Active VM Consolidation for Cloud Data Centers under Energy Saving Approach

Shailesh Saxena[1], Dr. Mohammad Zubair Khan[2], Dr. Ravendra Singh[3], Abdulfattah Noorwali[4]

[1]Research Scholar, Department of CS and IT, MJP Rohilkhand University, Bareilly, India

[2]Department of CS, Taibah University, Medina, KSA
mkhanb@taibahu.edu.sa
[3]Department of CS and IT, MJP Rohilkhand University, Bareilly, India
rsiet2002@gmail.com
Department of Electrical Engineering, Umm Al-Qura University, Makkah, Saudi Arabia

Abstract-Cloud computing represent a new era of computing that's forms through the combination of service-oriented architecture (SOA), Internet and grid computing with virtualization technology. Virtualization is a concept through which every cloud is enable to provide on-demand services to the users. Most IT service provider adopt cloud based services for their users to meet the high demand of computation, as it is most flexible, reliable and scalable technology. Energy based performance tradeoff become the main challenge in cloud computing, as its acceptance and popularity increases day by day. Cloud data centers required a huge amount of power supply to the virtualization of servers for maintain on- demand high computing. High power demand increase the energy cost of service providers as well as it also harm the environment through the emission of CO_2. An optimization of cloud computing based on energy-performance tradeoff is required to obtain the balance between energy saving and QoS (quality of services) policies of cloud. A study about power usage of resources in cloud data centers based on workload assign to them, says that an idle server consume near about 50% of its peak utilization power [1]. Therefore, more number of underutilized servers in any cloud data center is responsible to reduce the energy performance tradeoff. To handle this issue, a lots of research proposed as energy efficient algorithms for minimize the consumption of energy and also maintain the SLA (service level agreement) at a satisfactory level. VM (virtual machine) consolidation is one such technique that ensured about the balance of energy based SLA. In the scope of this paper, we explore reinforcement with fuzzy logic (RFL) for VM consolidation to achieve energy based SLA. In this proposed RFL based active VM consolidation, the primary objective is to manage physical server (PS) nodes in order to avoid over-utilized and under-utilized, and to optimize the placement of VMs. A dynamic threshold (based on RFL) is proposed for over-utilized PS detection. For over-utilized PS, a VM selection policy based on fuzzy logic is proposed, which selects VM for migration to maintain the balance of SLA. Additionally, it incorporate VM placement policy through categorization of non-overutilized servers as- ***balanced, under-utilized and critical***. CloudSim toolkit is used to simulate the proposed work on real-world work load traces of CoMon Project define by PlanetLab. Simulation results shows that the proposed policies is most energy efficient compared to others in terms of reduction in both electricity usage and SLA violation.

Keywords: *Cloud, VM consolidation, Dynamic Threshold, Fuzzy Reinforcement, Overutilized, Underutilized, Balanced, Critical.*

Manuscript received November 5, 2021
Manuscript revised November 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.11.47

1. Introduction

With the raise in users' computing power, the huge number of computing servers, increases the size of cloud data centers with this high consumption of electricity increases operational cost of the computing. Additionally, this high consumption of electricity influences the environment by the carbon footprint produce by it. Therefore, an efficient policies to reduce the electricity consumption in cloud data centers is required. These policies also responsible to identified the main sources for energy wastage in cloud data centers.

An inefficient usage of computing resources increases the level of energy wastage. A study [1] indicate that the dimensions for power usage of servers are very restricted. The same study defines power usage of resources in cloud data centers based on workload assign to them and indicates that idle server consume approx. 50% of its peak utilization power. Therefore, keeping servers in underutilized state is very inefficient regarding the consumption of electricity. To achieve this energy performance tradeoff, a VM consolidation policy try to run multiple instances of VMs on a single server through virtualization, as a result less number of servers is in working state with their highest utilization while rest are contribute in energy saving ratio..

However, the main challenge for this consolidation policy to adapt the dynamic and unexpected nature of the workload assign to VMs inside the physical server. Due to this the servers become underutilized or overutilized at any step of time. The overutilized server ncreased the performance degradation due to heavy load, while underutilized servers, causes inefficient energy consumption. A lots of research have been proposed to handle such type of situation in form of different management schemes for dynamic VM scheduling. The key idea of such schemes is to transfer of VMs from one server to another in order to maintain the efficient level of energy consumption, while increase the performance level of the system according to SLAs.

