International Journal of Electrical, Electronics and Computer Engineering **5**(1): 9-14(2016)

ISSN No. (Online): 2277-2626

Adaptive Load Detection Technique for Effective Virtual Machine Consolidation

Prof. Anand Motwani*, Sayed Qudsiyanaz Rafique**, Dr. P. N. Singh and Prof. Jyoti Sondhi*****

**Assistant Professor & Head, Department of Computer Science & Engineering,*
NRI Institute of Research & Technology, Bhopal, (Madhya Pradesh), INDIA
***M. Tech. Scholar, Department of Computer Science & Engineering,*
NRI Institute of Research & Technology, Bhopal, (Madhya Pradesh), INDIA
**** Principal/Director, NRI Institute of Research & Technology, Bhopal, (Madhya Pradesh), INDIA*
*****Assistant Professor, Department of Computer Science & Engineering,*
NRI Institute of Research & Technology, Bhopal, (Madhya Pradesh), INDIA

(Corresponding author: Sayed Qudsiyanaz Rafique)

(Received 03 December, 2015 Accepted 12 February, 2016)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: **Cloud Computing (CC) is a hi-tech way of delivering the shared pools of software, platform and hardware resources as service through internet. The delivery and access to the resources is dynamic, convenient, on-demand and on basis of pay-as-per usage. The virtualization technology along with concept of distributed computing and other technologies made this complicated provisioning possible. Due the growing computational needs, the major challenge faced by cloud stakeholders is maintaining power-performance trade-off by satisfying high Quality of Service (QoS) and Service Level Agreements (SLA). At cloud Data Centers (DCs), the suitable optimization in power consumption through Dynamic Virtual Machine (VM) Consolidation is seen as potential approach for reducing energy consumption while maintaining QoS and SLAs up to a good level. The approach dynamically regulates the number of active machines to match resource demands while keeping others in sleep/hibernate mode to save energy. The efficient determination of overloaded and under-loaded hosts is the primary step in Dynamic VM consolidation. Based on this,the effective decisions like migration of VMs to other hosts can be taken to optimize the power. The theme of this work is to propose the 'Adaptive Load Detection Technique for Effective Dynamic VM Consolidation' and provide the optimization for better performance and environment. The load detection technique proposed in this work is utilizinga statistical method for determining overloadedhosts. The statistical method determines adaptive thresholds to detect over load conditions. The technique is then combined with few VM selection strategies for obtaining Effective Dynamic VM Consolidation.In this work few statistical techniques for detecting overloaded hosts that are part of Energy-Efficient VM Consolidation Policies are studied. The comparison / survey of VM consolidation algorithms based on various heuristics are presented. For our simulations CloudSim toolkit version 3.0.2 is used with real world workload traces of VMs. The experimental results demonstrates that the proposed policy is scalable and offers better consolidation quality and energy saving while effectively dealing with firm QoS requirements negotiated by SLAs.**

Keywords: Cloud Computing, Dynamic Virtual Machine (VM) Consolidation, Load Detection, Quality of Service (QoS) and Service Level Agreements (SLA), Resource Provisioning, CloudSim.

I. INTRODUCTION

The Cloud Computing can be defined on the several bases including user experience, business model, infrastructure and the service delivery aspects. Cloud computing is an emergent way of provisioning IT service (software, platform & infrastructure) delivery to users over the web with pay-as-you-go pricing model. Such computing model increasingly adopted in many areas, such as e-commerce, retail industry, and academy [1]. The Cloud DCs contains thousands of computing nodes that consumes massive amounts of electrical energy. As per one of the estimation, by 2014 infrastructure and energy costs would contribute about (¾) Th whereas IT would contribute just (¼) Th to the overall cost of operating a DC [2].

The reason for this extremely high energy consumption lies in the ineffective usage of these resources and not just the quantity of computing resources and the power inefficient hardware. Most of the time servers work at 10-50% of their full capacity, leading to more energy consumption and thus extra cost and Carbon dioxide ($CO_{2)}$ emission. The underutilization and over-utilization of servers is really inefficient.

