the predictions, it is nevertheless appropriate that all previous periods are involved in the predictions.

Simple exponential smoothing has been extended to empower the forecasting of data with a trend. This technique includes level and trend smoothing equations (l_t and b_t), used by a forecast equation [3, 14-15]

$$\ell_t = a.y_t+(1-a)(\ell_{t-1} - b_{t-1}) \quad (1)$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)\, b_{t-1} \quad (2)$$

$$forecast\ equation = F_{t+1} = b_t + \ell_t, \quad (3)$$

where ℓ_{t-1} is the estimate of the level in time $t-1$, b_{t-1} is the estimate of the growth rate in time $t-1$, l_t and b_t are respectively the estimates of the level and the trend in time t, α $(0 \le \alpha \le 1)$ and β $(0 \le \beta \le 1)$ are respectively the smoothing parameters for the level and the trend, and y_t is the observation of series in time t. The initial values of the parameters can be based on past statistics or initial knowledge and experience [3]. The parameters α and β can be adjusted by analyzing the results for different values to decrease the difference between real and predicted values.

Using aggressive VM migration leads to extra overhead and energy consumption. The proposed method of this paper aims to prevent unnecessary VM migrations by predicting the status of host resources to improve the QoS and energy consumption. For example, assume 5 VMs with the same specifications are running on a host. The actual efficiency of different VMs at different times is shown in Table 1, and the amount of CPU utilization predicted by the proposed method is shown in Table 2. As shown in Table 2, the predictions are very close to the real numbers, and they will be very helpful in identifying the hosts that are overloaded and underloaded. It is clear that the predictions are not the same as what is happening in reality, and a percentage of errors in predictions is tolerable. The prediction with an acceptable difference is above the real threshold, and the hosts' threshold will not rise above that. To show that this approach can be useful for consolidating the VMs, consider this example in a data center.

Suppose the host threshold is 27%. The host threshold can have different values, and this percentage is just an example to illustrate how the proposed solution works. According to this threshold, at the time 3, the host is considered as an overloaded host, and consequently, some of its VMs must be selected and migrated to the other appropriate hosts. However, if the system can be aware or guess what is going to happen to this host in a short time, it may be able to make a better decision. According to the predictions in Table 2, the resources will be reduced at the time 5. Therefore, the proposed method avoids choosing this host as an overloaded

Table 1
VM CPU utilization

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8	Time 9
VM1	4.92	5.33	5.67	5.75	6.92	5	4.75	5	5.67
VM2	28.83	27.42	31	26.17	21	26.67	27.33	30.50	23.17
VM3	17.08	16.58	17.42	16.25	17.67	16.33	17	15.58	17
VM4	5	5.75	5.75	5.92	5.92	5.75	6.83	5.75	6
VM5	75.33	73.92	75.5	73.5	74.33	76.08	75.83	73.92	74.25
Average	26.23	25.80	27.22	25.52	25.17	26.32	26.18	26.32	25.03

Table 2
VM CPU utilization prediction

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8	Time 9
VM1	5.42	5.8	6.16	6.5	6.76	7.21	7.48	7.27	6.96
VM2	29.33	29.71	29.64	30.53	29.96	28.1	27.7	27.49	28.04
VM3	17.57	17.96	18.1	18.3	18.25	18.42	18.21	18.12	17.67
VM4	5.5	5.88	6.32	6.66	6.93	7.11	7.2	7.23	7.15
VM5	75.83	76.21	76.14	76.37	76.04	75.88	76.11	76.22	75.85
Average	26.73	27.112	27.272	27.672	27.588	27.344	27.34	27.266	27.134

Table 3
Data centers' utilization and their prediction

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8	Time 9
Host1	84	74	84	84	68	55	43	80	75
Host2	85	35	79	82	75	60	15	75	35
Host3	75	75	45	75	60	75	85	60	60
Host4	60	78	60	77	55	60	75	65	45
Host5	55	40	25	40	55	45	15	15	15
DCU	71.8	60.4	58.6	71.6	62.6	59	46.6	59	46
DCU Prediction	72.3	72.5	65.69	60.67	65.76	63.48	68.1	50.49	58.44

host, and allows it to continue without indicating as an overloaded host. Because the proposed method predicts that the host will return to the normal state in a short time. In other words, by predicting the future state of the host from the current state, the proposed method prevents unwanted migrations in the current state, and thus prevents the reduction in the QoS that occurs with the VM migrations.