Maximum available solutions [2][3][4] have focus on

centralized decision making approach for VM scheduling, where decision of scheduling is define by a single node (generally called manager) periodically. These approaches have an issue regarding poor scalability, as there will be a bottleneck condition on manager if the number of servers as well as VMs inside those servers grow rapidly. The proposed RFL based VM consolidation scheme develop a dynamic monitoring system (DMS) in order to increase scalability by enabling more than one physical server as a manager to define the scheduling decisions in a hierarchy. DMS have monitoring the system in two levels; first at data center level through GM (global manager) and second at physical server level through LM (Local Manager). In this manner every server contributes in the process of VM consolidation dynamically to avoid the performance degradation in the situation of over utilization as well as high energy consumption in an under-utilization situation. The PDM will decrease through migrating one or more VMs away from over-utilized server to others, while energy consumption will reduce through migrating all VMs away from under-utilized server and convert them into a sleeping mode. Such decision based scheduling algorithms perform following four decision making tasks on every server to adjust the dynamic and unexpected nature of workload in cloud environment:

- How any server must be defined as overutilized,
- How any VM is selected for migration to others from an overutilized server, and
- How the status of any server must be defined as balanced, underutilized, and critical.
- When any server must be convert its state from working to switch off mode.

These schemes have a most significant challenge, what is the optimal way to make the decisions more effectively in real time manner to minimize consumption of energy without violation of SLAs of the system. Additionally, the optimal solution also maintain scalability of the scheme, if number of servers increases in data center due to the high demand of computation.

In the next section, a related literature of the proposed work is reviewed. This review provided a help to identifying the advantages and the limitations of the reviewed algorithms. After that, a proposed DMS framework of servers' load categorization is define. In this section, a RFL based server overutilized detection algorithm and VM selection algorithm for migration with placement policy is describe to reduce energy consumption of cloud datacenter. After that, simulation of proposed algorithm and policies through Cloud-Sim simulator using the real load trace of ConMon project provided planet lab is present. In the last section, conclusion based on the simulation result is define. Additionally, a future scope of the proposed work is define in the last.

2. Review of Related Literature

One of the first proposal regarding VM consolidation based on dynamic decision making system was define by Nathuji and Schwan [5]. In this proposal, authors applied a 2-level decision making policy for the management of resources. At the first level, the servers coordinates locally and operates the power management policies for the guest VMs on each server. Linux Kernel with on-demand driver is an example of such type of policy. Additionally, decisions for changing the state of power consumption are made according to maintain QoS levels. At second level, the decision is made regarding the status of all servers and managing the characteristics of rack or blade level requirements. In this proposed consolidation of VMs, live migration is used for unloaded VM to away from the existing server and put that server in a power-saving mode. It also included power consumption and number of VM migration as the efficiency parameters. Gmach et al. [6] also proposed a 2-level management of VM consolidation to include the unexpected behavior of the load on VMs. At primary level, VMs are identified cause for performance degradation of the server, whereas at next level, migration of VMs between servers is controlled. At primary level, first placement controller used a trace which have information of resource usage by VMs instantiated within the servers of the data center, and optimize the allocation of resources using this historical data to meet the desired quality of service level agreements. At second level controller could be a reactive migration controller operate with a mathematical logic primarily based feedback. This master controller have a consultant to observe the servers' resource utilization on regular basis and triggers a mathematical logic primarily based feedback whenever any server have too low or too high utilization of resource. Once that consultant detects an under-utilized or over-utilized situation of the servers, the fuzzy module of controller identifies a procedure acceptable as a remedy in such situation. Beloglazov et al. [7], proposed a hierarchy structure to manage a distributed systems for unified design of distributed dynamic VMs for cloud data centers. In this work, they proposed hierarchical levels of decision making system that includes global and local two types of managers. The local manager resides on each server as a VM monitor(VMM), continuously monitor the status of VMs in terms of CPU and RAM usage, resize the VMs as per resource requirements, and also make a decision; which VMs should be migrated from any server and when. While the global manager (a master server), collects information from the local managers and maintain an overview of resource usage inside each server. The global manager also optimize the placements of VMs inside the data centers.

In [8] [9] authors, presented the review of several heuristic algorithms for detecting the status of servers i.e

overloading or underloading. Additionally they also show an analytical review of VM selection and placement policies used for migration in VM Consolidation strategy. In the same category, Barabagallo et al. [10] proposed a model of self-organizing server in P2P network for fully decentralized data center including another bio-inspired algorithm. This collaboration increases the energy efficiency of the system through redistribute the workload over servers. Their idea is to have multiple entities in the role of manager to observe and collect information about the deployed VMs. Then after this observed information used to make decisions about whether any VM should select for migrations or not.

