Energy Efficient Live Virtual Machine Provisioning at Cloud Data Centers - A Comparative Study

Shalini Soni
M. Tech. Scholar
Bhopal Institute of Technology & Science,
Bhopal

Vimal Tiwari
Assistant Professor
Bhopal Institute of Technology & Sci
Bhopal

ABSTRACT
Cloud computing offers utility-oriented IT services like: pervasive applications from consumer, scientific, and business domains based on a pay-as-you-go model. So, the workload in cloud environment is usually dynamic. At cloud data centers, different virtual machines (VMs) Provisioning techniques cause different CPU utilization. Therefore, VM Provisioning on PMs to improve resource utilization and reduce energy consumption is one of the major concerns for cloud providers. The problem of VM Provisioning includes queuing of VM requests, placing the VMs on hosts, and the optimization of the current VM allocation using Live Migration. The existing VM provisioning schemes are to optimize physical server and network resources

utilization, but many of them also focus on optimizing multiple resources utilization simultaneously. The setting up of utilization thresholds for resources is one of the common optimization techniques. The ultimate aim of Cloud providers is to optimize resource usage and reduce energy consumption with the obligation of providing high Quality of Service (QoS) to customers, while maintaining the Service Level Agreements (SLAs). We surveyed various Live Virtual Machine Provisioning techniques and presented the comparison among few benchmark techniques based on adaptive utilization thresholds, as contribution to Green Cloud computing solutions. A performance evaluation study and comparison is done using the CloudSim toolkit.

General Terms
Cloud Computing, Cloud Provider, Energy Efficiency, SLA, Virtual Machine Allocation,

Keywords
Adaptive Threshold, Cloud Computing (CC), Cloud Providers, Energy, Energy efficient, Quality of Service (QoS), Service Level Agreements (SLA), Virtual Machine (VM), VM Allocation, Green computing

1. INTRODUCTION
Cloud Computing technologies are gaining popularity due to attributes like dynamic scaling, on-demand provisioning and the pay-as-you-go model. In recent years, this computing paradigm has received wide adoptions by industrial, scientific and academic users. Datacenters normally meet different usage scenarios from users e.g. running a scientific simulation, which may be in form of a batch job with or without a specific deadline; or hosting a government or corporate web site for a long period of time, which requires a guaranteed Quality of Service (QoS). Recently, as the scale and performance of IT data centers grow, data centers often become less efficient in utilizing system resources [1]. Such ineffective utilization often increase operational costs and power consumption results in reduced system reliability and device lifetime. Another problem is significant CO2 emissions that contribute to the greenhouse effect [2].

Virtualization technology at data center is one of significant way to reduce power consumption. This technology allows server consolidation in a datacenter, thus reducing the amount of the hardware in use. Thus cloud providers can reduce the total energy consumption for servicing their clients within agreed Service Level Agreement (SLA) violations.

The problem of VM Provisioning includes queuing of VM requests, placing the VMs on hosts, and the optimization of the current VM allocation. The objective of the optimization is achieved through reallocation of VMs using Live Migration. Reallocation of VMs minimizes the number of physical nodes serving current workload, whereas idle nodes are switched off in order to lessen the power consumption. The algorithm for 'optimization of the current VM provisioning' looks through the list of hosts and detects for overloaded and under-loaded host. Then, the algorithm applies the VM Selection policy to select VMs that need to be migrated from the overloaded and under-loaded host. Finally the VMs are placed as per the VM placement algorithm.

For optimization problem, setting up of utilization thresholds as static is inefficient for systems with unknown and dynamic workloads in cloud environment. Such techniques do not adapt to workload changes and do not capture the time-averaged behavior of the system [3]. Here, in this work, we go through the techniques of VM Provisioning that utilizes the static and Adaptive Utilization Thresholds based techniques for determining host overload and under-load and 'Random Choice' policy for selection of VM for migration. The main idea behind utilizing adaptive-thresholds is preventing CPU utilization to become 100% which cause the SLA violations. Furthermore, a number of techniques of host overload and under-load detection are surveyed, analyzed and combined with one of the VM selection policies for performance evaluation purpose. The evaluation is done through simulations using famous cloud simulation toolkit: CloudSim.

