AUTO RESOURCE MANAGEMENT TO ENHANCE RELIABILITY AND ENERGY CONSUMPTION IN HETEROGENEOUS CLOUD COMPUTING

Moataz H. Khalil[1, 2], Mohamed Azab[2], Ashraf Elsayed[3], Walaa Sheta[1, 2], Mahmoud Gabr[3] and Adel S. Elmaghraby[1,2]

[1]CECS Department, University of Louisville, Kentucky, USA.
[2]The City of Scientific Research and Technology Applications, Egypt.
[3]Department of Mathematics & Computer Science, Faculty of Science, Alexandria University, Alexandria, Egypt.

ABSTRACT

A classic information processing has been replaced by cloud computing in more studies where cloud computing becomes more popular and growing than other computing models. Cloud computing works for providing on-demand services for users. Reliability and energy consumption are two hot challenges and tradeoffs problem in the cloud computing environment that requires accurate attention and research. This paper proposes an Auto Resource Management (ARM) scheme to enhance reliability by reducing the Service Level Agreement (SLA) violation and reduce energy consumed by cloud computing servers. In this context, the ARM consists of three compounds, they are static/dynamic threshold, virtual machine selection policy, and short prediction resource utilization method. The Minimum Utilization Non-Negative (MUN) virtual machine selection policy and Rate of Change (RoC) dynamic threshold present in this paper. Also, a method of choosing a value as the static threshold is proposed. To improve ARM performance, the paper proposes a Short Prediction Resource Utilization (SPRU) that aims to improve the process of decision making by including the resources utilization of future time and the current time. The output results show that SPRU enhanced the decision-making process for managing cloud computing resources and reduced energy consumption and the SLA violation. The proposed scheme tested under real workload data over the CloudSim simulator.

KEYWORDS

Cloud computing, Service Level Agreement, Reliability, energy consumption, Virtual machine migration, Resource management, utilization resource prediction.

1. INTRODUCTION

An emerged innovative paradigm for delivering applications, platforms, and computing resources (processing power/bandwidth/storage) is named cloud computing. This new paradigm uses a Pay-As-You- Go (PAYG) model to the delivery application to customers [1]. Clouds are presented as a platform for various types of applications with various Quality of Service (QoS) features, like performance, availability, and reliability. A Service Level Agreement (SLA) is a contract between cloud providers and customers that have the previous features. The failure to comply with QoS features can reduce the reliability of applications and incur SLA violations, resulting in fines to the cloud provider [2].

Cloud computing falls into three models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). IaaS Clouds afford a virtual computing environment, where computing ability is delivered by allocating Virtual Machines (VMs) to IaaS users on request. IaaS gives cloud users the chance to run their applications on the most appropriate virtual machine, also pay only for the actual resources that are used [3]. Based on the large scale and high complexity of cloud data centers, reliability, and energy efficiency are two important challenges in cloud computing that need careful attention and research [4]. The Reliability of cloud computing is known as the framework of security or the framework of resource and application failures. Giving to the hard complexity of the cloud architecture, failures became certain. It has been shown that the system with100K processors experience a failure every couple of minutes [5]. In a survey of 63 Data Centers done by [6], it has been reported that the mean downtime cost of each data center rose to $740,357 from $500,000 in 2010 (+38%). Improving cloud computing reliability is a vital factor in increasing users' requests, which, due to increasing cloud computing profit. Approximately 45% of the total operating expenses of IBM data centers go in electricity bills [7]. Therefore, saving the consumed energy is targeted by cloud computing providers. Where the profit can increase by decreasing the cost of electricity bills. The risk of non-reliability produces from overutilization (overloading) of cloud hosts. Also, the main source of high energy consumption is underutilization (under loading) of cloud hosts.

The reliability of the system increases as high utilization hosts is reduced by transferring running virtual machines to a lower utilization host. In Addition, the huge number of resources used to store backups or run replicas have a counteractive effect on the energy efficiency of cloud computing. On the other side, with increasing the virtual machine consolidation (defined as allocating running virtual machines in less number of physical hosts), the energy consumed of the system reduces. But a huge VM consolidation has a negative impact on cloud computing reliability. Now, a critical tradeoff between reliability and energy consumption of cloud computing is obvious. Both the reliability and energy efficiency of cloud computing increases asymmetrically. This trade-off opens up new opportunities and challenges in cloud computing by considering both these elements simultaneously. The arriving to equilibrium status between these two metrics consider NP-hard problem from different perspectives such as SLA and operational cost and environment. There is a special requirement for more research in the area of enhancing the relationship between system reliability and energy consumption in cloud computing [4].

Figure 1 represents the optimal status of a cloud computing system for both reliability and energy consumption. The status of physical machines (overloading or under loading) is detected by using a utilization threshold. A utilization threshold is classified as a single resource (like CPU only) or multiple resources (like CPU, memory, storage, network bandwidth) and it can be static or dynamic. If a host utilization is below the threshold, then this host is underutilization and all VM running on that host are moved to the other host, which is known as a host consolidation. If a host utilization is higher than the threshold, then this host is over-utilized therefore one or more VM need(s) to migrate by using the live VM migration.

The live VM migration technology is enabling cloud providers to reallocate VMs from overloading physical hosts to other physical hosts that are not suffering from overloading. At the same time, the cloud providers consolidate VMs into a few physical hosts and switch off unused machines to reduce energy consumption [9, 10]. The live migration technology assists in enhancing reliability and allows energy savings while keeping a satisfactory level of SLA [11]. The problems of detecting the status of host (overutilization or underutilization), selecting VM which will be migrated, and the host will be received migrated VM are the most related problems to VM migration [10, 12]. After overutilization hosts are detected, the decisions for two problems need to make, the problems are; (1) selecting which and how many virtual machines will migrate

from overutilization host to appropriate host, and (2) What is an appropriate (target) host? First, given that migration is expensive, the VM selection policies play a vital role to decrease the number of VM migrated and time consumed for migrating. Second, the target host is a suitable host that able to reallocate VM on it. For an overutilization scenario, the target host should not be over-utilized after migrated VM did. If there is not an active host able to reallocate VM migrated, one of the inactive hosts resumes operating and reallocating VM on it. For an underutilization scenario, all VMs consolidate in less number of active hosts without causing overutilization for these hosts. Idle hosts are switched to allow-power status to save energy [16]. VM selection policies are not implemented for an underutilization scenario. Consequently, VM migrations and hosts switch off are critical to saving energy and, at the same time, avoid unnecessarily huge migrations and limit host switches.

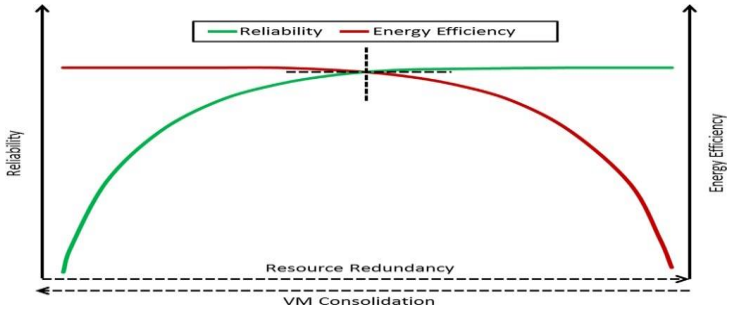

Figure 1. Reliability and Energy Efficiency trade-off in cloud computing systems.

One of the challenges of workload modeling in cloud computing is that cloud trace logs available for analysis are few. Google released a "cluster usage trace dataset" in November 2011 that stores wide machine resources and workload data proceedings on a huge production cluster [17]. The comparison study of GCD and gird trace log has been published in [18]. This study shows that computing jobs are a high priority. In additional, a machine is titled as idle machine [18], in a case the machine resources utilization is very full however 90% of utilization is attributed to low jobs. Therefore, this work will focus on high jobs which are the computing jobs of GCD. Google trace log data has some problems. For example, some data of cycle per Instruction (CPI), and memory accesses per instruction(MAI) are missed. In addition, the snapshots are taken every 5 minutes which has a bad effect on the prediction process. For example, the short prediction with a couple of time slots, for example, can be for 10 minutes. While the long prediction with six-time slots, for example, can be for 30 minutes.

This paper presents auto resource management (ARM) scheme for improving SLA and energy consumption in cloud computing, The ARM has three components; they are the threshold, VM selection policy, and VM placement host. For threshold, this paper proposes a static and dynamic threshold. For a static threshold, the paper presents a threshold extraction method. Our proposed extraction method is extended from previous work [19]; then we added K-means to extract the upper threshold. The normality of a cloud computing environment is dynamics; this paper proposes a new dynamic threshold method called Rate of Change (ROC). ROC calculates the threshold depending on the rate of change in resource utilization from past to current status. For VM selection policy, this paper presents a new VM selection policy called Minimum Utilization Non-Negative (MUN). MUN selects a VM such that reduces the number of VMs will need to migrate from this host. Also, this paper proposes a prediction model is named Short Prediction Resource Utilization (SPRU).The proposed prediction model depends on the Support Vector Regression (SVR) method. The SPRU involves the ARM scheme (ARM-SPRU) to make a

decision depending on the current and predicted resources utilization of hosts. With ARM-SPRU, the SPRU helps on detecting the overutilization and underutilization of hosts by predicting the host utilization in the near future. Also, SPRU helps to choose a placement host which will receive the migrated VMs. The joint use of current and predicted resource utilization allows the cloud providers to take decisions with better vision in the dimension of current and future. In this manner, this paper has four contributions:

1. For future resource utilization, the article proposes short prediction model for resources utilization in short term prediction.