In [11] authors proposed a VM consolidation scheme having classification of four sub-problems: full server observation, underloaded server detection, Selection of VMs to migrate and placement of that selected VMs. Proposed scheme used a utilization threshold policy (THR) to detect the overloaded servers. If the utilization electronic resources exceed the higher value of threshold, only some VMs are required to be migrated in order to reduce performance degradation of the system, while utilization is below the lower value of threshold, all VMs is to be migrated and that server switched to the energy saving mode to maintain the efficient use of energy in the system. Here the proposed framework simulated the results using four different polices to select VM(s) for migration: Random Selection Method, Selection using Single Threshold, Migrations in Minimum Time and Performance Growth Migration. The pliability of this proposal shown in the results of simulation. In [12] authors proposed a dynamic VM scheduling policy to improve the efficient usage of electricity as well as SLA violation in the data centers of cloud. Using the values of mean and standard deviation with respect to CPU usage of VMs, they proposed a decision making for the status of server i.e extra loaded or not. Additionally, they also used positive maximum correlation coefficient to select VMs for migration from extra-loaded servers to others. Simulation result shows that the proposed policies for the detection of extra loaded server have better performance in terms of SLA violation from the policies already simulated in CloudSim.

In [13], authors proposed a framework for resource scheduling using fuzzy logic feedback to incorporate the dynamic impact of real-time workload during scheduling the workload over resources. It provided real-time execution of all relevant experts while minimizing the impact of limited resources and uncertainties on the system performance.

3. Proposed Framework

In this section, we concentrate on detecting an overutilized physical servers, with a sub-problem of dynamic scheduling of VMs. In this concern a solution is proposed that is based on reinforcement with fuzzy logic learning. In cloud computing, workload nature is directly dependent on dynamic behavior of users' request, so detection of physical server overutilization can be considered as a dynamic decision making task. A load of any physical server depends on the utilization of CPU, RAM and bandwidth during the execution of the services. But in the proposed work the load is calculated only in terms of CPU and RAM utilization of the servers by assuming the ideal condition of bandwidth utilization. Accordingly, four categories are described for physical servers and each server will be in either one of the category; ***overutilized, balanced, underutilized and critical***. For the categorization of physical servers a threshold based on its utilization ratio is used, but defining the threshold value is not an easy task, as low value of threshold can convert the system into maximum overbooking server system with inefficient energy consumption of data centers. Similarly, a large value of threshold causes the increment in performance degradation with respect to 100% utilization of CPU and RAM of the servers. So a RFL (reinforcement with fuzzy logic) learning based DMS is employed with three type of CDC level utilization threshold; T_{Max} , T_{Avg} , T_{Min} initialized with 70%, 40% and 10% respectively according to the utilization performance SLA.

A server is overutilized on exceed the CPU and RAM demand of all VMs in the server with the maximum utilization threshold of the CDC, as result SLA of the CDC compromise. A server is balanced if the total CPU and RAM demand of VMs in the server is lies between maximum and average utilization threshold of CDC and causes to maintain the balance of SLA. A server is underutilized when the total demand of CPU and RAM of VMs in the server is lies between average and minimum utilization threshold of CDC and causes to maintain the SLA violation through placement of migrated VMs from other servers. A server is critical when it's total demand of CPU and RAM in less than minimum utilization threshold and causes to performance degradation due to less utilization. Next, some VMs from overutilized server need to be migrated on underutilized server in order to maintain the utilization performance. For saving the energy consumption, all VMs from critical server need to migrate on underutilization server in order to switch off the critical server. The steps of the decision making system (DMS) in CDC related to VM consolidation can be illustrated in algorithm 1.

To maintain the SLA for performance utilization and energy consumption, migration of VMs are needed from overutilized and critical servers respectively. In case of energy consumption SLA, the migration of all VMs from critical server is beneficial as their total demand of utilization is less than 10% of total utilization capacity.

Algorithm 1: RFL based VM Consolidation

Input: Load of PS_i ($1 \leq i \leq N$)
Output: Status of PS_i
Process:
1. Initialize running PS number N 2. For (i=1 to N); Create M number of VMs to allocate in each PS; 3. Initialize T_{Max}= upper threshold & T_{Min}= lower threshold, & $T_{Avg} = [T_{Max} - ((T_{Max} - T_{Max}) / 2)]$; 4. For(i=1 to N); Select TH_i using fuzzy inference and update according reinforcement; 5. IF ($PS_i > TH_i$) says PS_i is over-utilized; 6. Then run algorithm 2 (VM selection for migration) 7. Else ($T_{Max} \geq PS_i \geq T_{Avg}$) says PS_i is balanced; 8. No action required for balanced servers; 9. Else ($T_{Avg} \geq PS_i \geq T_{Min}$) says PS_i is underutilized; 10. Then used to place the migrated VMs until it become balanced; 11. Else ($PS_i < T_{Min}$) says PS_i is critical; 12. Then all VM migrate to underutilized PS; 13. IF PS_i is critical & all VMs migrate to another PS; 14. Then move PS_i to sleep mode // energy saving mode 15. End IF 16. Calculation of energy utilization and SLA violations; 17. End For