The remainder of the paper is organized as follows. In Section 2 we present the literature review (related work) followed by the power and system model in Section 3. The literature review prominently surveys and compares VM provisioning techniques based on static and adaptive threshold. Section 4, introduces the performance parameters of energy efficiency and presents the experimental results obtained through evaluation and analysis. The general simulation parameters are also mentioned in this. We make a conclusion and discuss possible directions for future research in Section 5.

2. LITERATURE REVIEW
The work [4] is emphasized on implementation, simulation and function validation of DVFS in CloudSim simulator. The close relationship between DVFS efficiency and hardware architecture is highlighted by the use of a scientific

application. The paper also demonstrates that the DVFS efficiency also depends on the built-in middleware behavior. But, the DVFS scheme reduces the dynamic power consumption by decreasing the supplying voltage and frequency, which results in a slowdown of the CPU and increased execution time.

Kyong Hoon Kim et al. [5] investigated power-aware provisioning of virtual machines for real-time services framework and proposed adaptive versions of DVFS. At Cloud data centers, the approach tries to model a real-time service as a real-time virtual machine request; and provisioning of virtual machines using Dynamic Voltage Frequency Scaling (DVFS) schemes. Several schemes to reduce power consumption by hard real-time (HRT) services and power-aware profitable provisioning of soft real-time (SRT) services, is proposed. The HRT services like financial analysis, distributed database, or image processing, which consists of multiple real-time applications or subtasks are considered. A user requests VMs by either HRT-VM or SRT-VM and Cloud resource brokers finds resources or VMs for such real-time services requested by users. The proposed algorithms are simulated using the CloudSim toolkit [14, 15] with an extension enabling power-aware simulations. The simulation results have shown that data centers can reduce power consumption for soft real-time services and increase their profit using proposed Adaptive-DVFS schemes regardless of power consumption.

Beloglazov et al. [3] presents some novel techniques for the auto-adjustment of the utilization thresholds based on a statistical analysis of historical data collected from the resource usage by VMs, during their lifetime. The main idea of the proposed adaptive-threshold algorithms is to adjust the value of the upper utilization threshold depending on the strength of the deviation of the CPU utilization. In case higher the deviation, more likely that the CPU utilization will reach 100% and cause an SLA violation. To calculate the upper CPU utilization threshold few statistical methods are used. These statistical methods to determine over-utilized and under-utilized hosts, and policies to select a VM to be migrated, can be combined to form various strategies. The destination hosts is chosen in order to minimize power consumption. Few of adaptive-threshold algorithms are based on statistical methods: Median Absolute Deviation (MAD), Local Regression (LR) and Interquartile Range (IQR). These are compared in next section.

The authors' in [6] propose an admission control and scheduling mechanism which maximize the resource utilization, profit and also ensure QoS requirements of users to met specified SLAs. In this work, it is assumed that the datacenter will receive two types of application workloads having different QoS requirements, i.e., transactional and non-interactive batch jobs.

The authors [7] in their tried to investigate - SLA and Energy-Efficient Dynamic Virtual Machine (VM) Consolidation techniques, that meets Quality of Service expectations and Service Level Agreements (SLA) requirements. The VM consolidation algorithms are analyzed based on various heuristics on legitimate host. By conducting a performance evaluation study of various existing energy efficient VM consolidation techniques, a comparative analysis and results are presented. The experiments are done using real world workload traces obtained from more than a thousand VMs using CloudSim toolkit.

In this work [10] authors conducted a survey of research in energy-efficient computing and proposed: energy-efficient resource allocation policies and scheduling algorithms considering QoS expectations and power usage characteristics of the devices; The proposed VM allocation is carried out in two steps: first VMs that need to be migrated are selected, and then "Modified Best Fit Decreasing" (MBFD) algorithm is used for placement. These algorithms sorts all VMs in decreasing order of their current CPU utilizations, and allocate each VM to a host that provides the least increase of power consumption due to this allocation. This allows controlling the heterogeneity of resources by choosing the most power-efficient nodes first. Considering **n** is the number of VMs that have to be allocated and **m** is the number of hosts the complexity of the allocation part of the algorithm is calculated as **n · m**.

