2017 28th International Workshop on Database and Expert Systems Applications

MuMs: Energy-Aware VM Selection Scheme for Cloud Data Center

Rahul Yadav, Weizhe Zhang, Huang Chen, and Tao Guo
School of Computer Science and Technology, Harbin Institute of Technology
Emails: rahul@stu.hit.edu.cn, wzzhang@hit.edu.cn, 857255934@qq.com, and gt8399@hit.edu.cn

***Abstract*—The energy consumption of data centers has been increasing continuously during the last years due to the rising demands of computational power especially in current Grid- and Cloud Computing systems, which directly influence the increment in operational costs as well as carbon dioxide (CO_2) emission. To reduce energy consumption within the cloud data center, it required energy-aware virtual machines (VMs) selection algorithms for VM consolidation at time host detected underloaded and overloaded and after allocating resources to all VMs from the underloaded hosts required to turn into energy saving-mode. In this paper, we propose energy-aware dynamic VM selection algorithms for consolidating the VMs from overloaded or underloaded host for minimising the total energy consumption and maximise the Quality of Service (QoS) include the reduction of service level agreements (SLAs) violation . To validate our scheme, we implemented it using CloudSim simulator and conducted simulations on the 10 different days real workloads trace, which provided by the PlanetLab.**

***Index Terms*—Cloud computing, data center, energy efficiency, VM selection, SLA.**

I. Introduction

In this modern digital era, the demand of computing power is increasing very fast, that required a large- scale of data centers. The statistical survey shown that the workloads transition from traditional data centers to cloud data centers by 2019, more than 86% workloads will be processed in cloud data centers, and 55%(2 billion) Internet users will use private cloud storage, up from 42%(1.1 billion) cloud users in 2014 [1]. These data centers require large amounts of electric energy to keep running the all-essential services to the cloud users, which lead the increment in carbon dioxide (CO_2) emissions as well as operating costs. It is not a good sign for cloud computing industry point of view as well as a social environment point of view.

Current research work conducted by researchers has been focusing on making data centers more sustainable and environment friendly, particularly in minimizing their electrical power consumption. In the Cloud Computing platform, a host has been installed more than one heterogeneous virtual machines (VMs) on parallel computing platform called data center. The virtualization technology gives administrative privileges to VM users within the guest operating system and they can customize their runtime resources according to their specific need. These VM can run different kinds of application simultaneously. The dynamic consolidation of VM is a significant method to optimal utilization of data center computing resources. The VM reallocated using live migration, according to their current resource requirement in order to minimize the number of active hosts, required to handle the workload. The idle hosts have switched to energy saving-mode with fast transition times to eliminate the static energy and reduce the total electric energy consumption. The hosts reactivated when the workload demand increases. This method has two main objectives: First, efficient electric power consumption and second, minimize the Service Level Agreements (SLAs) violation.

In this paper, we introduced a MuMs dynamic VM selection algorithm to balance the trade-offs among electric power consumption, the number of migrations, performance of host, and total number of hosts are shutdown. This scheme based on statistical analysis of hosts CPU utilization history. The main contributions of this paper are the following:

- We introduced a MuMs VM selection approach for selecting the VMs from overloaded hosts and under loaded hosts. The objective of this algorithm to optimal utilization of the energy and minimise the SLA violation.
- Evaluate the efficiency of MuMs approach by using CloudSim simulator and compare its performance against the performance of other proposed approach in literature.

The remaining part of this paper is organised as follows. In the Section 2, we discuss the important related work. Section 3, discusses the system architecture. Section 4, the proposed a energy-aware VMs selection algorithm for selecting VM from overloaded hosts or underloaded hosts. Section 5, describe the experiment setup and in the Section 6 introduced the efficiency metrics for measuring the overall performance of the algorithm. In the Section 7, evaluate the effectiveness of the proposed algorithm within the cloud environment. Finally, We conclude the summary of the this paper with in the Section 8.