But choosing of VMs to migrate away from an overutilized server is a challenging task. In this scope, a utilization-based VMselection policy is proposed regarding to make decision about those VMs required to migrated away when any physical server become overutilized. In this paper, VMs contribution in terms of CPU and RAM utilization of the server are used as main criteria to select the VMs for migration. Additionally, MaxU and MinU two general sub-criteria used here for the selection of VMs based on their maximum utilization and minimum utilization respectively. If an overutilized physical server used MaxU criteria to VM selection for migration, the probability of degrading the performance caused by overutilized will be decreased (as the highest loaded VM is always migrate), but instead it will increase its own contribution in inefficient consumption of electricity required to run the data center. It also has more chance to convert the overutilized server into balanced server with a less migration of VMs, which has a positive impact on the performance degradation caused by more migration, In contrast, if any server uses MinU criteria, it will increase the probability of performance degradation caused by overloading, but it's own contribution will decrease to define the efficient usage of energy inside the data center. On the contrary, using MinU criteria has a negative impact on the performance degradation due to large number of migration. The proposed VM selection policy for migration including their placement can be illustrated in algorithm 2.

Algorithm 2: VM Selection for Migration

Input: over-utilized PS_i OR critical PS_i ($1 \leq i \leq N$)
Output: VM for migration
Process:
1. get X (CU, MU, NumVM), utilization value of over utilized PS; 2. select O(MaxU, MinU), criteria of VM migration from one PS to another 3. set migrated VM placed on underutilized server until it become balanced 4. get NumVM, number of VMs in critical server 5. select all VMs to migrate move the host to sleep mode 6. set migrated VM placed on underutilized server until it become balanced 7. Repeat until SLA is achieved 8. END

A. Architecture of RFL Based DMS

A hierarchical architecture of proposed DMS shown in Figure 1, there is a Local managers at each server as VM level agents of DMS. Their main task is to observe the CPU and RAM usage of VMs deployed in the servers, maintain the number of VMs as per its demands and take decision of VM selection for migration. On the higher level of hierarchy, DMS also has a global manager as a master agent for making dynamic decision about the status of every server using informations collected from the local managers. Global manager contribute in active VM consolidation strategy by making the decision for VM migration and conversion the power states of the servers.

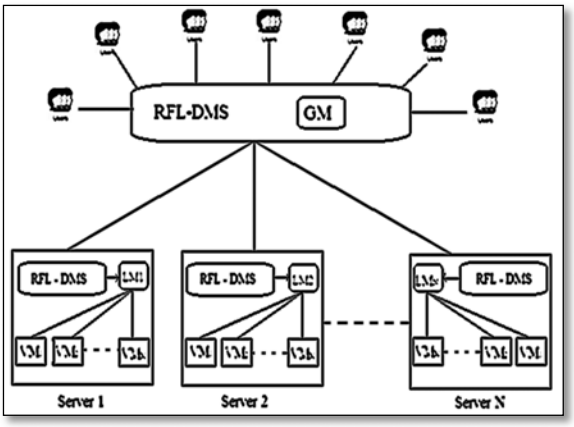


Figure 1: A System Architecture for RFL Based DMS

In the proposed RFL based DMS, the global manager perceive the utilization status of each server according the received information from the local managers periodically.

Then it make decision about the detection of overutilized servers. The global manager have an input queue, used to keep the state information of each server received at each time step *t*. Data kept in the queue is considered as input data for RFL based DMS for defining the threshold value for each server and send these values to the local managers respectively. After detecting the overutilized server, RFL based DMS shifted towards the local manager and then it makes the decision about the selection of VM(s) for migration to maintain the balance of efficient usage of electricity as well as performance of data centers.

B. Formulating the Proposed RFL Based VM Consolidation

Fuzzification[17][19] is an attractive approach to handle uncertain, imprecise, or un-modeled data and define an intelligent decision-making system on it. Detection of overutilization server and selection of VMs for migration from any server have the same level of uncertainty due to the uncertain nature of users' request for services. Threshold based detection of overutilized server is most effective method with the challenge of defining the threshold value (as discussed earlier). So RFL describe a fuzzy inference system with the inputs X_i: [CU, RU, NumVM] {i= 1 to N} to define the dynamic values of threshold TH: [TH_1, TH_2, …….TH_N] (different for each server) as an output. Membership function of the inputs and output variable define as;

CPU utilization (CU) = [Low , Medium , High]
[(<30), (30-70), (>70)]

RAM Utilization (RU) = [Low , Medium , High]
[(<30), (30-70) (>70)]

Number of VMs (NumVM) = [Less, Average, High]

Threshold (TH) = [T_{MAX}, T_{AVG}, T_{MIN}]

A sample of inference rules for define the threshold for each server in dynamic manner can be illustrate in figure 2-

If CU is H and RU is H and NumVm is L then TH is T_{MAX}
If CU is H and RU is H and NumVm is A Then TH is T_{MAX}
If CU is H and RU is H and NumVm is H Then TH is T_{MAX}
If CU is H and RU is M and NumVm is L Then TH is T_{MAX}
If CU is H and RU is M and NumVm is A Then TH is T_{MAX}
= = = = = = = = = = = =
= = = = = = = = = = = =
= = = = = = = = = = = =