This method can be also utilized to determine the underloaded hosts. Suppose that all data center hosts are in a normal state and their VMs are fully running without resource shortage. There is only one underloaded host, the utilization of which is lower than the specified threshold. The proposed method tries to relocate VMs from underloaded hosts to the other hosts, and makes the host go sleep state or shutdown to reduce energy consumption and improve efficiency. Whenever the data center needs resources, the proposed method powers on or wakes up the host. Switching on and off or sleeping and waking up the hosts is costly and time-consuming. When the method makes a host go sleep state or shutdown, and a few minutes later, due to the growing demand for VMs resources, the host should be restarted, some overhead costs, such as the reboot time, are imposed to the host. This causes the system to experience a shortage of resources and faces a level of service violation. In this situation, predicting the data center utilization (DCU) can help the method to identify the underloaded hosts, more accurately.

For example, consider a data center that has 5 hosts, which the utilizations of its hosts at different time intervals are shown in Table 3. The utilizations of the hosts are different and change during the time. At the time 3. as shown in Table 3, the utilization of the host 5 is lower than the threshold (30%), and thus the host should be considered as an underloaded host. However, prediction of the future of processor utilization is used for final decision. Since the overall data efficiency of the data center is still above the threshold in predictions, the method ignores them, and lets them continue because there is a possibility that the resources will be needed again. Hence, the host is not determined as an underloaded host. Consider time 7, when the utilization of the host 5 is lower than the threshold, and with looking at the predictions, a relative decrease in the total DCU can be considered. Therefore, this server is specified as an underloaded server if the other hosts have sufficient resources to accept the VMs of this host.

4. Simulation Result

The proposed method is executed on the CloudSim simulation toolkit [26]. The CloudSim toolkit is used to simulate the cloud data centers and their infrastructures. The experiments are carried out under following conditions:

1. The real data from the CoMon project (a monitoring system for Planetlab) are used for the experiments. The data includes the processor utilization data from more than 1,000 VMs located on hosts in more than 500 locations around the world. The data were collected on 10 random days in March and April 2011 [30].
2. The simulation environment includes 300 servers of common types available in data centers, includ-

ing Hp ProLiant DL360 G7, Hp ProLiant DL360 Gen9, and Hp ProLiant Ml110 G5. A random model is used to model the amount of memory and bandwidth, and the processor used the CoMon project workload model. Table 4 shows the specifications of the intended servers. Amazon EC2 VMs are also utilized (from https://aws.amazon.com/ec2/instance-types/), which include 5 different types of VMs. The details of these VMs are summarized in the Table 5.

The proposed method is implemented and its performance is evaluated compared to the other VM consolidation techniques, inter quartile range (IQR) and local regression (LR). These algorithms are evaluated in combination with MMT and MC. IQR applies statistical analysis for dynamic adaptation in the utilization threshold. LR estimates the future resources utilization to adjust the threshold accordingly. The first algorithm, known as LrMmt, which uses the LR algorithm with parameter 1.2 to detect the server overflow, and employs the MMT algorithm to select the VMs. The second algorithm, known as IqrMc, which uses the IQR algorithm with secure parameter 1.5 to detect the finite element, and exploits the MC algorithm to select the VMs. This comparison is based on different criteria including energy consumption, the number of VM migrations, SLA violations, and the number of host shutdowns. Table 6 shows the results obtained in one of the days.

Table 4
Hosts specifications

Host	HP ProLiant ML110 G5	HP ProLiant DL360 G7	HP ProLiant ML110 G9
Processor (Mips)	2300	3067	2660
Core	2	12	36
Memory (GB)	4	16	64
Network Bandwidth (Kbits/s)	10000000	10000000	10000000

Table 5
VM specifications

Virtual Machine	Large	Medium	Small	Micro	Nano
Processor (Mips)	2000	1000	1000	500	250
Memory (MB)	2048	2048	1024	1024	512
Network Bandwidth (Kbits/s)	100000	100000	100000	100000	100000

Table 6
Simulation output for day 03/03/2011

	The proposed Method	LrMmt	IqrMc
Number of hosts	300	300	300
Number of VMs	1052	1052	1052
Simulation time(s)	86400	86400	86400
Energy consumption(kWh)	18.54	19.37	23.13
Number of migrations	3154	4492	4521
SLA (%)	0.00081	0.00103	0.00074
Performance degradation (%)	0.02	0.02	0.01
Number of host shutdowns	322	360	458