2. For update static thresholds, the article presents an extraction static threshold method presents.

3. For the dynamic threshold, we suggest a new dynamic threshold method and comparison with published dynamic threshold methods.

4. For VM selection policy, we offer a new VM selection method and comparison with published VM selection methods.

The article is organized as follows: section 2 will discuss the related work. Section 3 consists a proposed prediction model, extraction static threshold value, dynamic threshold method, the conditions of over/under utilization host, VM selection policy, and VM placement host. Section 4 discusses the experiment design. The results and analysis will discuss in section 5. Conclusion and future work debate in section 7.

2. RELATED WORK

A number of researchers aim to improve reliability while others aim to reduce energy consumption. The goal of both researchers reduces the costs are paid by the user and increases the profit of data centers. Due to the large size of data centers, financial savings are substantial. In this section, a lot of the relevant proposed approaches for solving the tradeoff between reliability and energy consumption on cloud and distributed computing will be review. The proposed approaches classify into three groups: reactive management, proactive management, and hybrid management. For the reactive management group, the approaches depend on improving energy efficiency and SLA performance according to the current resources utilization of hosts as [12, 23]. The proactive management group, the approaches work depending on the predication resources utilization without care about the current resources utilizations. The hybrid management approaches depend on the combination of current and predicting resources utilization of hosts as [24, 25-28]. The vital advantage for the hybrid approaches group is reducing the unnecessary VM migrations which due to improving SLA performance and energy consumption.

For reactive approaches, in [12], the authors presented a competitive analysis study for improving energy efficiency and decreasing SLA violation, which causes reliability decay. This study compared the different VM selections policies with static and dynamic thresholds. For VM selection policies, the authors designed a set of VM selection policies: Random Selection (RS), Maximum Correlation (MC), the Minimum Migration Time (MMT), and Minimum Utilization (MU). Also, several dynamic thresholds have been developed and tested; Local regression (LR), Median Absolute Deviation (MAD), Interquartile Range (IQR), and Local Regression Robust (LRR).The experiments were executed over the PlanteLab VM workload trace. The experiments were evaluated by SLA, energy consumption and number of VM migrated. The results showed the role of the dynamic threshold to reduce the amount of energy and number of VM migrations. The

main shortage of this work is that the utilization of the workload implemented is low utilization. Therefore, is shown that the performance of different VM selection methods is the same. Also, the discussion of reliability performance is absent. An energy-aware task consolidation (ETC) technique is presented in [20]. The ETC minimizes energy consumption by restricting only a single resource (CPU) usage below a specified threshold. This work located a default CPU utilization threshold of 70% to show task consolidation management between virtual clusters. The simulation results demonstrated that ETC can valuable reduction for power consumption when managing task consolidation for cloud systems. The critical limitation of ETC is that ETC depended on CPU utilization only while ignoring other important resource utilization like memory and network resources. Also, cloud computing is a dynamic environment; ETC used a static threshold which is not acceptable with a cloud computing environment.

An automatic system that monitors tasks and detects hotspots called Sandpiper is presented in [21]. The sandpiper proposes a novel mapping of physical to virtual resources and commencing the necessary migrations. Also, the sandpiper combines three dimensions into a single volume metric as the product of CPU, memory and network utilization. The sand piper implements a black-box and gray-box approach based on a volume criterion. A black box and gray box (BG) sorts overloaded hosts based on their volume metric and the VMs in each host based on their volume-to-size ratio (VSR). BG chooses the host with the highest volume first. A VM which has the highest VSR will be migrated is the second step of the proposed approach. The BG scheme also approves the volume metric to choose target hosts, i.e., they are arranged in direction of growing volume to allocate. The major limitation of the BG scheme is that does not enable a reliable characterization of overloaded and under loaded hosts. The BG scheme also does not support migrating under loaded hosts an disrestricted to homogeneous physical machines. various VM selection policies have been established in [22]. A neural network classifier (NN), self-organizing map (SOM), and K-mean are used to select a VM to migrate. Neural networks are applied to classify virtual machines into three categories (high, middle, low), then forecast the category of the virtual machine depend on these features of the VM: current CPU utilization, CPU Utilization history mean, MIPs (Million Instructions Per Second), RAM. SOM is used to cluster input virtual machines into a set of groups of similar patterns according to virtual machine properties. The resulted VMs will be allocated on a grid representing their relative similarities in a graphical representation. The VM in migration will be selected according to the coordinates which reflects the VM properties. For k-means, the small or huge VMs utilization grouped together which due to simplifying the process of selecting VMs in the migration stage. The small or huge VMs utilization would be proposed to transfer from their host in sleep mode or not overloaded mode. The authors compared their VM selection policies and VM selection policies published in [12].The result showed that NN policy is the best on energy consumption, the number of migration, and reallocation VM mean time evolution. At the same hand, it is the worst case for SLA violations.

A Three-dimensional Virtual Resource Scheduling Method (TVRSM) for cloud computing energy reduction is suggested in [23]. the TVRSM consists of three phases: virtual resource allocation, virtual resource scheduling, and virtual resource optimization. The virtual resource allocation phase, based on the objective function and constraints, the resource scheduler locates all the VMs demanded by the customers to the appropriate PMs. According to the energy consumption model, in the phase of virtual resource scheduling, this phase uses the VM migration technology to migrate the VMs from the overload hosts to other hosts that have lower resource utilization or other idle hosts. In the virtual resource optimization phase, according to the resource utilization and energy consumption model, the resource scheduler improves the virtual resource s in the cloud data center. There sources scheduler requests the virtual resource optimization module, which migrates the VMs from the hosts with the smallest resources utilization too there

physical machines, and changes the original hosts to sleep mode? For detection overloading utilization hosts, they used the threshold proposed in [12] with updating the criterion of detecting overloading. The concept of the multi-dimensional load is proposed. The multi-dimension load depends on assigning a weight for CPU, memory, and network utilization. The results showed the major effects of TVRSM in energy consumption and the number of virtual machine migration. Nevertheless, the results ignored the performance of TVRSM with SLA.

For hybrid management approaches, a VM consolidation algorithm with multiple usage prediction (VMCUP-M) and minimum resource temperature (MRT) VM selection policy have been presented in [24]. VMCUP-M aims to improve energy efficiency. The multiple usage prediction methods (MUP) is the main core of VMCUP-M. An MUP refers to both the CPU and the memory resource and the horizon employed to predict long future utilization. In [24], a long term prediction applies to 6-time steps. The joint use of current and predicted utilization allows for a reliable characterization of over utilized and underutilized hosts, thus enabling cloud providers to increase their compliance with the SLA. For the MRT policy, it works by migrating a VM to reduce the resource temperature of a given host. While migration is expensive, the goal of MRT is to select only the VMs that contribute most to reduce the load of overloading host. The result of simulation on real-world workloads shown that the proposed MUP scheme can easily be integrated into existing VM selection, thresholds, and placement algorithms to increase the performance of a data center proposed in [12]. Also, the results shown VMCUP-M algorithm reduces energy consumption due to the decreased number of active hosts and VM migrations; thus achieving better compliance with the SLA than the state of the art. The two major limitations of the proposed approach are (1) the performance of all tested VM selection policies are similar which meaning that all VM migration is effective from underutilization hosts only. By another meaning that the workload used in simulations is low utilization resources, and (2) the paper depends on a long term prediction concept that is not suitable for the cloud computing environment. Because of this, 80% of the job lengths in the cloud are less than 1000 seconds [18]; the long term prediction is 1800 seconds for this work.

In [25], the authors presented a system that uses virtualization technology to allocate data center resources dynamically based on application demands and support green computing by optimizing the number of hosts in use. The concept of skewness is presented; it aims to measure the unevenness in the multidimensional resource utilization of a host. By minimizing skewness, a combination of different types of workloads improves the overall utilization of host resources. Also, the authors developed a set of heuristics policies that prevent overload in the system effectively with saving energy used. In addition, he authors proposed a short prediction mechanism based on the EWMA formula to forecast the CPU load on the DNS server in their university. They measured the load each minute and forecasted the load in the next minute. The limitation of this work is that the skewness concept may cause uncertainty for approaches when resources utilization data are large and do not add new features for the data. Another limitation is neglected detection of resources causing overloading where skewness mixed all utilization together.

The study in [26] proposes a Bayesian Network-based Estimation Model (BNEM) for a live VM migration (BN-VMC) method. A BNEM for dynamic VM migration has been presented by associating nine related features from various views with BN, as well as with the assistance of the superior ability of probability estimation and probabilistic reasoning by such BN methods. Depending on BNEM, the migration probability of the VMs deployed in different hosts with specific load patterns is estimated. The VM migration probability shows that the potential whole times of VM migrations are able to be predicted. The BN-VMC method consists of four phases. In the first phase, the estimation of the overload probability (EOP) algorithm is proposed to

determine if the hosts have overload risks. In the second phase, the migration and capacity aware migration selection (MCAMS) algorithm is proposed to migrate some VMs from the hosts with overload risks, as well as enable the overload alert condition to not be satisfied. Third, the BN-VMC method selects the VMs with low memory requests and a substantial effect on the potential times of VM migrations to leave. In the third phase, the migrated VMs are redeployed to the new destination hosts using a proposed algorithm called migration and power-aware best fit decreasing (MPABFD).Finally, contract the running hosts and migrate all VMs to under loaded ones and switch these hosts to sleep to decrease energy consumption. To turn off more hosts, the BN-VMC algorithm used the iteration strategy, thus switching hosts with the lowest resource utilization to sleep.