II. Related Work

Even although, the literature work on the energy-aware dynamic VMs consolidation for the cloud data centers has been massive. The most important things have been noticed in previous works, that the dynamic VM consolidation approach, also called by other names such as dynamic VM management and dynamic resource allocation. However, the selection of

2378-3915/17 $31.00 © 2017 IEEE
DOI 10.1109/DEXA.2017.43

VM has always been part of this approach to make an efficient decision. The authors Pahlavan et al. [2] have proposed a VM consolidation algorithm for energy-aware in the cloud data center considering structural features such as racks and network topology. Moreover, they also focused on the structure of the network and cooling system in the cloud data center hosting the physical machines when consolidating the VMs. Therefore, some of the racks and routers employed, without compromising the SLA, so that low traffic routers or idle routers and cooling equipment can be turned off in order to minimize the electricity consumption. The author Zhu et al. [3] introduced a static CPU utilization threshold for detecting the overloaded host if the utilization of the host more than 85% of its total capacity, this host detected as overloaded host. This approach is not suitable for dynamic workload because it does not adapt to workload changes. However, nowadays, lots of current research work has been focusing on decision-making based on statistical analysis of historical data. The authors El-nozahy et al. [4] have examined the problem about the energy management for resources in a single web-application platform with constant SLAs (response time) and the balancing of the load controlled by the application. The authors Beloglazov et al. [5] have proposed cloud computing architectural framework and provisioning the data center resources in energy-efcient way, while meeting SLA requirements. They made two parts of the VM consolidation problem: (1) the submission of new requests for VM provisioning and allocate VMs on hosts, (2) the significant use of the current VM allocations. For solving the problem of VM placement on hosts, they use MBFD algorithm. This modeling rstly sort all VMs current CPU utilization in decreasing order, and allocate each VM to a host, which gives efficient energy consumption environment. In another work, they [6] introduces a heuristic based energy-aware approach. This approach focused on the statistical analysis of CPU utilization history to determine an upper threshold for detecting the overloaded hosts.

The authors Bobroff et al. [7] have introduced a dynamic server migration approach, which improve the amount of required resources and SLA violation rate. It can forecast dynamic workloads over intervals less than the time scale of demand variability.

III. System Architecture

Cloud computing is on-demand network access based virtualized platform, which provides large-scale of data center resources to the users. The submission of multiple requests for VMs provisioning which allocated to the hosts simultaneously. The allocation of VMs on the hosts according to the CPU utilization of host. The energy consumed by the CPU is linearly proportion to its utilization [6]. Therefore, the total resources of the host and usage VMs resources categorized by a single parameter called the CPU performance. The CPU performance defined in Millions Instructions per Second (MIPS). The efficient consolidation of the VM will reduce the electric energy consumption as well as the SLA violation rate. When the running VM cannot get its provision resources

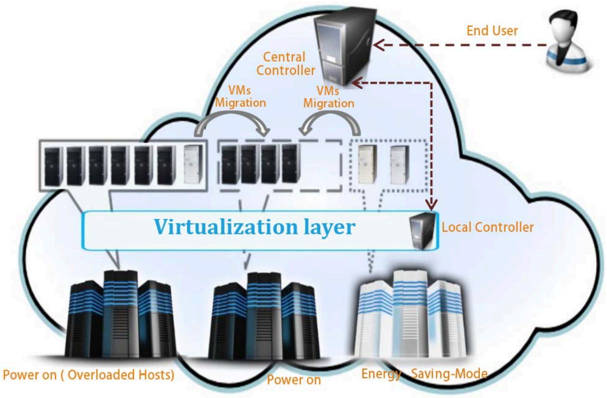

Fig. 1: Cloud Computing System Architecture.

(such as, MIPS, memory) from the cloud data center will occur SLA violation. In this case, a cloud service provider must be pay some penalty to cloud serves users. When its confirmed about an overloaded host. The next step is to select VMs for migration from overloaded host to appropriate host and it will have applied iteratively to the host until does not reorganized as overloaded.