If CU is M and RU is L and NumVm is A Then TH is T_{AVG}
If CU is M and RU is L and NumVm is H Then TH is T_{AVG}
= = = = = = = = = = = =
= = = = = = = = = = = =

If CU is L and RU is L and NumVm is A Then TH is T_{MIM}
If CU is L and RU is L and NumVm is H Then TH is T_{MIN}

Figure 2: Inference Rule for RFL based Dynamic Threshold

Similarly, there are different policies of VM selection with its own advantages and limitations. Therefore, a fuzzy inference system is again used here to select the policy of VM selection according to the merit of its benefits. Now the proposed RFL again used the X_i: [CU, RU, NumVM] {i= 1 to N} as an inputs of the fuzzy inference to define the decision for VM selection VMS: [VMS_1, VMS_2,……,VMS_N] (different for each server) as an output. Membership function of the inputs are similar as above, but output variable is define the selection of VM migration policy like MaxU (a VM whose utilization is maximum from all other VMs in a server will be select for migration) and MMT (a VM takes minimum migration time will be select for migration) as;

VM selection (VMS) = [MaxU, MMT]

A sample of inference rules for define the VM selection policy for overutilized server according to the benefits of the policy can be illustrate in figure 3-

If CU is H and RU is H and NumVm is L Then VMS is MaxU
If CU is H and RU is H and NumVm is A Then VMS is MaxU
If CU is H and RU is H and NumVm is H Then VMS is MaxU
= = = = = = = = = = = =
= = = = = = = = = = = =
If CU is M and RU is M and NumVm is L Then VMS is MaxU
= = = = = = = = = = = =
If CU is M and RU is L and NumVm is L Then VMS is MMT
= = = = = = = = = = = =
= = = = = = = = = = = =
If CU is L and RU is L and NumVm is A Then VMS is MMT
If CU is L and RU is L and NumVm is H Then VMS is MMT

Figure 3: Inference Rule for RFL based VM Selection

C. Load Estimation

Estimation of load is one of the most challenging tasks in the cloud. A server's load depends on the usage of CPU, RAM and bandwidth of the service links [14]. In the proposed active VM consolidation, the load is estimated through the usage of CPU and RAM under the idle condition of bandwidth. The main task of the proposed scheme is to eliminate the critical server(s) from the data center through migrating all VMs away from it and switch that server(s) in energy saving mode. Therefore, the load of any servers is estimated through the summation of the CPU and RAM usage of the VMs under that server [15],

$$Load(PS) = \sum_{i=1}^{m} \frac{(CU+RU)i}{2} \quad (1)$$

where CU and RU is the utilization ratio of the server according to the total capacity of that server,

$$CU\ (\%) = \frac{Used\ Core}{Total\ Core\ Capacity} \quad (2)$$

$$RU\ (\%) = \frac{Used\ RAM}{Total\ RAM\ Capacity} \quad (3)$$

D. Performance Metrics

According to the study in [9], we employed a generic and

application-independent metric for SLA violations based on the maximum (near about 100%) utilization of resources like CPU and RAM requested by users' inside a VM is satisfied. Therefore, to evaluate the proposed RFL based VM consolidation approach; energy-consumption (EC), SLAT (SLA violation time per active server), PDM (performance degradation due to migrations), ESV (Energy SLA violation) are used as SLAs metrics. Additionally a reward feedback is also used to define reinforcement behavior of the system to adopt the uncertainty of the users' request.

Energy Consumption: To calculate the value of energy consumption for each server a linear function is used based on the utilization of CPU and RAM of the server. Equation (4) is expressed the linear approximation of power consumption of any server as:

$$P = P(U_{CPU,\,RAM}) = P_{IDLE} + (P_{PEAK} - P_{IDLE}) * (U_{CPU,\,RAM}) \quad (4)$$

As discussed earlier, P_{IDLE} is define as a fraction value of P_{PEAK}, now equation (4) can be expressed as follows:

$$P = f * P_{PEAK} + P_{PEAK}(1 - f) * U_{CPU,\,RAM} \quad (5)$$

For the simulation of idle power condition, the value of P_{peak} is set on 230 W while the value of f has been set to 0.5 [20]. In this way, the total consumption of energy in a server during a time step t can be expressed through equation (6) as:

$$EC = \int_{t_1}^{t_2} P(U_{CPU}(t))\, dt.$$

SLA Violation: SLA violations (SLAV) is used to define the level of performance degradation in cloud data centers and it can be calculated from both SLAT and PDM metrics as-

$$SLAV = SLAT * PDM \quad (7)$$

Where

$$SLAT = \frac{i}{N}\sum\frac{Ts}{Ta} \quad (8)$$

and

$$PDM = \frac{1}{M}\sum\frac{Cd}{Cr} \quad (9)$$

Here, N is total number of servers in the data center; Ts is the time interval, in which any server achieve its maximum level of utilization and Ta is the total working period of any server. While M indicates the number of VMs inside the data center, C_d is degradation in performance of VMs caused by its migrations and Cr is total required capacity by VMs during its whole working period. According to the definition of QoS, resources like CPU, RAM, bandwidth and storage should also be taken into account. Since the CPU and RAM are the main resources, generally described in the service request of cloud users. So the proposed approach, measured the performance only on the basis CU and RU of the servers.