Virtualization technology is one of the ways to deal with such inefficiencies as this allows Cloud providers to create multiple Virtual Machine (VMs) instances on a physical server, consequently improving the utilization of resources. Still, a lot of resources are left unutilized to support the scalability property of Cloud Computing. To address the problem of underutilization and over-utilization of servers the dynamic virtual machine (VM) Consolidation to least number of servers (hosts) in accordance with the current resource requirements is suggested [3]. The consolidation can be achieved by using one of the virtualization capabilities known as Live VM Migration. Using live migration, VMs can be dynamically consolidated to minimum number of active physical nodes while dealing with workload constraints. The energy consumption can be attained by switching idle nodes to sleep or hibernation (i.e. low-power modes), thus eliminating the idle power consumption. The current desktop and server CPUs can consume less than 30% of their peak power in low-activity modes, leading to dynamic power ranges of more than 70% [9]. Dynamic VM consolidation involves two basic processes: detecting over-utilized and under-utilized hosts; and selecting VMs from those hosts for live migration to minimize the number of active hosts. But to achieve optimized effect on energy conservation subject to performance constraints, it is important that how and when to do such consolidation remains an open problem.

The focus of this work is on proposing effective dynamic VM consolidation using adaptive thresholds of load at hosts. The calculation of adaptive thresholdis based on a statistical technique known as InterQuartile Deviation. The technique is applied by a Cloud provider at Infrastructure as a Service (IaaS) based Cloud DCs under QoS constraints. The determination of under-utilized hosts and over-utilized host is the first step in VM consolidation problem. In Cloud environment users dynamically provision VMs and deploy applications with heterogeneous requirements over it. In this way cloud servers face mixed workloads at different times. So, dynamic decision of determining over and underutilized hosts is one of critical processes.

Section 2, discussesdifferent approaches including adaptive thresholds based approaches, for determining the under and over utilized hosts. Host overload determination methods are sometimes referred to as VM Allocation Policies. Next, an “Effective Dynamic VM Consolidation” policy / scheme based on proposed “Adaptive Load Detection Technique” which is a statistical approach for determination of over-loaded hosts is proposed in Section 3. Then, in Section 4 we examine performance characteristics of online algorithms for the problem of energy and performance efficient dynamic VM consolidation. Also, the competitive ratio of the proposed scheme is analyzed by comparing it with few such consolidation policies on basis of parameters like quality of consolidation, energy consumption and average SLA violation. Finally the paper is concluded in Section 5.

II. RELATED WORK

The context of energy efficient resource management has been shifted to DCs due to growth of virtual computing environments. It is complex to estimate resource demands as diverse application workload are dynamically creating and destroying in Cloud DCs. Current state-of-the-art Cloud infrastructure such as Amazon EC2 [5] neither support energy-efficient resource allocation that considers consumer preference for energy saving schemes, nor utilize sophisticated economic models to set the right incentives for consumers to reveal information about their service demand accurately [6]. Consequently, providers cannot accomplish efficient service allocation, which meets consumer requirements and expectations with regards to their energy saving goal for Green Cloud computing. Here we surveyed few dynamic VM consolidation techniques that involve overload detection based on adaptive utilization thresholds.

Beloglazov and Buyya [12, 16] have presented a novel technique 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. Further this technique is used in dynamic VM consolidation and ensures SLAs. It is based on adaptive utilization thresholds to determination of overload detection. The works proposed includes (i) Median Absolute Deviation (MAD) (ii) InterQuartile Range (IQR) (iii) Local Regression based methods to determine Adaptive Utilization Threshold.The algorithm is heuristically adapted to handle non-stationary workloads. The extensive work has been simulated using CloudSim toolkit with more than a thousand VMs and varied number of hosts.

To predict the future workload and proactively optimize the resource allocation for implementing an energy-aware dynamic VM consolidation framework, the authors [13] applied weighted linear regression. The system mainly focused on web applications and the SLAs are defined in terms of the response time. In another work Bobroff *et al.* [5] proposed a predicting technique to determine overloaded server. To forecast time-series analysis of historical data is used.