Fig. 1 shows the comparison of the LrMmt, the IqrMc, and the proposed method based on the number of VM migrations. The number of the migrations in the proposed method is 3154 migrations, which is 30%, and 31% lower than the number of the migrations in LrMmt and IqrMc, respectively. In the proposed method, the number of migrations has significantly diminished by eliminating relocations from the hosts that are not underloaded. When the method has to decide whether a host is overloaded or not, it predicts the state of the host's resources in the next point based on the amount of resource consumption in the past.

Figure 1
Number of VM migrations

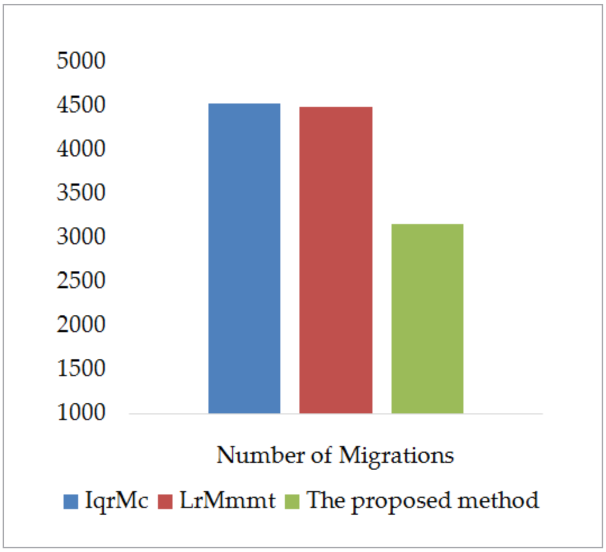


Fig. 2 shows the comparison of the LrMmt, the IqrMc, and the proposed methods based on their energy consumptions. The proposed method is 5% and 20% better than LrMmt and IqrMc, respectively. workload prediction in the proposed method prevents turning on the new hosts, and thus reduce the energy consumption. When the number of allocated resources of a host exceeds the threshold, one or more VM(s) should be migrated to the other hosts. If a method cannot find a proper host for the migrated VM(s) a new host must be turned on for the VM(s) that cause increasing the energy consumption of the data center. The proposed method predicts the future status of the resources of the host, and if it detects that it will return to the normal state within a short time, it will not consider it as an overloaded host. Therefore, the proposed method prevents VM migrations to avoid further costs.

Figure 2
Energy consumption (kWh)

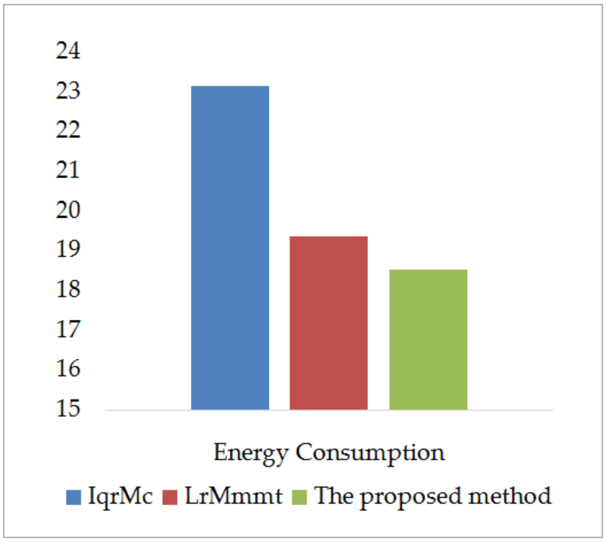
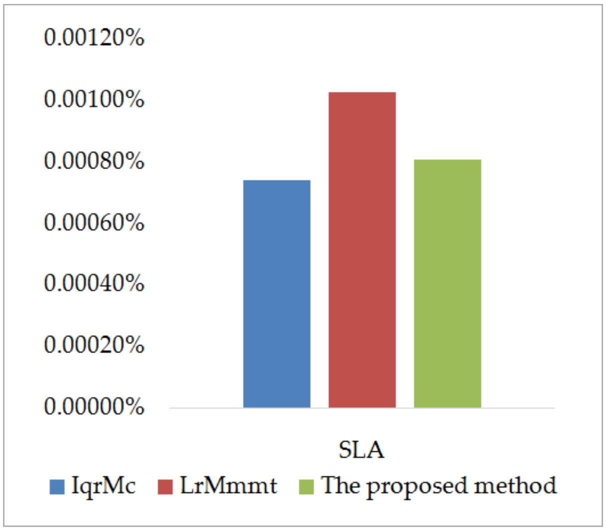