Reward Feedback: It is a function that provide a feedback in numeric form to elaborate the current state of the environment. It is also define as a mapping of the next state with the current state to indicate what is good in an immediate sense. In the proposed architecture, this reward feedback returns a value to DMS indicating the effectiveness of the current decision that utilized to make the next decision. In this manner an efficient usage of electricity and performance in/of datacenter is achieve. In fact, the proposed RFL based DMS define a global feedback for setting the threshold values for each server to adjust the dynamic nature of workload on the servers. A standard performance metric[8] called ESV (energy SLA violation) is used to show the performance of the proposed work. ESV is calculated through the product combination of SLA parameters (SLAT and PDM) and EC as shown in equation (10):

$$ESV = SLAT \times PDM \times EC. \quad (10)$$

Consequently, the value of global reward feedback can be calculated by taking the inverse of the ESV value as shown in equation (11):

$$R = 1/ESV \quad (11)$$

4. Simulation and Results

The simulation and evaluation of the proposed RFL based DMS system is carried out through CloudSim toolkit [16]. To show the effectiveness of the proposed approach, simulation is define over 800 physical servers situated over different locations and allocated them a real life workload data that generated by PlanetLab through CoMon project related to monitoring the infrastructure of PlanetLab [9][17[18]. Servers are configured using two categories in equal manner, that means half of the servers are configured in one category of HP ProLiant Server i.e. *G4(110)* and another half are configured in another category of HP ProLiant Serbver i.e. *G5(110)*. Table 1 and Table 2 are used to express the characteristics of these servers and VMs deployed under these servers respectively.

Table 1: Server Level Characteristics of Data Center

Type of Servers	No. of Servers	Core Type	CPU (MHz)	RAM
G4(ML 110)	400	Dual	1860	4 GB
G5(ML 110)	400	Dual	2660	4 GB

Table 2: VM Level Characteristics of Servers

VM Type	Core Type	CPU (MIPS)	RAM
Type 1 (High Instance)	Single	2500	2048
Type 2 (Medium)	Single	2000	1024
Type 3 (Small)	Single	1000	1024
Type 4 (Micro)	Single	500	512

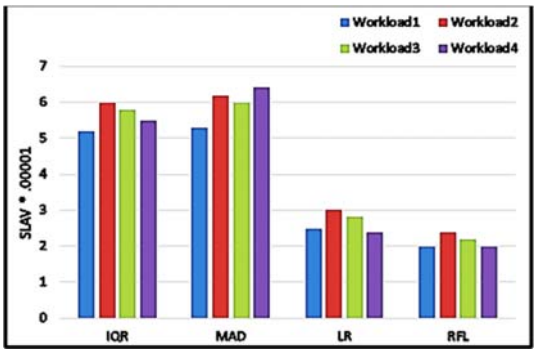
Here a real workload data trace of 10 days is used, that is collected from the VMs of thousands servers deployed over 500 different locations [9]. This real world traces of workload data contains the value of utilization at the interval of every 5 min. Table 3 is used to shows the characteristics of the workload on four different days taken from the workload data trace.

Table 3: Workload Characteristics (CPU & RAM utilization)

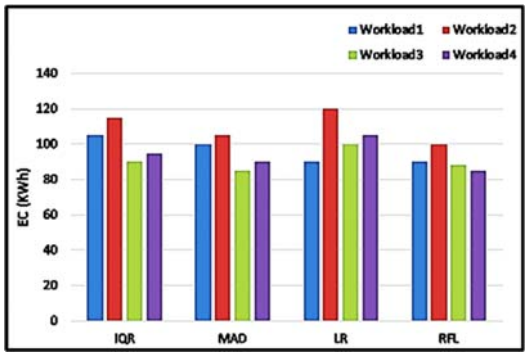
Data	Number of VMs	Mean	St. dev.	Median	Qrt 1	Qrt 2
Workload 1	898	11.88%	16.83%	6%	2%	14%
Workload 2	1516	9.06%	12.78%	5%	2%	13%
Workload 3	1463	11.39%	16.55%	7%	2%	16%
Workload 4	1233	12.56%	15.07%	6%	2%	17%

To show the efficiency, the proposed work is compared with three heuristic algorithms IQR, MAD and LR which have been proposed and simulated on cloudsim [9]. The first two algorithms are based on auto adjustment of the CPU utilization threshold using qualitative analysis of historical data of VM utilization. The third algorithm works on the observation of previous k values regarding the CPU utilization for estimating the CPU utilization in near future. In comparison with the RFL, none of these have efficient future decision makings regarding the utilization of servers on the consequence of the current utilization. In addition, they don't have much ability to adjust the unexpected consequences caused by estimating action in long term.