In [14], on basis of utilization analysis from resource utilization log and two prediction methods: "Linear Predicting Method" (LPM) and "Flat Period Reservation-Reduced Method" (FPRRM), the resources are allocated dynamically to achieve the energy efficiency objective. On using M/M/1 queuing theory for predicting methods, better response time and less energy-consumption goals are achieved. Experimental evaluation performed on CloudSim [17, 18] simulator to demonstrate the proposed methods. The diverse application workloads that are created at data centers are not considered in this as it is utilizing logs.

Motwani et al. stated that the most common and simplest algorithms, to detect the non-overload and overload states of the hosts involve setting up of CPU utilization threshold. Authors' also stated that these static thresholds are unsuitable for unknown and dynamic workloads, because it do not adapt to workload changes and do not capture the time-averaged behavior. For proposed VM consolidation, a dynamic heuristic of maximum utilization is adapted to detect the host overload, in the work [4, 10]. At the next step proposed VM selection algorithm is iteratively applied until the host is considered as not being overloaded. The proposed policy is named as "Less Migration Time".

Here Adaptive Utilization Threshold based Techniques presented in [15] are studied and briefly described, as these are most relevant to our work. Several heuristics have been proposed in literature for determining over-utilized and under-utilized hosts based on Adaptive Utilization Threshold is also presented by authors in their previous works. Adaptive Utilization Threshold provides better approximations for determination over-utilized and under-utilized hosts. The results reveal that it reduces energy consumption at DCs while maintaining QoS.

III. PROPOSED WORK

For describing the proposed work we first discuss the dynamic VM consolidation as a three step process:

1. Deciding if a server is considered to be under or overloaded, so VMs should be migrated from it, to other active or reactivated hosts to avoid violating the QoS requirements.
2. Selecting VMs for live migration from an overloaded host, as all VMs from under-loaded hosts would be migrated to switch this host to low power mode.
3. Placing the selected VMs on other active or reactivated hosts using live migration.

Here we proposed a dynamic heuristic method for determining CPU utilization threshold. The threshold determined is adaptive and is based on statistical data analysis of collected during the lifetime of VMs. As the workloads are dynamic so the workload data may contain outliers that come from non-normal distributions. The proposed algorithm adjusts the value of the utilization threshold of CPU depending on the strength of the deviation its utilization. The idea of adjusting utilization threshold is based on the concept that: on higher deviation the CPU utilization is likely to be reaching 100% causing an SLA violation. So on higher the deviations the value of upper utilization threshold should be lower.

When the process of detecting overload or under-load detection is invoked, it compares the current CPU utilization with dynamically obtained threshold. Based on this dynamic heuristic of maximum utilization threshold, the algorithm detects a host overload. After determining the adaptive utilization threshold the VM consolidation applies as it is, i.e. VM selection and Placement.

A. About Proposed Quartile Deviation or Semi-Inter Quartile Range

Quartile deviation is a measure of dispersion based on upper quartile (Q_3) and lower quartile (Q_1) of a series. The difference is the range between the two quartiles and is called InterQuartile range. While the Quartile deviation is half of the difference between the two quartiles (upper and lower quartile). The Quartile Deviation is also known as Semi-InterQuartile Range (see Equation 3.1).

$$\text{Inter Quartile Range} = | Q_3 - Q_1 |$$

$$\text{Semi InterQuartile Range} = \left|\frac{Q3-Q1}{2}\right| \qquad \dots(1)$$

Using IQD, The CPU utilization threshold calculated on basis of equation as defined below:

$$Tu = 1 - s.\ IQD \qquad \dots(2)$$

Where ***s*** is the safety parameter and it defines, how strongly the system tolerates host overloads.

The safety of the method is adjusted using this parameter ***s***. If ***s*** is having lower value than higher the tolerance to variation in the CPU utilization. This would possible cause increasing the inefficient utilization and SLA violation caused by VM consolidation. After threshold calculation, the algorithm works similarly to the static threshold algorithm by comparing the current CPU utilization with the calculated threshold.