The cloud computing Platform target shown in Fig 1. This platform has two key players first: A central controller and Second: A local controller, which similarly described in [5]. The central controller directly connected to end users as well as local controller. The resources management of the data center done by the central controller, which allocates VMs to hosts in the data center based on a predened policy. The resizing of the VM according to their resource requirements also done by a central controller, and it decides which and when VMs will migrate from one host to another host. A local controller resides in virtualization layer, which directly connected to the central controller, and it has a responsibility to monitor the current state of hosts and sending all gathered information to the central controller. We proposed a MuMs VM selection approach to optimize the electric energy consumption under the requirement of SLAs. This approach selects VMs from overloaded host and migrated to the other appropriate host, which fulfil its requirements. We are using Power Aware Best Fit Decreasing (PABFD) VM placement algorithm discussed in [6]. The turning the idle hosts into energy saving-mode, which is helpful in reducing the energy consumption in the data center.

We consider four different types of hosts are Fujitsu M1, Fujitsu M3, Hitachi TS10, and Hitachi SS10. The features of these hosts are shown in Table 2. We are getting the energy consumption of considered servers from the SPECpower [9].The electric energy consumption of these hosts at different workload are shown in Table I.

IV. Maximum Utilization Minimum Size Energy Model

The Maximum Utilization Minimum Size (MuMs) scheme based on the idea proposed in [6] for selecting of the VMs

TABLE I: The Electric Energy Consumed by the Considered Servers at Different Level of Workload in Watts(W)

Server	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Fujitsu M1	13.3	18.3	21.1	23.4	26.5	29.6	34.7	40.7	46.8	57.4	60
Fujitsu M3	12.4	16.7	19.4	21.4	23.4	26.1	29.7	34.8	41	47.1	51.2
Hitachi TS10	37	39.9	43.2	45.5	48.8	52.8	57.8	65.1	73.8	80.8	85.2
Hitachi SS10	36	38.8	41.2	43.7	46.3	49.4	53.1	58.8	64.2	67	69.7

at the time the hosts detected overloaded or under-loaded . This scheme selects only those VM whose utilization of CPU is high and its size must be minimum than other VMs. Moreover, this scheme estimated as the amount current VM CPU utilization is divided by the total size of the RAM allocated to this VM. The equation of the proposed VM selection policy within the Cloud environment are describe as follows:

Qi= $\{ q \mid q$ is a total number of VMs on the i^{th} host$\}$

$$MuMs = \frac{CurrentVmUtilization(q)}{TotalAllocatedSize(q)} \quad (1)$$

Where, Q_i represent the sets of all VMs on the ith host within the data center. CurrentVmUtilization represent the current CPU utilization of the VM and TotalAllocatedSize is a size of ram allocated to the VM.

Algorithm 1 MuMs VMs Selection

1: **Input:**vmList
2: **output:** Selected VM
3: vmToMigrate ← $NULL$
4: minMetric ← MIN
5: **for each** VM in vmList **do**
6:     size ← vmRam has get from host
7:     vmUtilaization ← $\frac{vmCpuMIPS}{vmMips}$
8:     metric ← $\frac{vmUtilization}{size}$
9:     **if** metric $>$ minMetric **then**
10:         minMetric ← metric
11:         vmToMigrate ← VM
12: **return** vmToMigrate

The pseudo-code in Algorithm 1, get a set of VM list of the overloaded host and check all VM sizes and its CPU utilization (lines 6-7). Afterword, based on VM size and VM CPU utilization calculate maximum utilization with minimum size called *MuMs* by using (2) (line 8). Next, compare the all VM of the vmList and Select a VM for migration who's size is minimum and CPU utilization is maximum than other VMs in the vmList of overloaded host (lines 9-12).

V. Experiment Setup

The deploying real large-scale virtualised infrastructure is very expansive and difficult for doing a repeatable experiment to analysis and compare the result of the proposed algorithm. Therefore, to repeatable experiment of the proposed algorithms, simulation is a best choice for evaluating VM selection policy. We have chosen the CloudSim toolkit [10] for analysis and compare the performance of the proposed MuMs VM selection scheme. This is a modern Open Source simulator, which provides an IaaS cloud computing framework that enable us to do repeatable experiments to analysis and compare the result on large-scale virtualized cloud data centers .