The work in [27] extended the work developed by [28] to scale up and down resources with a smarter approach to save energy consumption and try to lower the SLAs violation. It implements paddings as a security measure, on the number of resources forecast by the PRedictive Elastic Resource Scaling (PRESS) model. Also, it uses SLAs violations analysis to tune paddings. Moreover, it added a forecasting model that predicts the target host that receives migrated VMs from over utilization hosts to avoid service degradation. The discussed published paper applied linear regression, multiple linear regression, EWMA formula, and Bayesian model. This paper is first one applies the SVR to predict the CPU and memory utilization.

3. METHODOLOGY

In this section, the prediction model will discuss. Also, the proposed extraction static threshold value, the proposed method of dynamic threshold, VM selection policy, the condition of detection overutilization and underutilization host, and VM placement method.

3.1 Short Prediction Resource Utilization Model

A More accurate prediction approach can develop using machine learning techniques and data mining. Data mining and artificial intelligence techniques like support vector machine (SVM), decision trees, Genetic Algorithm (GA), Neural Networks (NNs), fuzzy logic, etc., have been well applied in corporate financial distress predicting. The SVM gained acceptance due to many attractive characteristics and good generalization performance on a wide range of challenges. Also, SVM symbolizes the structural risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle employed by conventional regression methods. SRM minimizes an upper bound of the generalization error, while ERM minimizes the error on training data. Newly, the application of SVM to time-series forecast called support vector regression (SVR) has also displayed many breakthroughs and acceptable performance for different challenges like financial market prediction [29], electric utility load forecasting applications [30], environmental parameter estimation [31], control system and signal processing prediction [32], and machine reliability forecasting [33,34,35]. One of the main features of SVR is that rather than reducing the observed training error, SVR aims to minimize the generalized error bound so as to get generalized performance. This generalized error bound is the combination of the training error and a regularization term that dominance the complexity of the hypothesis space.

The resource utilization prediction in cloud computing is classified to long prediction and short prediction. Numerical analysis for characteristics of cloud and grid computing has been published in [18].The previous numerical analysis concluded cloud tasks are quite shorter than Grid tasks where over 80% of Google task lengths are shorter than 1000 seconds. For gird, most of the jobs are longer than 2000 seconds. So, we support using a short prediction for the cloud computing environment. In this article, a short prediction model with 2 steps for resource utilization proposes. The proposed model depends on the SVR method. The proposed model is named short prediction resource utilization (SPRU). At the first time step, the proposed model predicts the utilization of resources using current time resource utilization as input. At the second time step, SPRU uses current and output of resources utilization of the first time step as input, then predicts utilization after 2 time steps. Table 1 summarizes SPRU prediction.

Table 1. SPRU Prediction Steps.

m step, *U(h)* host utilization, *d* resource (CPU, memory), *t* time,		
Step (m)	Input for SPRU	Output of SPRU
m= 1	$Ud(h)t$	$Ud(h)t$+1
m= 2	$Ud(h)t$+1	$Ud(h)t$+2

3.2 Static Threshold Extraction Method

In the cloud computing environment, the resources utilizations of hosts are changed based on applications requested by users. Also, the size of requested applications is changed according to the time of submitting during the day, if it is on week day or weekend, morning or night, and...etc. So, this paper proposes a method for the extraction of the upper static threshold value. The proposed method depends on the historical data of resource utilization of hosts and time. The static threshold is a hard limit of resource utilization. The proposed method is extended to our work published in [19]. The published work in [19] presented a prediction model based on a support vector machine (SVM) for each resource separately. The SVM is built on SRM principle rooted in the statistical learning theory. It provides better generalization capabilities, due to, the SRM achieves through minimizing the upper bound of the generalization error. The separated plan is created by the SVM model that has a set of support vector points. The support vector points determine the boundary of the plan using points from data used.

In this paper, K-means implemented to our previous work to cluster the support vector points in the side of update machines. The support points are clustered into three groups: high, middle, and low groups. The high cluster represents the overutilization hosts; the low cluster represents the underutilization hosts. The least value of resource utilization in the overutilization host group is considered the upper static threshold for the resource.

3.3 Dynamic Threshold Rate of Change (RoC)

The static thresholds are hard to apply for an environment with dynamic and various workloads like cloud computing, in which changed types of applications share a real resource. The system should be able to automatically update with the behavior of workload requested based on the workload patterns exhibited by the applications. Therefore, a dynamic threshold capable of auto modified the utilization thresholds based on a preceding of the historical data composed during the lifetime of VMs is proposed.

The main idea of the proposed dynamic threshold is to adjust the value of the upper utilization threshold depending on the rate of change of resource utilization. The proposed dynamic threshold is named Rate of Change (RoC). The RoC implements through two stages. At the first stage, the rate of change between the Previous utilization and current utilization calculates then multiples on the threshold of previous, then adds the previous threshold. Equation 1 shows the previously calculated amount. The RoC calculates in the second stage. The RoC calculates as the difference between the current utilization and safety parameter multiples by the difference between the current utilization and the output of equation 1. The RoC can calculate at any time slots except the first slot time where the pervious utilization equal zero. At the first time slot, the static threshold applies. Equation 2 explains how RoC is calculated in the second phase. Roc is most like as cover for host utilization.

$$T^d = T^d_{t-1} + T^d_{t-1} * \frac{U^d_t(h) - U^d_{t-1}(h)}{U^d_{t-1}(h)} \quad (1)$$

$$RoC_d(h) = U^d_t(h) - s * \left(U^d_t(h) - T^d\right) \quad (2)$$

Where $Roc(h)$ is the new threshold at stating time, $U^d(h)$ is the host utilization for resource d at time t, s is a safe parameter equal to 0.1, $U^d(h)$ is the host utilization for resource d at time t+1, $Roct^d(h)$ is the new threshold for host h for current time t, T^d is the threshold of resource d, T^d is the threshold of resource d, at time t-1.

3.4 Overutilization Host Detection Condition under SPRU (OHD-SPRU)

The SPRU model improves the detection of over utilization host by detecting the host not only according to the current utilization of resources but also according to the predicted resource utilization of the host using the SPRU model. In other words, the detection process is done by a broader perspective for host resource utilization, which develops a chance for better resources management and reduces the number of VM migrations. If the host is over utilized in the current status and will be over-utilizing in the near future, this host will need to migrate one of VMs are running over it. The host is considered over utilized for resourced, if the following condition is satisfied:

$$U^d_t(h) \geq T^d \ \& \ U^d_{t+1}(h) \geq T^d \quad (3)$$

Where U(h) is the host utilization for resource d at time t, $U^d(h)$ is the host utilization for resource d at time t+1, T^d is the threshold of resource d.

The reminder states for the hosts are (1) high utilization currently and low utilization in the future, (2) low utilization currently and low utilization in the future, and (3) low utilization currently and high utilization in the future. For the second state, the host's utilization will decrease, therefore the migration will be not required. For the last state, the host's utilization will be increased but the host's utilization might go down again in the near future.

3.5 Underutilization Host Detection Condition under SPRU(UHD-SPRU)

In the next step of overutilization host detection, the cloud manager starts collecting running VMs in less of the hosts without causing overutilization again for hosts receiving VMs. With the SPRU model, for the underutilization host, the current resources utilization and future resources utilization are equal OR lower. The next equation explains the condition of detecting the underutilization host.

$$U_t^d(h) \leq U_{t+1}^d(h) \quad (4)$$

3.6 VM Selection Policy

After the overutilization host is detected, the resource which causes over utilizing becomes the source of risk due to host failing. To decrease the host utilization, selecting a VM the overutilization host is required to apply. A new VM selection policy called Minimum Utilization Non-Negative (MUN) proposes. The policy aims to reduce the critical resource utilization on the host by migrating the selected VM. While the migration cost is expensive, the objective of the proposed policy reduces the number of VM migrated, by selecting VM that has a higher impact on reducing utilization of critical resources on the host The proposed The policy has two steps. In the first step, the difference between the host utilization and threshold of critical resource calculates as shown in equation 5. In the second step, for resource causing overutilization, the resource utilization of each VM is subtracted from the difference between host utilization and threshold. The VM with a minimum non-negative result is selected. If a cloud computing system has *m* hosts (h), each host has dimension *d* of resource *r*. The policy selects a VM that satisfies the following condition of equation 6.

$$MUN(h^d) = U_t^d(h) - T^d \quad (5)$$

$$find\left(\min\left(MUN(h^d) - U^d(vm)\right)_{nn}\right) \quad (6)$$

Where U(h) is the host utilization for resource d at time t, T^d is the threshold of resource d, U^d (vm) is the virtual machine utilization for resource d, and nn denotes for non-negative values.