B. VM Selection Policies

The next step after the determination of overloaded host is determination of the particular VMs to migrate from that host. This problem of selection is solved by iteratively applying VM selection algorithms until the host is considered as not being overloaded. The VM selection policy used with proposed Load Detection Technique are Minimum Utilization (Mu) and Minimum Migration Time (Mmt).

The described policies are applied iteratively until the host is considered as being not overloaded. Our framework outperformed by the use of proposed policy with two of the VM selection policies.

C. VM Placement

For VM placement it is reasonable to apply a heuristic, such as the Best Fit Decreasing (BFD) algorithm [7], which has been shown to use no more than 11/9. OPT + 1 bins (where OPT is the number of bins provided by the optimal solution) [8]. Considering N is the number of hosts, the complexity of the algorithm is 2N.

IV. EXPERIMENTAL SETUP & RESULT ANALYSIS

A. Experimental Setup

From various studies using CloudSim [17, 18] simulation software involving PlanetLab [5] workload, we found that the performance of proposed work showed significant improvements in terms of energy, over existing OpenStack Neat ‘s [11] ‘Adaptive threshold-based:Median Absolute Deviation (MAD) and InterQuartile Range (IQR)’ overload detection algorithm along with (Mmt) and (Mu) VM selection algorithm respectively. All are also compared with DVFS (Dynamic Voltage and Frequency Scaling) [19]. The technique stands distinctive in improving three parameters including energy consumption at hostswhile maintaining SLAs.

The general simulation parameter involves Planet Lab workload of 03.March.2011. The number of Hosts and VM are 800 and 1052 respectively. The simulation ran for 24 hours with scheduling interval of 15 minutes. Other configurationsof CPUs and VMs are adapted from [4, 10]

B. Result Analysis

The Results along with Performance Metrics are discussed below:However, most of these metrics are valid for other types of service performances also. Table 1 shows the total energy consumption, mean time before a VM Migration and average SLA violation.

Total Energy Consumption: It is defined overall energy consumed by the physical resources of a DC as a result of running application workloads.

Mean Time Intervals between VM Migrations: It determines the quality of consolidation as it is inversely proportional to number of active physical hosts. Lesser the number of active physical hosts lesser will be the energy consumption and better the consolidation.

Average SLA Violation: It is the average SLA violation caused during overall life cycle of running application workloads. The Consolidation Policies should not affect on this.

Table 1: Comparison of Policies on Various Parameters.

Parameters	**DVFS**	**IqrMu**	**Proposed IqdMu**	**MadMmt**	**Proposed IqdMmt**
Energy Consumption (kWh)	817.6	212	198	201	178
Mean Time before a VM Migration (sec)	0	19.66	19.68	15.04	17.47
Average SLA violation		10.16	9.90	10.05	9.60

V. CONCLUSION & FUTURE SCOPE

Today's computational needs at Cloud DCs resulted in greater power consumption, increased operational costs and high carbon emissions in environment. Maintaining power-performance trade-off by satisfying high Quality of Service (QoS) and Service Level Agreements (SLA) is one of the major challenges faced by cloud stakeholders. The work studied and explored energy-efficient cloud computing, compare and analyzes various SLA and Energy-Efficient Dynamic Virtual Machine (VM) Consolidation. In this work few statistical techniques for detecting overload and under-loaded hosts that are part of Energy-Efficient VM Consolidation Policies are also studied. The main focus of this work is to propose the 'Adaptive Load Detection Technique' to detect the overloaded and under-loaded hosts. The load detection technique proposed in this work is adaptive and based on a famous statistical method: InterQuartile Deviation. When this load detection is applied along with few VM selection approaches for consolidating VMs, better optimization in energy efficiency and performance is obtained.