In another work Beloglazov and Buyya [11] have presented a novel technique for dynamic consolidation of VMs which ensures Service Level Agreements (SLA) and based on adaptive utilization thresholds. Authors approach is based on a Markov chain model that optimally solves the problem of host overload detection under the specified QoS goal, for any known stationary workload and a given state configuration. Non-stationary workloads are also handled by the algorithm. The extensive work has been simulated on more than a thousand VMs.

A technique of setting upper and lower utilization thresholds for hosts and keeping the total utilization of the CPU by all the VMs between these thresholds is proposed in [12]. The utilization thresholds are used to decide the time to migrate VMs from a host and applied to host overload detection. But it is stated that, the fixed values of utilization thresholds are unsuitable for an environment with dynamic and unpredictable workloads, in which different types of applications can share a physical resource [3]. The system should be able to automatically adjust utilization behavior on the workload patterns exhibited by the applications.

3. POWER AND SYSTEM MODEL

3.1 Power and System Model

We used the power model and energy consumption model as described in [8, 9, 10] which defines the power consumption as a function of the CPU utilization P(u) as shown in equation (3.1) and total energy consumption by a server as defined in equation (3.2), where Pmax is the maximum power consumed; k is the fraction of power consumed by an idle server which is approximately 70% of power consumed by fully utilized CPU; and u is the CPU utilization.

$$P(u) = k.P_{max} + (1 - k) \cdot P_{max} \cdot u = P_{max} \cdot (0.7+0.3.u) \quad \textbf{(3.1)}$$

The CPU utilization may change over time due to the dynamic workload. Thus, the CPU utilization is a function of time and is represented as u(t). Energy consumption by a physical node (E) is represented as

$$E = \int_t P(u(t))dt \quad \textbf{(3.2)}$$

System model mentioned in [7, 8, 10,] is utilized of for this work employs the IaaS environment represented by a large-scale data center consisting of M heterogeneous physical nodes. The type of the environment entails no knowledge of application workloads and time for which VMs are provisioned. Multiple independent users submit requests for provisioning of n heterogeneous VMs characterized by requirements of: processing power, RAM and network

bandwidth.

In this work, Green Service Negotiator is added as an improvement. It is required components to support the energy-efficient resource management. Green Negotiator negotiates with the consumers/brokers to finalize the SLAs between the Cloud provider and consumer depending on the consumer's QoS requirements and energy saving schemes. The global manager resides on the master node and collects statistics from the local managers to maintain system's resource utilization view and based on the judgment made by the local managers, the global manager issues VM migration commands to optimize the VM placement. Virtual Machine Manager (VMM) performs actual migration of VMs and takes decisions to alter the power modes of the nodes.