TABLE II: The Characteristics of the Hosts

Server	CPU	Core	Clock Speed	Memory
Fujitsu M1	Xeon 1230	4	2.7 GHz	8 GB
Fujitsu M3	Xeon 1230	4	3.5 GHz	8 GB
Hitachi TS10	Xeon 1280	4	3.5 GHz	8 GB
Hitachi SS10	Xeon 1280	4	3.6 GHz	8 GB

TABLE III: Amazon EC2 VM Types

VM Types	MIPS	Memory
Hight-CPU instance	2500	850 MB
Extra-Large instance	2000	3750 MB
Small instance	1000	1700 MB
Micro instance	500	613 MB

In our cloud computing simulation setup, we are installing 800 heterogeneous hosts which have real congurations. These hosts are Fujitsu M1, Fujitsu M3, Hitachi TS10, and Hitachi SS10. The features of these servers are shown in Table II. The electric energy consumption of these servers at different workload are shown in Table I.

The servers CPU clock speed are mapped onto MIPS ratings: each core of the server Fujitsu M1 is mapped 2700 MIPS, 3500 MIPS each core of the Fujitsu M3 server, 3500 MIPS each core of the Hitachi TS10 server , and 3600 MIPS each core of the Hitachi SS10 server . The network bandwidth of the each server is modeled to have 1GB/s. The Corresponding VM types supported as Amazon EC2 VM types as shown in Table III

The main key factor of the simulation based experiment must be conducted using real workload trace of the data center server, which will help to applicable on real cloud environment. For obtaining this objective, we have used the data provided by PlanetLab as a part of CoMon project [11]. We have used more than a thousand heterogeneous VMs CPU utilization data from more than 500 heterogeneous servers places all around the world. The features of the data for each day are discussed in [6]

VI. Efficiency Metrics

Evaluating the results and compare the effectiveness of the algorithm, we are using different kind of the metrics. The first kinds of the metric called total energy consumed by

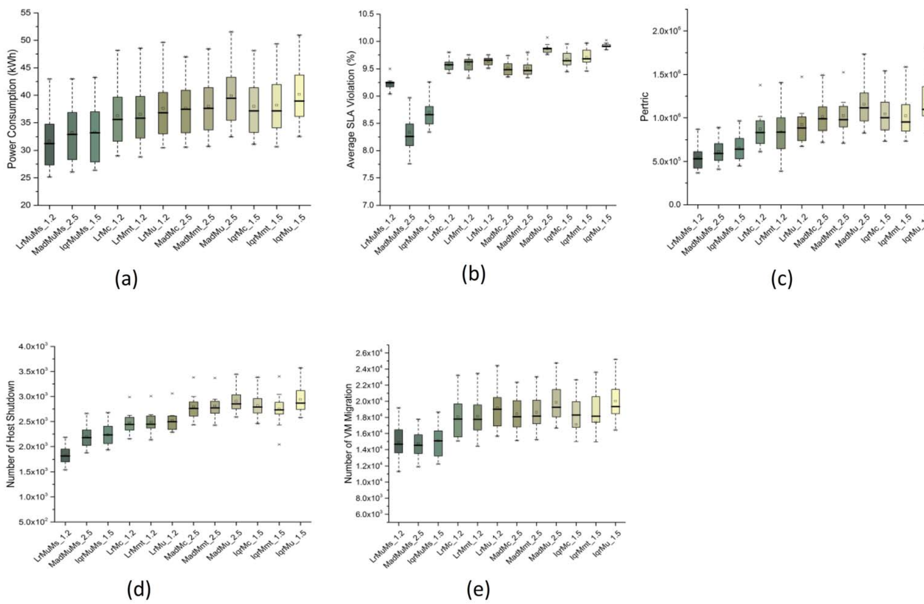

Fig. 2: Evaluating the effectiveness of *MuMs* algorithm. (a) Energy Consumption Comparison (b) Average SLA Violation (c) Performance Metric (d) Number of Hosts Shutdown (e) Number of VM migration.

the data center resources at different workload. The second type of efficiency metric is the average percentage of the SLA violation, it only occurs when provisions VMs are not getting requested resources (or average computing power of the shared host is not allocated to the requested VMs). This metric directly influence on the level of Quality of Service (QoS) which is not negotiated between cloud provider and its users. If an SLA violation occurs, then there is responsibility of cloud service provider must be pay some penalty for it to users.