3.7 VM Placement under SPRU (PABFD-SPRU)

The SPRU will have a vital role in determining the target host. The power-aware best fit decreasing (PABFD) algorithm published in [12] for multiple resource VM placement extents, which focuses on reliability and energy efficiency. The SPRU model involved in PABFD (PABFD - SPRU) aims to get a new placement for migrated VM by choosing the target host that is not over-utilizing now and in the near future. PABFD-SPRU works for reducing the probability of the target host that becomes over utilized which reduces the number of VM migration in the near future.

According to the combination of OHD-SPRU, UHD-SPRU, and PABFD-SPRU, the proposed auto resource management scheme will be depending on future predications by the SPRU model in addition to the current status of cloud computing. The developed scheme is called auto resource management with SPRU (ARM- SPRU). For the ARM, the overutilization and underutilization hosts are detected depend on only the current resources utilization.

4. EXPERIMENTAL DESIGN

To gauge the efficiency of our proposed approach in a practical cloud scenario, the CloudSim simulation toolkit has been used. The most popular simulation tool for the cloud computing environment is CloudSim. GridSim is the basic core of CloudSim. The Base programming language for CloudSim is Java. CloudSim has advantage of open source platform, so it is easy to extend based on Java.

4.1 Experiment Setup

For hardware, the experiment environment setup used in our model, a data center is considered heterogeneous where a data center is consisting of two types of hosts; the total number of hosts is 800. The data center has 800 hosts. Half of them is HP ProLiant ML110 G4 servers, and a reminder of them is HP ProLiant ML110 G5 servers. HP ProLiant ML110 G4 servers a configuration with1, 860 MIPS per core. HP ProLiant ML110 G5 servers has configuration with 2,660 MIPS per core. all servers have 2 cores, 4 GB of memory and 1 GB/s of network bandwidth. The Standard Performance Evaluation Corporation (SPEC) [36] has been followed to derive the power consumption of active2 servers in the simulation. Table 2 recaps a data center specification. Four virtual machines with a various specification are used following Amazon EC2 instances [23,37]: High-CPU Instance (2500 MIPS, 0.87 GB), Extra Large Instance (2000 MIPS, 1.74 GB), Small Instance(1000 MIPS, 1.74 GB), and Micro Instance (500 MIPS, 613 MB). At the first step of the simulation, VMs are distributed according to the resource requirements defined by the VMs. Table 3 summarizes virtual machine configurations.

Table 2: Data Center Configuration.

Host	MIPS	Number of Cores	Memory	Bandwidth
HP ProLiant ML110 G4	1860	2	4 GB	1 GB /s
HP ProLiant ML110 G5	2660	2	4 GB	1 GB /s

Table 3: Virtual Machine Configuration.

VM Type	MIPS	Memory
Extra Large	2000	1.74 GB
High-CPU	2500	0.87 GB
Small	1000	1.74 GB
Micro	500	613 MB

For software, real-world workloads represented VM utilization. Google Cluster Data (GCD) dataset for a 29-day period in May 2011 has been used as an example for a real workload in this work. The GCD workload has 670983 jobs, and each job has one or more tasks. The total number of tasks is 144841618. The GCD workload contains the normalized value of the average number of used cores and the utilized memory. The CPU and the memory utilization of VMs have been created such that the tasks of each job was combined by summing their CPU and memory consumption every five minutes in a period of 24 hours. In this work, we focused on the computing jobs which have high priority. So, computing jobs have been extracted from GCD workload. The computing jobs are high resources consumption in GCD contrast from other jobs [18]. The utilization of CPU and memory of the computing job has been filtered by 7% to 100%, produces a total of 1278 VMs. Table 4 summarizes the characteristics of the extracted workload.

4.2 Evolution Metrics

The SPRU evaluated with Mean Square Error (MSE), Squared Correlation Coefficient (R2), and Percentage of Predictions (PRED 25)) metrics. For R2 and PRED25, the values closer to 1 are better. For MSE, the value closest to zero is better. In the ARM scheme, every 5 minutes over 24 hours which is the system period, the resource utilization measured. The proposed scheme evaluates by the following metrics:

1.Number of active hosts for 24 hours.
2.Energy Consumption for all hosts during simulation time.
3.A number of VM migration.
4.Average SLA volition.

Table 4: Workload Trace Characteristic.

Workload	VMs	Resource	Mean (%)	St. dev (%)	Median (%)
GCD	1278	CPU	32.41	2.66	29.98
		Memory	17.34	4.93	11.44

The power of hosts has two parts. They are; the maximum power of the hosts and CPU utilization of hosts. The energy consumption defines as the difference between host powers for time cascaded. The power and energy consumption figured as shown in (7) & (8)[38].

$$PM_i(t) = k * PM_{i,max} + (1-k) * PM_{i,max} * U_{i,cpu}(t) \quad (7)$$

$$E = \int_{t_0}^{t_1} P_i(t)\, dt \quad (8)$$

Where $PM_{i,max}$ is the maximum power consumption of host i,k is the fraction of power consumption when the host i is in idle state and $U_{i,cpu}(t)$ is the CPU resource utilization of the host on time t. E is the energy consumed by host i from start time t0 to end time t_1.

It is required to explain a workload independent metric that able to use for evaluating the SLA delivered to any VM deployed. SLA violation level consists of two compounds [12]; (1) the percentage of the time, which is the active hosts have the CPU utilization of 100%, called SLA violation Time per Active Host (SLATAH). SLATAH calculates as appeared in (9). (2) the overall performance degradation by VMs migrations, Performance Degradation due to Migrations (PDM). PDM is computed as shown in (10). If a host is experiencing 100% utilization, the performance of the applications is limited by the free host capacity; therefore, VMs are not being provided with the necessary performance. So, monitoring the SLATAH is required SLA violation is defined as shown (11).

$$SLATAH = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{si}}{T_{ai}} \quad (9)$$

$$PDM = \frac{1}{M}\sum_{j=1}^{M}\frac{C_{dj}}{C_{rj}} \quad (10)$$

$$SLA = SLATAH * PDM \quad (11)$$

Where N is the number of hosts, Tsi is the total time during which the host experienced the utilization of 100% leading to an SLA violation, Tai is the total of hosts in active mode. M is the number of VMs, Cdj is the estimate of performance degradation of the VM caused by migration, Crj is the total CPU capacity requested by VM during its lifetime.

5. RESULTS & ANALYSIS

In this section, the experiments of results will discuss. The next experiments classified into three classes. Firstly, the analysis of SPRU will be discussed. Secondly, the performance of the proposed auto resource management (ARM) will be discussed and with published VM selection policies in [12] and our proposed selection method MUN under the static threshold (ST), LR, and RoC dynamic thresholds. Also, the parametric analysis between the performance of LR and RoC thresholds will be discussed. Thirdly, the ARM-SPRU will be discussed in the same manner of previous experiments to show the effect of the performance as SPRU on our proposed auto resources management. The second class will be evaluated under energy consumption, the number of VM migration, average SLA volition, and a number of the active host during simulation lifetime. For the third class, The ARM-SPRU will be assessed by comparing its performance with the results of the first-class experiment. By applying our extraction threshold method on GCD trace, the static threshold (ST) for CPU and memory are 0.8290, 0.7651 respectively. LR is selected because of LR is outperforms IQR, MAD, and LRR, which are published in [12] according to energy consumption and average SLA Violation.

5.1 Short Prediction Resource Utilization Model Analysis

The extracted jobs trace data form GCD is divided into 70% for training and 30 % for testing. Table 5 shows the evaluation metric results of SPRU. It is observed from table 5 that, with increasing steps, the performance of SPRU improves. The main reason for improving SPRU is that SVR gets better results with higher dimensions (more features) as SVM; At step two, SPRU has two inputs (features) rather than only one input (feature) likes in the first step. Figures 2 & 3 show examples of the performance of SPRU for CPU and memory resources at the first and the second steps. From Figures 2 &3, it is witnessed that the SPRU model satisfies suitable similarity between observed and predicted resource utilization during simulation life time. At the first time for second step prediction, we observed that the performance of SPRU is low due to the weak correlation between features and, with time increases, the correlation between features improves.

5.2 Auto resource management (ARM) Performance Analysis

In the two next sections, the performance of VM selection policies will be assessed under the static threshold, LR, and RoC dynamic threshold according to the current status of the host's resources utilization.

5.2.1 Static Threshold

This section analyses the performance of the proposed VM selection policy MUN compared with the performance of RS, MU, MC, and MMT policies under the static threshold (ST). Figures 4a, 4b, 4c, and 4d show that our proposed policy MUN has the best performance in energy consumption, the number of VM migrations, average SLA violation, and high stability for the number of active hosts compared with other policies. The main advantage of MUN is that works to reduce the number of VM migrations during simulation time. The MUN depends on selecting a VM which has a value of resource utilization causing overutilization for subscribed host equal to the difference between resource the threshold and host utilization currently (for example, if the host has VMs with CPU utilization 0.4, 0.3, 0.2 and host is over threshold by 0.18, MUN will select VM with 0.2 CPU utilization).

According to figure 4b, MUN has the least number of VM migrations, so the value of PDM becomes lower which has an effect on SLA violation. For energy consumption, the rate of change in power becomes lower which helps in reducing energy consumption according to equation 7. The number of VM migration for MUN is less than RS by 23%, where RS is the second-ranked policy according to the performance. Also, MUN policy has the least number and the most stable active hosts during all lifetime of simulation.