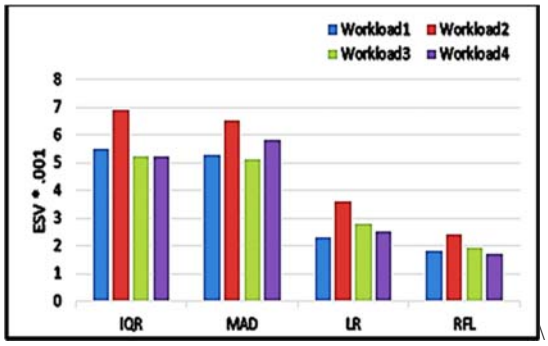
In simulation, the performance of proposed RFL is compared with LR, IQR and MAD is defined through the values of ESV, Energy Consumption (EC), SLA Violation (SLAV = PDM ×SLAT) and number of migrations simulated over the workloads of four different days (see Table 3). The simulated result of proposed RFL with other shown in Figure 4 in form of bar graphs. As Figure 4(a) and 4(b) shows, RFL have less SLA violation as well as energy consumption from others. Figure 4(c), shows that RFL have better performance than others, in terms of reducing the level of ESLAV. It also define the minimum number of migration for maintain the balance of SLA, see Figure 4(d).



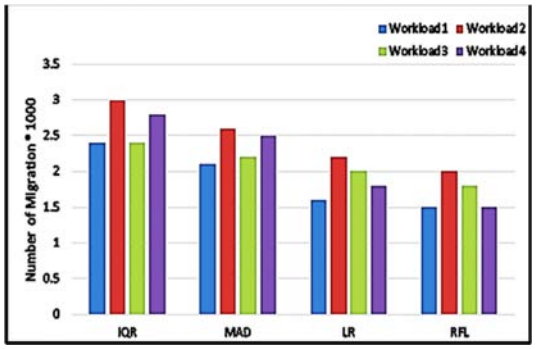
(a) SLA Violation



(b) Energy Consumption



(c) Energy SLA Violation



(d) Number of Migration

Figure 4: Performance of RFL in term of SLAV, EC, ESV and Number of Migration.

5. Conclusion

In this paper, a new policy of servers' categorization according to its utilization level in CDC (cloud data center); specially underutilized servers into three states. The proposed framework have a hierarchical decision making system based on reinforcement with fuzzy logic learning (RFL) to define which server is overutilized or not. It proposed a fuzzy inference system (FIS) with reward feedback to define the threshold for each server dynamically. Additionally, it also have a FIS to select the VMs for migration from one server to another. The main objective of this proposal is reduce energy consumption while balanced SLA in cloud data center. The simulation results shows that the proposed approach achieve a little improvement in energy-SLAV tradeoff form the existing policies in CloudSim.

In future, we will try to calculate the workload of cloud data centers in near future using some forecast algorithms like prophet etc. On that forecast, define a VM scheduling in such manner, that only required server will be in active mode or rest will be in sleep (electricity saving) mode, to avoid the migration of VMs.

References

[1] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso, "Power provisioning for a warehouse-sized computer", In *Computer Architecture News*, volume 35, pages 13–23. ACM, 2007.

[2] Eugen Feller, Louis Rilling, and Christine Morin, "Energy-aware ant colony based workload placement in clouds", In *Proceedings of the 12th Int. Conference on Grid Computing (IEEE /ACM)*,pages26–33.IEEE Computer Society, 2011.

[3] Jianhua Gu, Jinhua Hu, Tianhai Zhao, and Guofei Sun, "A new resource scheduling strategy based on genetic algorithm in cloud computing environment", in *Journal of Computers*, 7(1):42–52, 2012.

[4] Helmut Hlavacs and Thomas Treutner, "Genetic algorithms for energy efficient virtualized data centers", in *Network and service management (cnsm), 2012 8th international conference and 2012 workshop on systems virtualiztion management (svm)*, pages 422–429. IEEE, 2012.

[5] Ripal Nathuji and Karsten Schwan, "Virtualpower: Co-ordinated power management in virtualized enterprise systems", in *ACM SIGOPS Operating Systems Review*, volume 41, pages 265–278. ACM, 2007.

[6] Daniel Gmach, Jerry Rolia, Ludmila Cherkasova, Guillaume Belrose, Tom Turicchi,and Alfons Kemper, "An integrated approach to resource pool management: Policies, efficiency and quality metrics", in *Dependable Systems and Networks With FTCS and DCC, 2008. DSN 2008. IEEE International Conference on*, pages 326–335. IEEE, 2008.