A. *Performance Metric (Pertric)*

In order to maximize the overall performance with minimize the electric energy consumption, average SLA violation, and number of the hosts, which re-turn on from the energy saving-mode. If the number of the hosts is increasing who is going to energy saving-mode and again re-switch on for allocation of the VMs, which is directly impact on the energy consumption of the data center. For taking this concern, we introduced a performance metric describe as follows:

$$Pertric = ASLA * HS * E \tag{2}$$

Where, Pertric represent as the overall performance metric, *HS* represent the total number of the hosts shutdown after apply this algorithms, and E is a total electric energy consumption of the data center. *ASLA* represent the average SLA violation percentage in the data center

VII. Simulation Result and Analysis

To evaluate the performance of MuMs VM selection scheme, we are using real time CPU utilization data of heterogeneous servers. We have simulated our MuMs VM selection scheme with the overloaded hosts detection algorithms described in [6]. These overloaded hosts algorithms are Median Absolute Deviation (MAD), Linear Regression (LR), and Inter Quartile Range (IQR). For evaluating the proposed policy we consider real workload.

A. *Real Workload*

The real workload dataset provided by the PlanetLab as part of CoMon project. In CoMon project, data of thousand of VMs CPU utilization all around the world collected every five minute interval and stored in the different extension files. We selected this real dataset for evaluating the proposed policy. Analysis the proposed policy by using real workload discusses following sub-sections.

1) Evaluating the Energy Consumption : The total electric energy consumption of the host's resources in the data center depends on the CPU utilization, primary memory, network devices, and disks. But lots of the research work shows that the hosts CPU consume more electric energy than the other resources in the hosts [12] [13] . Therefore, we more focused on the CPU utilization of the hosts .

In this Section, we analysis and comparison of energy consumption simulation results using proposed MuMs VMs selection policy with the old hosts overloaded detection algorithms (such as, Median Absolute Deviation(MAD), Linear Regression (LR), and Inter quartile Range(IQR)) proposed in [6]. The Fig. 2a showing the hourly electric energy consumption using MuMs VMs selection policy is lesser than other old algorithms of the data center.

As the Fig. 2a showing that the energy consumption using

LrMuMs , MadMuMs, and IqrMuMs is less than the others old algorithms proposed in [6]. By the using of the proposed policy, the data center consuming 13% less energy than the traditional.

2) Evaluating Average SLA Violation : To maintaining the QoS is an important aspect of the cloud computing environment. The required QoS are determined by SLAs [14]. In this section, we analysis and compare the percentage of average SLA violation occurs during the overloaded hosts. The cloud users do not want to fell SLA violation and the performance degradation situation. If this situation occurs than Cloud Service Provider (CSP) must need to pay the penalty to the users. Thus, the user point of view as well as CSP point of view, decrement in SLA is a good news. The Fig. 2b showing that the percentage of average SLA using MuMs VMs selection policy is 10% less than the traditional algorithms.

3) Evaluating Performance Metric (Pertric) : In this section, we discuss about the overall performance of the cloud data center by using proposed MuMs VM section policy. The overall performance calculated by the Pertric metric proposed in section 6.A. The main objective to propose this metric to analysis the all aspects of energy-awareness in the cloud data center such as minimization in electric energy consumption, average percentage SLA violation, and number of reactivated hosts for placing the new VMs.

In the Fig. 2c showing the effectiveness of MuMs VMs selection policy relatively other old VMs Selection policies (such as Minimum Migration Time (Mmt), Maximum Correlation (Mc), and Minimum Utilization (Mu)) proposed in [6].