The performance evaluation of proposed work, on basis of various parameters is done under Cloud Simulator tool (CloudSim version 3.0.2). The parameter also includes a parameter that determines the quality of consolidation. Better VM consolidation quality, energy optimization along with lesser value of average SLA violation is finally attained. The work has social significance as it is reducing carbon footprints.

As a future scope, the work suggests to study more such approaches of determining host overload and under-load. The work suggests studying the performance of proposed and other energy efficient VM consolidation techniques over varied workloads for scalability and QoS objective. The work motivates to perform competitive analysis of these algorithms and addressing the individual problems of VM selection and placement under VM Consolidation.The proposed framework would be further implemented and tested with real cloud platforms like Open Stack.

REFERENCES

[1]. R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, "Cloud computing emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, *Future Gener. Comput." Syst.* **25**(6) (2009) 599–616, Elsevier Science, Amsterdam, the Netherlands.

[2]. Belady C. "In the data center, power and cooling costs more than the it equipment it supports" 2007. URL: http://www.electronics-cooling.com/articles/2007/feb/a3/.

[3]. Clark C, Fraser K, Hand S, Hansen JG, Jul E, Limpach C, Pratt I, Warfield A " Live migration of virtual machines. Proceedings of the 2nd Symposium on Networked Systems Design and Implementation" (NSDI 2005), USENIX, Boston, MA, USA, 2005.

[4]. AnandMotwani, Vaibhav Patel and Vijay M Patil "Power and QoS Aware Virtual Machine Consolidation in Green Cloud Data Center" *International Journal of Electrical, Electronics and Computer Engineeing* **4**(1): 93-96(2015).

[5]. N. Bobroff, A. Kochut, and K. Beaty, "Dynamic Placement of Virtual Machines for Managing SLA Violations," *Proc. IFIP/ IEEE 10th Int'l Symp. Integrated Network Management (IM)*, pp. 119-128, 2007.

[6]. Verma A, Ahuja P, Neogi A. pMapper: Power and migration cost aware application placement in virtualized systems. Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware (Middleware 2008), Springer, Leuven, Belgium, 2008; 243–264.

[7]. National Institute of Standards and Technology (2011) NIST cloud computing reference architecture: Version 1. NIST Meeting Report.

[8]. Amazon. Amazon Elastic Compute Cloud (EC2). http://aws.amazon.com/ec2/.

[9]. IBM Bluemix. https://console.ng.bluemix.net/.

[10]. HeenaKaushar, PankajRicchariya, AnandMotwani, "Comparison of SLA based Energy Efficient Dynamic Virtual Machine Consolidation Algorithms", *International Journal of Computer Applications* (0975 – 8887), Volume **102**– No.16, September 2014.

[11]. SearchCIO (2009) Amazon gets SAS 70 Type II audit stamp, but analysts not satisfied. http://searchcloudcomputing.techtarget.com/news/1374629/Amazon-gets-SAS-70-Type-II-audit-stamp-but-analysts-notsatisfied, Accessed 11 Aug 2012.

[12]. 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.

[13]. B. Guenter, N. Jain, and C. Williams, "Managing Cost, Performance, and Reliability Tradeoffs for Energy-Aware Server Provisioning," *Proc. IEEE INFOCOM*, pp. 1332-1340, 2011.

[14]. Yuxiang Shi; Xiaohong Jiang; Kejiang Ye, "An Energy-Efficient Scheme for Cloud Resource Provisioning Based on CloudSim," Cluster Computing (CLUSTER), 2011 *IEEE International Conference on*, vol., no., pp.595, 599, 26-30 Sept. 2011.

[15]. 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 Inter Science, Concurr. Comput. : Pract. Exper.*, **24**(13):1397-1420, September 2012.

[16]. Anton Beloglazov and RajkumarBuyya, "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.

[17]. 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. **11**, no., pp.1, , 21-24 June 2009.

[18]. Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, Cesar A. F. De Rose, and RajkumarBuyya, "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.

[19]. Etienne Le Sueur and Gernot Heiser, "Dynamic Voltage and Frequency Scaling: The Laws of Diminishing Returns", NICTA and University of New South Wales.