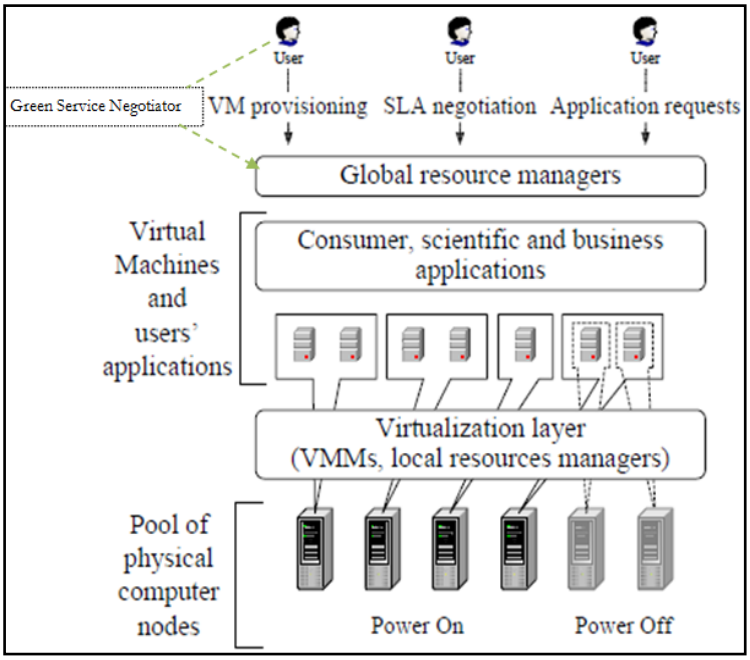

Figure 4.1 System Model for Proposed Work

4. PERFORMANCE COMPARISON

4.1 Experimental Setup

The simulations have been done on CloudSim [14, 15] toolkit to evaluate and compare few VM provisioning techniques discussed in this work. It is a modern simulation framework aimed at Cloud computing environments. The general simulation parameters chosen for experimentation purpose follow real workload traces from more than a thousand Planet Lab [16] VM

Table 1: Simulation Parameters

Simulation Variables	Values
No. of Hosts	800
Types of Host with configurations	**Type -1:** 2 CPU cores, 1860 MIPS/core, 4GB RAM, 1GB storage, 1Gbps Network Bandwidth.
	Type -2: 2 CPU cores, 2660 MIPS/core, 4GB RAM, 1GB storage, 1Gbps Network Bandwidth
No. of VMs. (PlanetLab Workload)	1054 ("2010303") Each VM is randomly assigned a workload trace from real workload data.

VM Types & their Requirements	4. Type 1: 2500 MIPS & 870 MB RAM, Type 2: 2000 MIPS & 1740 MB RAM, Type 3: 1000 MIPS & 1740 MB RAM, Type 4: 500 MIPS & 613 MB RAM, Network Bandwidth: 100 Mbps & 2.5 GB of storage.
Utilization measurement interval	900 seconds.
Simulation Time	24 hours

Here in this paper comparative study of VM Provisioning methods, which is combination of following Adaptive Utilization Threshold – and Random VM Selection method [10] policies, is done. The conventions used are shown in Table -2. The VM selection policy 'Random Choice' proposed in [10] is chosen for experimentation purpose. The Random Choice (RC) policy selects a VM to be migrated according to a uniformly distributed discrete random variable.

$$X^d = U(0, |Vj|); \quad (4.1)$$

whose values index a set of VMs Vj allocated to a host j. The policy is chosen to see the effect of only adaptive utilization thresholds based allocation when no optimized VM selection is applied.

Table 2: Conventions Used

POLICIES	CONVENTION
Static Threshold + Random Choice	**ThrRs**
Median Absolute Deviation + Random Choice	**MadRs**
Interquartile Range + Random Choice	**IqrRs**
Local Regression + Random Choice	**LrRs**

4.2 Performance Metrics

(a) Total Energy Consumption: It denotes the total energy consumption by physical servers of a data center caused by application workloads. Energy consumption is calculated according to the power model and energy consumption model mentioned above.

(b) Combined Energy and SLA Violation: a combined metric is utilized with the objective of minimizing the energy consumption while maintaining the level of SLA violation [10]. The combined metric is given below:

ESV = E. (SLAV)

Where E is total energy consumption and SLAV is SLA violation. SLA violation in an IaaS environment is measured as the percentage of time, during which active hosts have experienced the CPU utilization of 100%. ESV is combined Energy and SLA Violation.

(c)Number of VM Migrations:

Live VM migration may puts failures and performance issues in applications running in a VM. So it is one of the important metric for measuring performance of our policy and it should be less.