Table 5. Performance evaluation for SPRU model.

Resources	CPU		Memory	
Step	One Step	Two steps	One step	Two steps
MSE	0.00586	0.00485	0.00232	0.00212
PRAD (25%)	0.984	0.984	0.99	0.99
R^2	0.925	0.942	0.931	0.933

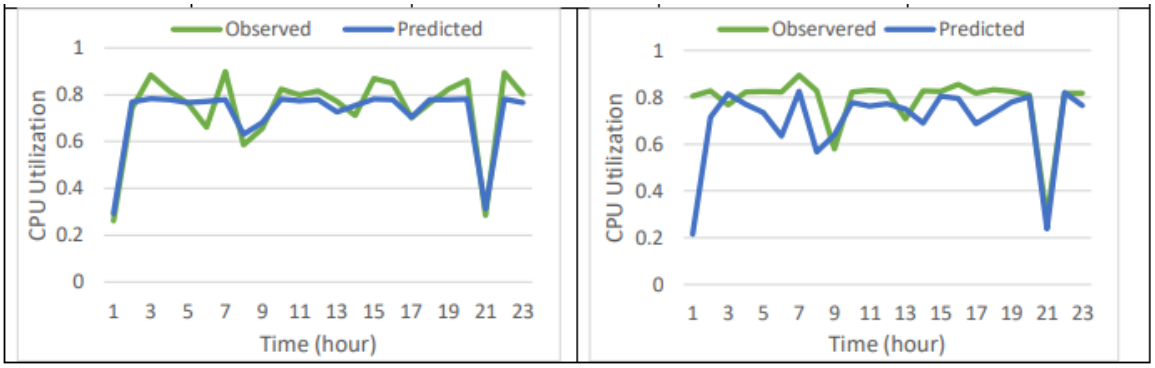

Figure 2. CPU Utilization with Simulation Time for one and two steps.

5.2.2 Local Regression Threshold

The dynamic threshold is more suitable for a cloud computing environment because cloud computing is a dynamic environment. The comparison analysis of our proposed MUN with other policies will be discussed within the LR dynamic threshold. The main idea of LR is local regression which helps on fitting a trend polynomial to the last number of observations of resource utilization. The degree of the polynomial is one to reduce the bias at the boundary. Figures 5a, 5b, 5c, and 5d explain the performance of different VM selection policies with LR. The MUN has the best performance compared to other policies with LR threshold. In addition, LR significantly outperforms the static threshold. With comparing the results of MUN with LR (MUN-LR) and the results of MUN with static threshold (MUN-ST), the figures discover that the energy consumption of MUN-LR is less than MUN-ST by 5.4%, the number of VM migration of MUN-LR less than MUN-ST by 40.7%, and SLA violation for MUN-LR is less than MUN-ST by 1.8%.

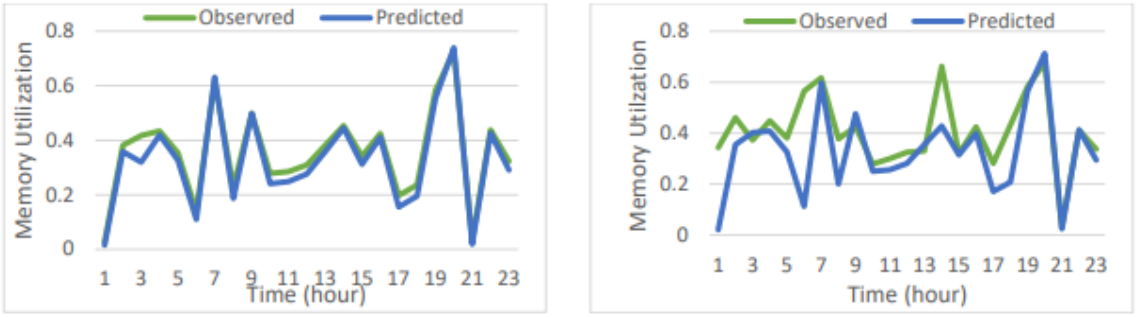

Figure 3. Memory Utilization with Simulation Time for one and two steps.

Fig. 4. Performance of various VM Selection Policies for GCD under Static Threshold: (a) Energy Consumption; (b) Number of VM Migration; (c) Average SLA violation; (d) Number of Active host.

The most reduction is done by MUN-LR where the main objective of MUN decreases the number of migration for each time slot. Also, the energy consumption reduces by a higher ratio than SLA violation for MUN-LR due to MUN policy. With decreasing the number of VM migration, the rate of power change becomes low, and the rate of power change is only effected on energy consumption as explained in equation On the other hand, SLA violation depends on both PDM and SLATAH.PDM is affected by decreasing the number of VM migration, but SLATAH is related by host resources utilization which it does not reduce by a high ratio like PDM reduction. Figure 4d shows the number of active hosts during the simulation lifetime. It notes that the behaviour of all VM selection policies is more stable than the behaviour of policies under the static threshold. The stable behavior of policies due to the rate of power change is limited, also, the number of VM migrations is the lowest.

5.2.3 Rate of Change Threshold

From the previous experiment of LR with different VM selection policies, it is clear that the dynamic threshold has a great effect on reducing energy consumption and the average SLA violation. Also, the difference between the performances of VM selection policies under LR became less than compared with the performance of the same policies under a static threshold. For energy consumption, the difference between the high and low VM selection policies is 15.89 kWh for the LR and 65.47 kWh for the static threshold. For SLA violation, the difference between the high and low VM selection policies is 0.25% for the LR and 0.34% for the static threshold.

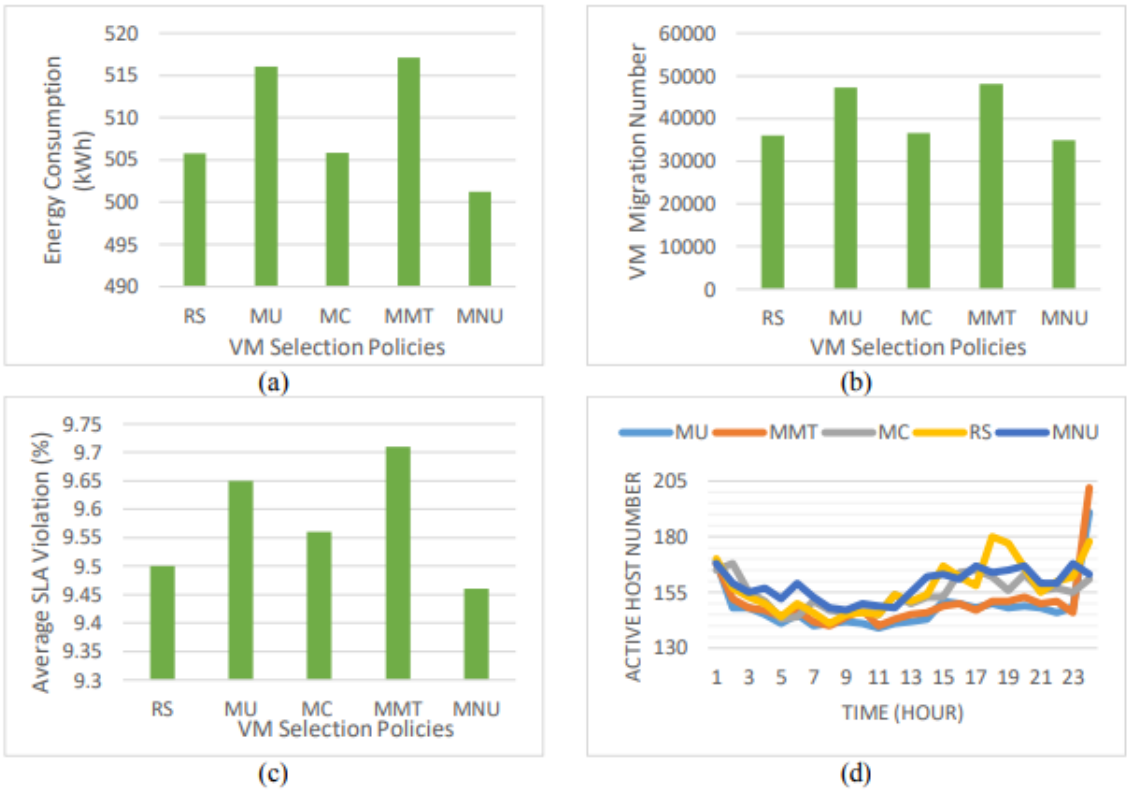

Fig. 5. Performance of various VM Selection Policies for GCD under LR Threshold: (a) Energy Consumption; (b) Number of VM Migration; (c) Average SLA violation; (d) Number of Active host.

The previous results were a stimulus to enhance the reduction of energy consumption and average SLA violation by designing a new dynamic threshold RoC. Figures 6a, 6b, 6c, and 6d display the performance of RoC for energy consumption, the number of VM migration, the average SLA violation, and the active host during the simulation lifetime. RoC has a great impact on reducing energy consumption and the SLA violation. In addition, the performance of all VM selection policies becomes similar. This similarity helps to apply any one of them with RoC. For all policies, the energy consumption reduces by 21% ~ 18.6% from the energy consumption with LR, and the SLA violation reduces by 13.7% ~ 11.4%.