[7] Anton Beloglazov and Rajkumar Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers", in *MGC@Middleware*, page 4, 2010.

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

[9] Anton Beloglazov and Rajkumar Buyy, "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*, 24(13):1397–1420, 2012.

[10] Donato Barbagallo, Elisabetta Di Nitto, Daniel J Dubois, and Raffaela Mirandola. "A bio-inspired algorithm for energy optimization in a self-organizing data center", In *Self-Organizing Architectures*, pages 127–151. Springer, 2010.

[11]. A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generation Computer Systems*, pp.755-768, May. 2012.

[12]. A. Beloglazov and R.Buyya , "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," *Proceedings of the 8th international workshop on middleware for grids, clouds and e-science ACM*, pp. 4, 2010.

[13] P. Rattanatamrong and J. A. B. Fortes, "Fuzzy scheduling of real-time ensemble systems," in *Proceedings of the International Conference on High Performance Computing and Simulation (HPCS '14)*, pp. 146–153, IEEE, Bologna, Italy, July 2014.

[14] C. M. S. Magurawalage, K. Yang, L. Hu and J. Zhang, "Energy-efficient and network-aware offloading algorithm for mobile cloud computing," Computer Networks, vol. 74, pp. 22–33, 2014.

[15] F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu et al., "Energy-aware VM consolidation in cloud data centers using utilization prediction model," IEEE Transactions

on Cloud Computing, vol. 7, no. 2, pp. 524–536, 2019.
[16] Rodrigo N Calheiros, Rajiv Ranjan, Anton Beloglazov, César AF De Rose, and Rajkumar Buyya, "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", *Software:Practice and experience*, 41(1):23–50, 2011.
[17] Mohd A. H. Monil and R. M. Rahman, "VM consolidation approach based on heuristics fuzzy logic, and migration control", in Journal of Cloud Computing: Advances, Systems and Applications (2016) DOI 10.1186/s13677-016-0059-7
[18] KyoungSoo Park and Vivek S Pai, "Comon: a mostly-scalable monitoring system for planetlab", in *ACM SIGOPS Operating Systems Review*, 40(1):65–74, 2006.
[19] Zhijia Chen, Yuanchang Zhu, Yanqiang Di, and Shaochong Feng, "A Dynamic Resource Scheduling Method Based on Fuzzy Control Theory in Cloud Environment" in Journal of Control Science and Engineering (Hindawi) Volume 2015, Article ID 383209, http://dx.doi.org/10.1155/2015/383209

Author's Profile

Shailesh Saxena is now pursuing his Ph.D in Computer Science at Faculty of Engineering, M. J. P. Rohilkhand University, Bareilly, India. He received the Master of Technology degree in CSE from U. P. Technical University, Lucknow, India, in 2012. His main area of research is computing. He published several research papers on Distributed Computing, Grid Computing and Cloud Computing. Now he continues his research on Energy-Saving Computing or Green Computing.

Mohammad Zubair Khan received the Master of Technology degree in CSE from U. P. Technical University, Lucknow, India, in 2006, and the Ph.D. degree in computer science and information technology from the Faculty of Engineering, M. J. P. Rohilkhand University, Bareilly, India. He was the Head and an Associate Professor with the Department of CSE, Invertis University, Bareilly, India. He has more than 18 years of teaching and research experience. He is currently an Associate Professor with the Department of Computer Science, Taibah University, Medina KSA. He has published more than 60 journals and conference papers. His current research interests include data mining, big data, parallel and distributed computing, the theory of computations, and computer networks. He has been a member of the Computer Society of India since 2004.

Dr. Ravendra Singh is presently working as a Professor in Department of CS & IT, Faculty of Engineering, M.J.P. Rohilkhand University, Bareilly, India. He has more than 22 years academic experience including 17 years of research. He has authored over 90 publications and 4 books. In past he did maximum research in the field of parallel and distributed computing or computer network regarding task allocation & scheduling. Currently his research includes Artificial intelligence, Machine learning, Deep learning and its applications.

Abdulfattah Noorwali (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Western Ontario, London, ON, Canada, in 2017. His thesis title was Modeling and Analysis of Smart Grids for Critical Data Communication. He is currently the Chairman of the Electrical and Computer Engineering Department, Faculty of Engineering and Islamic Architecture, UmmAl-Qura University, where he is an Assistant Professor. He is also a Senior Consultant with Umm Al-Qura Consultancy Oasis, Institute of Consulting Research and Studies (ICRS), Umm Al-Qura University, where he is the Chairman of Vision Office of Consultancy. He has authored many technical articles in journals and international conferences. His research interests include smartgrid communications, cooperative communications, wireless networks, the Internet of Things, crowd management applications, and smart city solutions.