Table 5.3 Comparison of ThrRs, MadRs, IqrRs and LrRs

Parameters	Thr	Mad	IqrR	LrR
Energy Consumption	195.	189.1	194.	176.
Energy - SLA violation	1.64	1.65	1.57	1.73
No. of VM migrations	8992	8951	8990	9283

4.3 Simulation Results and Analysis

The comparison between ThrRs, MadRs LrRs and IqrRs on basis of three parameters is shown in Table – 3. The graphs of Energy Consumption, Combined Energy and SLA violation (ESLAV) and Number of VM Migrations are shown below for each of these schemes in Figure 2, 3 and 4 respectively. Random selection VM Selection policy is used to understand that which of the statistical adaptive technique is performing better.

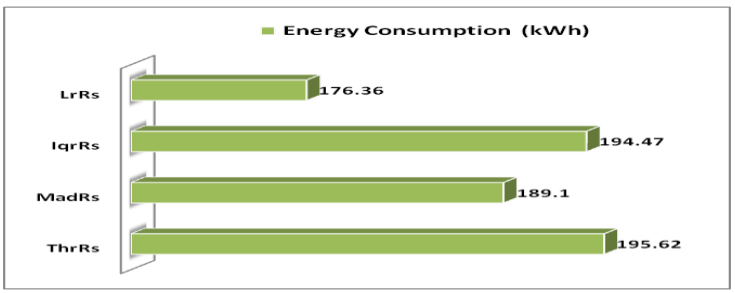

Figure 2: Energy Consumption (KWh)

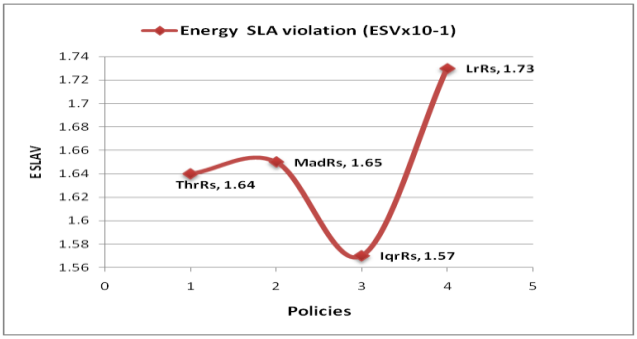

Figure 3: Energy SLAV (x10^{-1})

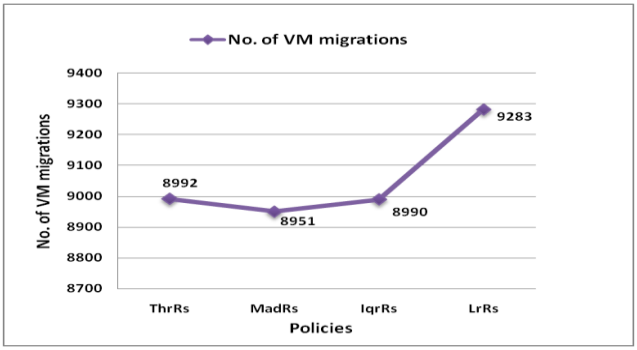

Figure 4: Number of VM Migrations

MadRs is performing better in terms of VM migration while IqrRs having minimum ESLAV. Almost all provisioning policies are giving near values when Random Selection method for VM migration is used. Such results show significance and need of optimal VM selection policies for optimal Live VM migration.

5. CONCLUSION AND FUTURE WORK

This paper focuses on Energy Efficient Live Virtual Machine Provisioning and present: (a) survey on Energy efficient Optimization of the VM Provisioning"; and for this purpose (b) Analysis of VM Allocation algorithms based on adaptive utilization threshold that are derived using statistical methods is done. (c) Comparative analysis and results by conducting a performance evaluation study of various techniques using real world workload traces.

In Cloud environment, the hosted applications at data centers are having heterogeneous requirements and vary over time. The Clients require strict QoS guarantees, which are documented in the form of SLAs. Though varied application requirements make VM provisioning algorithms complex, but they can be exploited to improve energy-efficiency.

The focus of this work is to study VM provisioning based on Adaptive utilization threshold based allocation strategies that can be applied in a virtualized data center by a Cloud provider. Such thresholds are applied with the purpose of maintain QoS and SLAs. To achieve a scalable solution for handling thousands of users, the existing techniques are tested using a series of simulation experiments on the CloudSim platform using real world workload. The results shown that the ESLAV is least in IQR based technique while MadRs is performing better in terms of VM migration.