The energy consumption reduction is higher than the average SLA reduction because of the energy consumption is only affected by VM migrations which cause power change of hosts. At the same time, SLA violation is affected by VM migrations degradation and the size of times that hosts become full utilization. In Figure 6d all VM selection policies have the same number of active hosts, which supports the similarity of the performance of policies in energy consumption, and SLA violation. The core reason for enhancing energy consumption and SLA violation is that RoC is more sensitive than LR. Because RoC is calculated By only the last utilization value , LR is calculated by the last 10 utilization values. In the following sections, the ARM-SPRU performance will be discussed for VM selection policies under both static and dynamic thresholds.

5.3 ARM-SPRU Performance Analysis

In this section, the SRPU prediction model will be evaluated. The good performance of SPRU is the motive to attach SPRU with an auto resource management (ARM) scheme to become ARM-SPRU. The main objective of ARM-SPRU optimizes the minimization of energy consumption and the average SLA violation. The ARM-SPRU will be analyzed for all VM selection policies under both static and dynamic thresholds.

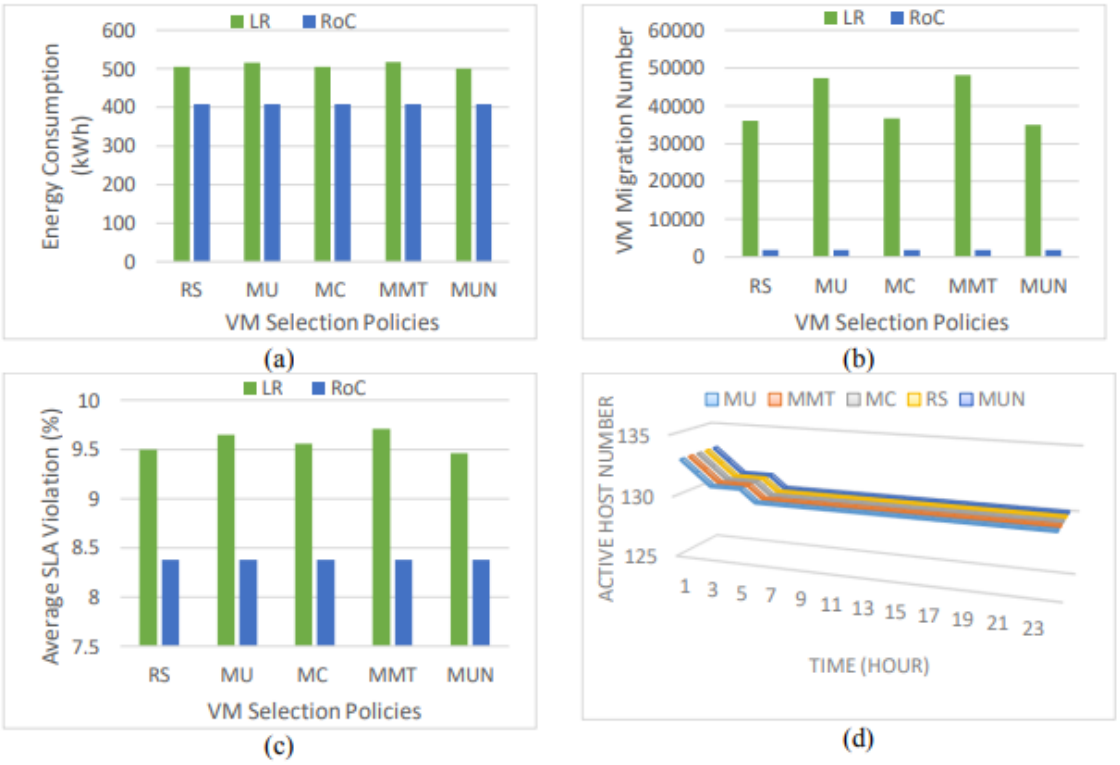

Fig. 6. Performance of various VM Selection Policies for GCD under RoC Threshold: (a) Energy Consumption; (b) Number of VM Migration; (c) Average SLA violation; (d) Number of Active host.

5.3.1 Static Threshold

The overutilization detection with SPRU defines the host as over utilizing if the current and the future host resources are over utilizing. Under utilization host detection with SPRU defines as the future resources utilization of host is equal to or less than the current resources utilization. Additionally, SPRU has the role to decide the target host by making sure that the target host will

not be over-utilizing after receiving the migrated VMs. Figures 7a, 7b, 7c, and 7d show the performance of ARM-SPRU compared to the ARM scheme over the energy consumption, the number of VM migration, the average SLA violation, and the active host's number. The following figures show the great effect of SPRU on an ARM scheme. The VM migration reduces by 95.5% ~ 97.5% for all policies. The VM migrations decrease has a vital impression on energy consumption and the SLA violation. Figure 7a shows that energy consumption reduces by 33% ~ 24 % for all VM selection policies.

The average SLA violation reduces by 13% ~ 11.7 % for all policies. In ARM-SPRU, all policies have the same performance in energy consumption due to PABFD–SPRU. Once the underutilization hosts are detected, the PABFD–SPRU chooses target hosts that are not going to over utilizing in the future, which helps target hosts to have high stability for power change in the rest of simulations. Figure 7d supports our explanation that the most effective reduction in the number of active hosts a red one in starting the simulation when the major VM migrations for saving energy are done. The average SLA reduces as a result of decreasing PDM who produces from decreasing the number of VM migrations. Specifically, our proposed VM selection policy MUN then RS has higher reduction SLA values rather than other policies.

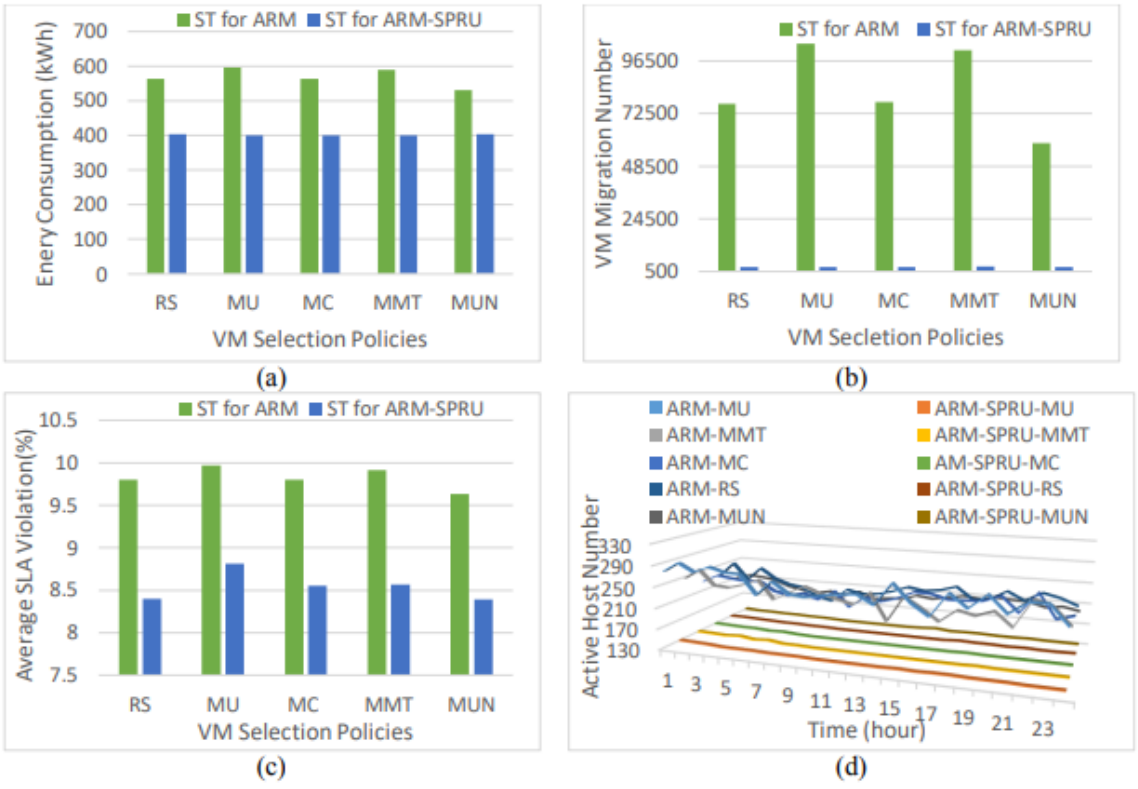

Fig. 7. Performance of various VM Selection Policies for GCD under Static Threshold: (a) Energy Consumption; (b) Number of VM Migration; (c) Average SLA violation; (d) Number of Active host.

5.3.2 Local Regression Threshold

According to the good performance of LR in the ARM scheme, LR will be tested with the ARM-SPRU scheme. Figures 8a, 8b, 8c, and 8d explain the performance ARM-SPRU under the LR threshold compared with the ARM scheme over the energy consumption, the number of VM

migration, the average SLA violation, and the number of active hosts. The ARM-SPRU succeeds to get higher minimization for the energy consumption and the SLA violation than energy consumption and the average SLA violation used ARM scheme for LR threshold. SPRU benefits the ARM scheme under the LR by performing a good prediction for the host's resource utilization which is due to reduce the number of VM migrations. The reduction of VM migration is mainly affected by reducing energy consumption and the average SLA violation. Also, the good performance for SPRU helps to reduce the number of active hosts compared to the number of active hosts in ARM. The energy consumption and the average SLA violation reduces by 18.7% ~ 21.3%, 21% ~ 18.6% respectively. Specifically, MUN has a low SLA violation compared with Other policies. Figure8a shows that energy consumption is constant for all VM selection policies. The main reason for constant energy consumption is that the rate of change power is constant as supported by figure 8d where the active host's number is similar for all VM selection policies with ARM-SPRU.