Fig. 3 shows the comparison of LrMmt, IqrMc, and the proposed method in terms of the SLA. The proposed method performs better than the both LrMmt and IqrMc methods. The proposed method performs better in terms of the SLA parameter than the LrMmt, because of improving the number of VM migrations.

Figure 3
SLA violation



The proposed method detects the host's threshold by predicting its threshold based on the host's current conditions and its history to prevent unwanted migrations, and thus addresses the performance degradation that occurs due to the VM migrations. The proposed method offers a higher level of SLA than the IqrMc, caused by hosting more VMs. Consequently, more VMs tend to compete for resources which in some cases cause a shortage of resources.

5. Conclusions

One of the main concerns of cloud data centers is to improve energy efficiency and to increase the profitability of these data centers, while keeping the quality of service at an acceptable level. In the proposed method of this paper, the problem of virtual machine (VM) consolidation was focused by considering host overload and host underload detection. The proposed method mapped the future state of the data center by predicting the hosts' workloads, and then based on these predictions, decided to specify a host as an overloaded or underloaded host. In the proposed method, an exponential smoothing technique was also leveraged to predict host utilization to make a better decision. Therefore, improper migrations were avoided, and the energy consumption was reduced. On the other hand, more accurate recognition of the underloaded hosts led to prevent continuous shutdowns of the hosts, and thus prevented a lot of VM relocations.

References

1. Abadi, R. M. B., Rahmani, A. M., Alizadeh, S. H. Challenges of Server Consolidation in Virtualized Data Centers and Open Research Issues: A Systematic Literature Review. Journal of Supercomputing. 2020, 76, 2876-2927. https://doi.org/10.1007/s11227-019-03068-1
2. Ahmad, R. W., Gani, A., Hamid, S. H. A., Shiraz, M., Yousafzai, A., Xia, F. A Survey on Virtual Machine Migration and Server Consolidation Frameworks for Cloud Data Centers. Journal of Network and Computer Applications, 2015, 52, 11-25. https://doi.org/10.1016/j.jnca.2015.02.002
3. Alidrees, A. Forecasting Consumer Consumption Behavior of Water Bottles Using Generlized Linear Model for Supply Chain Resilience. The University of Texas at El Paso. 2020, Open Access Theses & Dissertations. 3135. https://scholarworks.utep.edu/open_etd/3135
4. Anglano, C., Canonico, M., Guazzone, M. FCMS: A Fuzzy Controller for CPU and Memory Consolidation under SLA Constraints. Concurrency and Computation: Practice and Experience, 2017, 29(5), e3968. https://doi.org/10.1002/cpe.3968
5. Arianyan, E., Taheri, H., Sharifian, S. Novel Energy and SLA Efficient Resource Management Heuristics for Consolidation of Virtual Machines in Cloud Data Centers. Computers and Electrical Engineering, 2015, 47, 22-240. https://doi.org/10.1016/j.compeleceng.2015.05.006
6. Asghari, A., Sohrabi, M. K. Combined Use of Coral Reefs Optimization and Reinforcement Learning for Improving Resource Utilization and Load Balancing in Cloud Environments. Computing. 2021. https://doi.org/10.1007/s00607-021-00920-2
7. Asghari, A., Sohrabi, M. K., Yaghmaee, F. Online Scheduling of Dependent Tasks of Cloud's Workflows to Enhance Resource Utilization and Reduce the Makespan using Multiple Reinforcement Learning-based Agents, Soft Computing, 2020, 24, 16177-16199. https://doi.org/10.1007/s00500-020-04931-7
8. Asghari, A., Sohrabi, M. K., Yaghmaee, F. A Cloud Resource Management Framework for Multiple Online Scientific Workflows Using Cooperative Reinforcement Learning Agents. Computer Networks, 2020, 179, 107340. https://doi.org/10.1016/j.comnet.2020.107340
9. Asghari, A., Sohrabi, M. K., Yaghmaee, F. Task Scheduling, Resource Provisioning, and Load Balancing on Scientific Workflows Using Parallel SARSA Reinforcement Learning Agents and Genetic Algorithm. The Journal of Supercomputing, 2020. https://doi.org/10.1007/s11227-020-03364-1
10. Beloglazov A., Buyya R. Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers. Concurrency and Computation: Practice and Experience, 2012, 24(13), 1397-1420. https://doi.org/10.1002/cpe.1867
11. Beloglazov A., Buyya, R. Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centers under Quality of Service Constraints,"