As a future direction, this work suggests the study and development of other Energy Efficient SLA aware Resource Provisioning techniques over varied workloads to make the data centers scalable and reliable in terms of QoS. Also the individual problems like provisioning of VM requests, VM placement optimization, and dynamic VM consolidation can be analyzed and modified individually. Other statistical techniques can be applied: (i) to obtain optimal and near optimal solutions to predict the future workloads. (ii) To individual problems of Resource provisioning. The work may prove potential and encourage the researchers to perform competitive analysis of these algorithms to get theoretical performance.

6. REFERENCES

[1] Jian-Sheng Liao, Chi-Chung Chang, Yao-Lun Hsu, Xiao-Wei Zhang, Kuan-Chou Lai, Ching-Hsien Hsu, "Energy-Efficient Resource Provisioning with SLA consideration on Cloud Computing", 2012 41st International Conference on Parallel Processing Workshops, IEEE 2012.

[2] Anton Beloglazov and Rajkumar Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers", 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, IEEE Computer Society.

[3] Anton Beloglazov, Rajkumar Buyya, "Optimal online deterministic algorithms and adaptive heuris-tics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers", Wiley InterScience, Concurr. Comput. : Pract. Exper., 24(13):1397-1420, September 2012.

[4] Tom Guérout, Thierry Monteil, Georges Da Costa, Rodrigo Neves Calheiros, Rajkumar Buyya, Mihai Alexandru, "Energy-aware simulation with DVFS", Simulation Modelling Practice and Theory 39 (2013) 76–9, 2013 Elsevier B.V.

[5] Kyong Hoon Kim, Anton Beloglazov, and Rajkumar Buyya, "Power-Aware Provisioning of Virtual Machines for Real-Time Cloud Services", CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE Concurrency Computat.: Pract. Exper. 2011;

[6] Saurabh Kumar Garg , Adel Nadjaran Toosi, Srinivasa K. Gopalaiyengar, Rajkumar Buyya, "SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter", Elsevier - Journal of Network and Computer Applications 45(2014)108–120.

[7] Heena Kaushar, Pankaj Ricchariya and Anand Motwani. Article: Comparison of SLA based Energy Efficient Dynamic Virtual Machine Consolidation Algorithms. International Journal of Computer Applications 102(16):31-36, September 2014.

[8] Anton Beloglazov and Rajkumar Buyya, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers", MGC '2010, 29 November - 3 December 2010, Bangalore, India. Copyright 2010 ACM 978-1-4503 0453-5/10/11.

[9] Anton Beloglazov and Rajkumar Buyya, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers", MGC '2010, 29 November - 3 December 2010, Bangalore, India. Copyright 2010 ACM 978-1-4503-0453-5/10/11.

[10] Anton Beloglazov, Jemal Abawajy, Rajkumar Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing",

Future Generation Computer Systems, Volume 28, Issue 5, May 2012, Pages 755-768.

[11] Anton Beloglazov and Rajkumar Buyya, "Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centers under Quality of Service Constraints", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 7, JULY 2013.

[12] Beloglazov A, Abawajy J, Buyya R. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Generation Computer Systems 2011; doi:10.1016/j.future.2011.04.017.

[13] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," Journal of the American Statistical Association, vol. 74, no. 368, pp. 829–836, 1979.

[14] Buyya, R.; Ranjan, R.; Calheiros, R.N., "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," High Performance Computing & Simulation, 2009. HPCS '09. International Conference on , vol., no., pp.1,11, 21-24 June 2009.

[15] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, Cesar A. F. De Rose, and Rajkumar Buyya, "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms," Software: Practice and Experience (SPE), Volume 41, Number 1, Pages: 23-50, ISSN: 0038-0644, Wiley Press, New York, USA, January, 2011.

[16] The PlanetLab platform. http://www.planet-lab.org/.