5.3.3 Rate of Change Threshold

Figures 9a, 9b, 9c, and 9d explain the performance of ARM-SPRU with RoC over the energy consumption, the number of VM migration, the average SLA violation, and the number of active hosts. For figure 9a, energy consumption is the lowest overall previous experiments. By another meaning the performance of

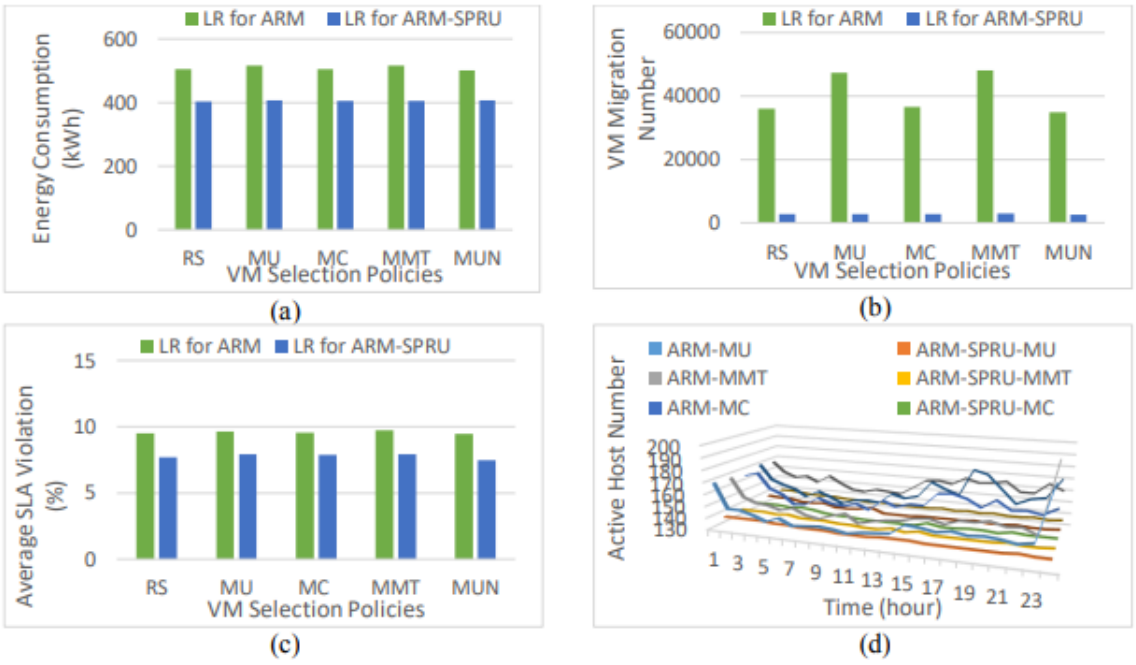

Fig. 8. Performance of various VM Selection Policies for GCD under LR Threshold: (a) Energy Consumption; (b) Number of VM Migration; (c) Average SLA violation; (d) Number of Active host.

ARM-SPRU with RoC threshold is the best from the perspective of energy consumption. On the other hand, the higher reduction of energy leads to an increased number of VM migrations due to the increase in the average SLA violation. Figure 9d observes that the number of active hosts with the ARM scheme is less than with ARM-SPRU. analysis energy consumption by more details for this case needs to study, the number of VM migrations for the MUN every 300 seconds will explain. Figure 10 demonstrates that the ARM-SPRU makes 3X VM migrations has been

performed by the ARM at the starting period. This huge migration contributes to reducing energy consumption, but at the same time causes increasing VM migration degradation which is reflected in increasing the SLA violation. The SLA violation increases by 0.73%. The core reason for the huge migration is that RoC is more sensitive than LR.

6. CONCLUSION & FUTURE WORK

In this article, the tradeoff between reliability and energy consumption in cloud computing has been addressed. To reach to the balanced status between the reliability and the energy consumption, an ARM proposed. Also, the optimization of ARM performance is studied by supporting the ARM with a prediction model called SPRU. For both the ARM and the ARM-SPRU, the analysis and comparison of a new VM selection policy MUN and the RoC dynamic threshold have been shown. The performance of ARM and ARM-SPRU has been evaluated by energy consumption, the number of VM migration, the average SLA violation, and the active host's number.

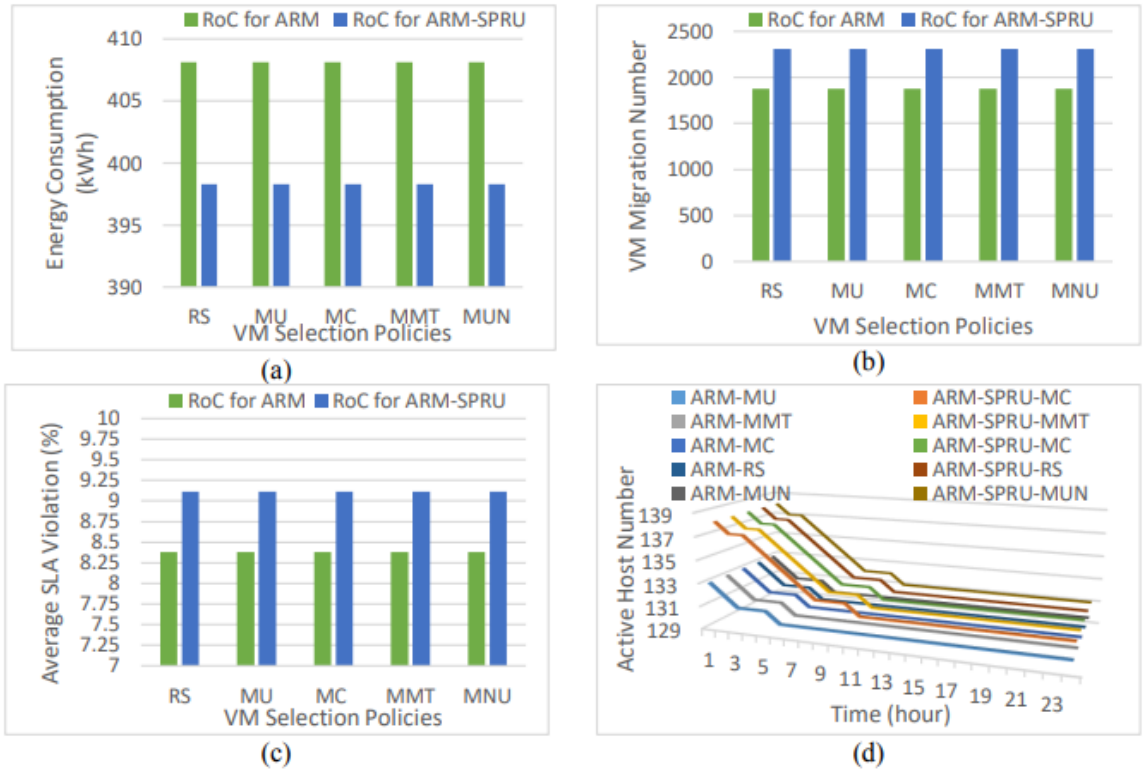

Fig. 9. Performance of various VM Selection Policies for GCD under RoC Threshold: (a) Energy Consumption; (b) Number of VM Migration; (c) Average SLA violation; (d) Number of Active host.

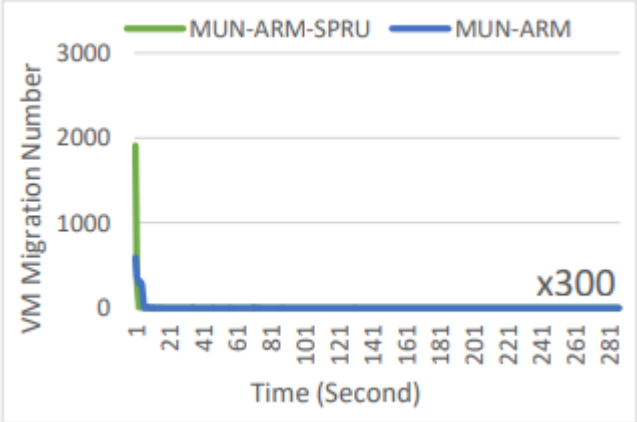

Fig. 10. The MUN Performance over Number of VM Migration.