4) Number of Hosts Shutdown and VMs Migration : The cost of dynamic live migration of VMs is always high which includes some amount of processing power on the allocated host, and performance degradation [5] [14]. Therefore, one of the objective is to minimise the total number of VMs migrations. In this section, we analysis and comparison the simulation results on the number of hosts shutdown and VMs migrations. If the number of the reactivated hosts are increasing, which lead the maximise in electric energy consumption. It means the hosts reactivated for allocation of new VMs and shutdown after some time when it detected underloaded.

In the experiment environment, we usage only 800 hosts but the number of hosts shutdowns is more than it cause by reactivation of hosts. In Fig. 2d showing that the using proposed MuMs policy also, minimize 13% of the reactivation of hosts relatively to traditional VMs selection policy.

The number of the migration is directly proportional to the performance degradation. If the total number of VMs migration decrease then the performance degradation also decrease which is good for users point of view and CSP point of view both. In Fig. 2e shows, the comparison of proposed policy VMs migration and other old policy describe in [6].

VIII. Conclusion

The Cloud data centers all around the world are growing, according to the computing demand, which is growing rapidly. Therefore, to keep running these data centers required huge amount of electric energy, which lead high operational cost and high carbon dioxide (CO_2) emission. The high emission of CO_2 is directly bad impact on the social environment of earth. This paper introduced a MuMs VMs selection policy, which is helpful to minimize the electric energy consumption and reduced the SLA violation of the cloud data centers. The implementation of this policy, we usage the CloudSim simulator to obtained the experiment results for analysis and comparison with other old algorithms such as MAD, IQR, and LR discussed in [6].

IX. Acknowledgement

The National key R&D Program of China under grant No. 2016YFB0800801, the National Science Foundation of China (NSFC) under grant No. 61672186, 61472108, support this work.

References

[1] (2015) Cisco global cloud index: Forecast and methodology, 20142019. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf

[2] S. Esfandiarpoor, A. Pahlavan, and M. Goudarzi, "Structure-aware online virtual machine consolidation for datacenter energy improvement in cloud computing," *Computers & Electrical Engineering*, vol. 42, pp. 74–89, 2015.

[3] X. Zhu, D. Young, B. J. Watson, Z. Wang, J. Rolia, S. Singhal, B. McKee, C. Hyser, D. Gmach, R. Gardner *et al.*, "1000 islands: Integrated capacity and workload management for the next generation data center," in *Autonomic Computing, 2008. ICAC'08. International Conference on*. IEEE, 2008, pp. 172–181.

[4] E. M. Elnozahy, M. Kistler, and R. Rajamony, "Energy-efficient server clusters," in *International Workshop on Power-Aware Computer Systems*. Springer, 2002, pp. 179–197.

[5] A. Beloglazov, R. Buyya, Y. C. Lee, A. Zomaya *et al.*, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances in computers*, vol. 82, no. 2, pp. 47–111, 2011.

[6] 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*, vol. 24, no. 13, pp. 1397–1420, 2012.

[7] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on*. IEEE, 2007, pp. 119–128.

[8] 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*, vol. 28, no. 5, pp. 755–768, 2012.

[9] "All published specpowerssj2008 results," https://www.spec.org/power_ssj2008/results/power_ssj2008.html, accessed: 2017-05-12.

[10] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya, "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and experience*, vol. 41, no. 1, pp. 23–50, 2011.

[11] K. Park and V. S. Pai, "Comon: a mostly-scalable monitoring system for planetlab," *ACM SIGOPS Operating Systems Review*, vol. 40, no. 1, pp. 65–74, 2006.

[12] D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing environments via lookahead control," in *Autonomic Computing, 2008. ICAC'08. International Conference on*. IEEE, 2008, pp. 3–12.

[13] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2. ACM, 2007, pp. 13–23.

[14] F. Farahnakian, A. Ashraf, T. Pahikkala, P. Liljeberg, J. Plosila, I. Porres, and H. Tenhunen, "Using ant colony system to consolidate vms for green cloud computing," *IEEE Transactions on Services Computing*, vol. 8, no. 2, pp. 187–198, 2015.