IEEE Transactions on Parallel and Distributed Systems, 2013, 24(7), 1366-1379. https://doi.org/10.1109/TPDS.2012.240

12. Beloglazov A., Buyya R. OpenStack Neat: A Framework for Dynamic and Energy-Efficient Consolidation of Virtual Machines in OpenStack Clouds. Concurrency and Computation: Practice and Experience, 2015, 27(5), 1310-1333. https://doi.org/10.1002/cpe.3314
13. Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., Brandic, I. Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility, Future Generation Computer Systems, 2009, 25(6), 599-616. https://doi.org/10.1016/j.future.2008.12.001
14. Gardner, E. S., Jr. Exponential Smoothing: The State of The Art. Journal of Forecasting, 1985, 4, 1-28. https://doi.org/10.1002/for.3980040103
15. Gardner, E. S., Jr. Exponential Smoothing: The State of The Art-Part II. Journal of Forecasting. 2006, 22(4), 637-666. https://doi.org/10.1016/j.ijforecast.2006.03.005
16. Gharehpasha, S.,Masdari, M., Jafarian, A. The Placement of Virtual Machines Under Optimal Conditions in Cloud Datacenter. Information Technology and Control, 2019, 48(4), 545-556. https://doi.org/10.5755/j01.itc.48.4.23062
17. Gholipour, N., Arianyan, E., Buyya, R. A Novel Energy-Aware Resource Management Technique using Joint VM and Container Consolidation Approach for Green Computing in Cloud Data Centers, Simulation Modelling Practice and Theory, 2020, 104, 102127. https://doi.org/10.1016/j.simpat.2020.102127
18. Handlin, H., Thrash, B. Using A Total Cost of Ownership (TCO) Model for Your Data Center,2013. http://www.datacenterknowledge.com/archives/2013/10/01/using-a-total-cost-of-ownership-tco-model-for-your-data-center/
19. Horri, A. Mozafari, M. S., Dastghaibyfard, G. Novel Resource Allocation Algorithms to Performance and Energy Efficiency in Cloud Computing. Journal of Supercomputing, 2014, 69(3), 1445-1461. https://doi.org/10.1007/s11227-014-1224-8
20. Jangiti, S., Sriram VS, S. EMC2: Energy-Efficient and Multi-Resource- Fairness Virtual Machine Consolidation in Cloud Data Centres. Sustainable Computing: Informatics and Systems, 2020, 27, 100414. https://doi.org/10.1016/j.suscom.2020.100414
21. Khoshkholghi, M. A., Derahman, M. N., Abdullah, A., Subramaniam S., Othman, M. Energy-Efficient Algorithms for Dynamic Virtual Machine Consolidation in Cloud Data Centers. IEEE Access, 2017, 5, 10709-10722. https://doi.org/10.1109/ACCESS.2017.2711043
22. Li, Z., Yan, C., Yu, L., Yu, X. Energy-Aware and Multi-Resource Overload Probability Constraint-based Virtual Machine Dynamic Consolidation Method, Future Generation Computer Systems, 2018, 80, 139-156. https://doi.org/10.1016/j.future.2017.09.075
23. Liu, F., Ma, Z., Wang, B., Lin, W. A Virtual Machine Consolidation Algorithm Based on Ant Colony System and Extreme Learning Machine for Cloud Data Center. IEEE Access. 2020, 8, 53-67. https://doi.org/10.1109/ACCESS.2019.2961786
24. Mapetu, J. P. B., Kong, L., Chen, Z. A Dynamic VM Consolidation Approach Based on Load Balancing using Pearson Correlation in Cloud Computing. Journal of Supercomputing. 2020. https://doi.org/10.1007/s11227-020-03494-6
25. Martino, C. D., Sarkar, S., Ganesan, R., Kalbarczyk Z. T., Iyer, R. K. Analysis and Diagnosis of SLA Violations in A Production SaaS Cloud, IEEE Transactions on Reliability, 2017, 66(1), 54-75. https://doi.org/10.1109/TR.2016.2635033
26. Melbourne Cloud LAB. CloudSim: A Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services. http://www.cloudbus.org/cloudsim/. Accessed on December 7, 2020
27. Miller, R. Ballmer: Microsoft Has 1 Million Servers (July 2013). http://www.datacenterknowledge.com/archives/2013/07/15/ballmer-microsoft-has-1-million-servers/
28. Mills, M. The Cloud Begins with Coal. Big Data, Big Networks, Big Infrastructure, and Big Power. An Overview of the Electricity Used by the Digital Ecosystem, Digital Power Group, 2013.
29. Moges, F. F., Abebe, S. L. Energy-Aware VM Placement Algorithms for the OpenStack Neat Consolidation Framework. Journal of Cloud Computing. 2019, 8, 2. https://doi.org/10.1186/s13677-019-0126-y
30. Park, K. S., Pai, V. S. CoMon: AMmostly-Scalable Monitoring System for PlanetLab. SIGOPS Operating System Review, 40(1), 200. https://doi.org/10.1145/1113361.1113374
31. Salimian, L., Esfahani, F. S., Nadimi-Shahraki, M. H. An Adaptive Fuzzy Threshold-based Approach for Energy and Performance Efficient Consolidation of Virtual Machines. Computing, 2016, 98(6), 641-660. https://doi.org/10.1007/s00607-015-0474-5
32. Shaw, S. B., Singh, A. K. Use of Proactive and Reactive Hotspot Detection Technique to Reduce the Number of