The analysis of ARM explained that our proposed VM selection policy MUN outperforms other policies under the static threshold. For the dynamic threshold, RoC has better performance than LR and static threshold in perspective of the energy consumption and the SLA violation. For the ARM-SPRU, the SPRU has the ability to reduce the energy consumption and the average SLA violation better than the ARM scheme under the static threshold. Also, the analysis of the ARM-SPRU under the dynamic thresholds provides that the performance of LR is better than the RoC on the reliability perspective. Whereas, the performance of RoC is better than the LR on the energy consumption perspective. Table 6 recaps the Outcomes of the performance analysis of the ARM and the ARM-SPRU. The sign (-) mentions that any VM selection policy can choose to apply. Because all policies have the same performance for energy consumption and the SLA violation. The sign (√) refers to the best threshold that can apply with the ARM or the ARM_SPRU. For future work, the proposed plan is to test the performance of ARM and ARM-SPRU in real cloud computing systems. Additionally, the proposed plan includes designing a method that aims to improve the reliability of cloud computing on the scheduling stage.

Table. 6 The Summary of the Performance analysis of The ARM & ARM-SPRU.

Scheme	**Threshold**	**The Best VM Selection Policy**		**The Best Threshold**	
		Energy Consumption	SLA Viol.	Energy Consumption	SLA Viol.
ARM	ST	MUN	MUN		
	LR	MUN	MUN		
	RoC	-	-	√	√
ARM-SPRU	ST	-	MUN, RS		
	LR	-	-		√
	RoC	-	-	√	

REFERENCES

[1] Linlin Wu, Saurabh Kumar Garg, Steve Versteeg, and Rajkumar Buyya, "SLA-Based Resource Provisioning for Hosted Software-as-a-Service Applications in Cloud Computing Environments", IEEE TRANSACTIONS ON SERVICES COMPUTING, Vol. 7, No. 3, JULY-SEPTEMBER 2014, pp. 465-485.

[2] Deborah Magalhães, Rodrigo N. Calheiros, Rajkumar Buyy, Danielo G. Gomes," Workload modeling for resource usage analysis and simulation in cloud computing", Journal of Computers and Electrical Engineering, Vol. 47, 2015, pp. 69–81.

[3] S. Bhardwaj, L. Jain, and S. Jain, "Cloud computing: A study of infrastructure as a service (IaaS)," International Journal of Engineering and Information Technology, vol. 2, 2010, pp. 60–63.

[4] Yogesh Sharma, Bahman Javadi, Weisheng Si, Daniel Sun, "Reliability and energy efficiency in cloud computing systems: Survey and taxonomy", Journal of Network and Computer Applications, Vol.74, 2016, pp .66–85.

[5] Christian Engelmann, Al Geist, "Super-Scalable algorithms for computing on 100,000 processors", In the processing of the 5th international conference on Computational Science, Vol. 1,2005, pp.313-321.

[6] Ponemon Institute, "Cost of Data Center Outages", 2016, pp 1-21.

[7] Steven L. Sams, "Discovering hidden costs in your data center – a CFO perspective", 2011, pp 1-4.

[8] P. Getzi Jeba Leelipushpam, Dr. J. Sharmila," LIVE VM MIGRATION TECHNIQUES IN CLOUD ENVIRONMENT – A SURVEY", in the processing of IEEE Conference on Information and Communication Technologies (ICT), 2013, pp.408-413.

[9] A.-C. Orgerie, M. D. D. Assuncao, and L. Lefevre, "A survey on techniques for improving the energy efficiency of large–scale distributed systems," ACM Computing Surveys, vol. 46,2014, pp. 1–31

[10] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," IEEE Transactions of Parallel and Distributed Systems, vol. 24, 2013, pp. 1107–1116.

[11] T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J.-M. Pierson, A. V. Vasilakos, "Cloud computing: Survey on energy efficiency," ACM Computing Surveys, vol. 47, 2014, pp. 1–36.

[12] A. Beloglazov and R. Buyya, "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, pp. 1397–1420.

[13] Anton Beloglazov and Rajkumar Buyya, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers", in proceeding of the 8th International Workshop on Middleware for Grids, Clouds and e-science, 2010, No.4.

[14] Rajyashree, Vineet Richhariya," Double Threshold Based Load Balancing Approach by Using VM Migration for the Cloud Computing Environment", International Journal of Engineering and Computer Science, Vol.4, No.1, 2015, pp. 9966-9970.

[15] Shivani Gupta, Damodar Tiwari, Shailendra Singh, "Energy Efficient Dynamic Threshold Based Load Balancing Technique in Cloud Computing Environment", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6, No.2, 2015, pp. 1023-1026.

[16] I. Takouna, E. Alzaghoul, and C. Meinel, "Robust virtual machine consolidation for efficient energy and performance in virtualized data centers," in The IEEE International Conference on Green Computing and Communications (GreenCom), September 2014, pp. 470–477.

[17] Traces of Google Workloads, http://code.google.com/p/ google cluster data, 2015.

[18] Sheng Di, Derrick Kondo, Walfredo Cirne," Characterization and Comparison of Cloud versus Grid Workloads", IEEE International Conference on Cluster Computing, 2012, pp. 230-238.

[19] Moataz H. Khalil, Walaa M. Sheta, Adel S. Elmaghra by, "Categorizing hardware failure in large scale cloud computing environment", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2016.

[20] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-box and gray-box strategies for virtual machine migration," in The 4th USENIX conference on Networked systems design and implementation, 2007, pp. 229–242.

[21] Hend A.Selmy, Yousra Alkabani, Hoda K. Mohamed, "Energy Efficient Resource Management for Cloud Computing Environment", 9th International Conference on Computer Engineering & Systems (ICCES),2014, pp. 415-420.

[22] Wei Zhu, Yi Zhuang, Long Zhang, "A three-dimensional virtual resource scheduling method for energy saving in cloud computing", Journal of Future Generation Computer System, Vol. 69, 2017, pp. 66-74.

[23] Nguyen Trung Hieu, Mario Di Francesco, Antti Yl¨ a-J¨a¨ aski, " Virtual Machine Consolidation with Multiple Usage Prediction for Energy-Efficient Cloud Data Centers", IEEE TRANSACTION ON SERVICES COMPUTING, VOL. PP, NO. 99, May 2017, pp. 1-14.

[24] Sheng Di, Derrick Kondo, Walfredo Cirne," Characterization and Comparison of Cloud versus Grid Workloads", IEEE International Conference on Cluster Computing, 2012,pp.230-238.

[25] F. Farahnakian, P. Liljeberg, and J. Plosila, "Lircup: Linear regression based cpu usage prediction algorithm for live migration of virtual machines in data centers," in The 39th Euromicro Conference Series on Software Engineering and Advanced Applications, September 2013.

[26] Zhihua Li, Chengyu Yan, Xinrong Yu, Ning Yu, "Bayesian network-based Virtual Machines consolidation method", Future Generation Computer Systems, Vol.69, 2017, pp. 75-87.

[27] Shen Z, Subbiah S, Gu X, Wilkes J.," Cloudscale: elastic resource scaling for multi- tenant cloud systems", In Proceedings of the second ACM symposium on cloud computing, ACM, 2011.

[28] Gong Z, Gu X, Wilkes J, "Press: Predictive elastic resource scaling for cloud systems", In Processing of International Conference on Network and Service Management (CNSM) IEEE, 2010 ; pp. 9–16.

[29] F. E. H. Tay and L. J. Cao, "Application of support vector machines in financial time series forecasting" The International Journal of management Science Omega, Vol. 29, 2001 pp. 309–317.

[30] W. Lu, W. Wang, A. Y. T. Leung, S.-M. Lo, R. K. K. Yuen, Z. Xu, and H. Fan, "Air pollutant parameter forecasting using support vector machines", In Processing of International Joint Conference on Neural Networks (IJCNN '02), May 12–17, Vol. 1 2002, pp. 630–635.

[31] Gustavo Camps- Valls, Lorenzo Bruzzone, José L. Rojo-Álvarez, Farid Melgani," Robust Support Vector Regression for Biophysical Variable Estimation from Remotely Sensed Images", IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, Vol. 3, No. 3, JULY 2006, pp. 339-343.

[32] J. A. K. Suykens, J. Vandewalle, and B. De Moor, "Optimal control by least squares support vector machines", Journal of Neural Networks, Vol. 14, No. 1,2001, pp. 23–35.

[33]J. Yang and Y. Zhang," Application research of support vector machines in condition trend prediction of mechanical equipment", In Processing of International Symposium on Neural Networks (ISNN 2005), May 30–June 1, Vol. 3498,2005 pp. 857–864.

[34] W.-C. Hong, P.-F. Pai, C.-T. Chen, P.-T. Chang, "Recurrent support vector machines in reliability prediction", In processing of International Conference on Natural Computing (ICNC 2005), Vol. 3610, 2005, pp. 619–629.

[35] W.-C. Hong and P.-F. Pai, "Predicting engine reliability using support vector machines", The International Journal of Advanced Manufacturing Technology, Vol. 28, No. 1-2, Feb.2006, pp. 154–161.

[36] Nguyen Trung Hieu, Mario Di Francesco, Antti Yl¨a-J¨a¨ aski, " Virtual Machine Consolidation with Multiple Usage Prediction for Energy- Efficient Cloud Data Centers", IEEE TRANSACTION ON SERVICES COMPUTING, Vol. PP, No. 99, May 2017, pp. 1-14.

[37] Amazon EC2, https://aws.amazon.com/ec2/instance-types/, 2015.

[38] Wei Zhu, Yi Zhuang, Long Zhang, "A three-dimensional virtual resource scheduling method for energy saving in cloud computing", Future Generation Computer System, Vol. 69, 2017, pp. 66-74.