Virtual Machine Migration and Energy Consumption in Cloud Data Center. Computer and Electrical Engineering, 2015, 47, 241-254. https://doi.org/10.1016/j.compeleceng.2015.07.020

33. Varasteh A., Goudarzi, M. Server Consolidation Techniques in Virtualized Data Centers: A Survey. IEEE Systems Journal, 2017, 11(2), 772-783. https://doi.org/10.1109/JSYST.2015.2458273
34. Witanto, J. N., Lim, H., Atiquzzaman, M. Adaptive Selection of Dynamic VM Consolidation Algorithm using Neural Network for Cloud Resource Management. Future Generation Computer Systems, 2018, 87, 35-42. https://doi.org/10.1016/j.future.2018.04.075
35. Xiao, H., Hu, Z., Li, K. Multi-Objective VM Consolidation Based on Thresholds and Ant Colony System in Cloud Computing. IEEE Access, 2019, 7, 53441-53453. https://doi.org/10.1109/ACCESS.2019.2912722
36. Zhang, F., Liu, G., Fu, X., Yahyapour, R. A Survey on Virtual Machine Migration: Challenges, Techniques, and Open Issues, IEEE Communications Surveys & Tutorials, 2018, 20(2), 1206-1243. https://doi.org/10.1109/COMST.2018.2794881
37. Zhou, Z., Abawajy, J., Chowdhury, M., Hu, Z., Li, K., Cheng, H., Alelaiwi, A., Li, F. Minimizing SLA Violation and Power Consumption in Cloud Data Centers using Adaptive Energy-aware Algorithms. Future Generation Computer Systems. 2018, 86, 836-850. https://doi.org/10.1016/j.future.2017.07.048
38. Zhou, Z., Hu, Z., Yu, J., Abawajy, J., Chowdhuri, M. Energy-Efficient Virtual Machine Consolidation Algorithm in Cloud Data Centers. Journal of Central South University, 2017, 24(10), 2331-2341. https://doi.org/10.1007/s11771-017-3645-z
39. Zhou, Z, Yu, J, Li, F, Yang, F. Virtual Machine Migration Algorithm for Energy Efficiency Optimization in Cloud Computing. Concurrency and Computation: Practice and Experience. 2018, 30(24), e4942. https://doi.org/10.1002/cpe.4942
40. Zolfaghari, R., Sahafi, A., Rahmani, A. M., Rezaei, R. Application of Virtual Machine Consolidation in Cloud Computing Systems. Sustainable Computing: Informatics and Systems, 2021, 30, 100524. https://doi.org/10.1016/j.suscom.2021.100524