Supported by tetfund TERTIARY EDUCATION TRUST FUND

AN ENERGY-AWARE CLASS-BASED LOAD BALANCING MODEL FOR RESOURCE MANAGEMENT IN CLOUD COMPUTING

Olasupo O. Ajayi*, Florence A. Oladeji, Charles O. Uwadia
Computer Sciences Department, University of Lagos, Nigeria
*Corresponding author: olaajayi@unilag.edu.ng

Received: December 10, 2016 **Accepted:** February 28, 2017

Abstract: Cloud Computing is a model for enabling ubiquitous and on-demand access to shared resource pool. It represents a paradigm shift from traditional personal computing to computing as a pay-per-use utility. Cloud Computing is not without its challenges and despite tremendous progress in recent years, issues relating to security, resource provisioning and high availability still continue to plague it. In this paper, we focus on multi-objective resource management schemes; which are schemes that seek to manage multiple system or user requirements with little or no compromises. Multi-objective in Cloud Computing may include guaranteeing resource availability while adhering to Service Level Agreements or effectively utilizing resources while conserving energy. These objectives are usually divergent and prove a challenge for researchers as an improvement in one objective usually results in a corresponding wane in another or several others. We therefore propose a new approach using class-based migration policy for resource management, which is able to evenly balance workloads among systems and better conserve energy. Results of simulations carried out and compared to the state of the art, show that the proposed approach conserved energy and balances workloads better.

Keywords: Cloud computing, energy conservation, load balancing, migration, workload consolidation

Introduction

Currently computing and data are moving from onsite computers such as desktops, personal computers to large data centers located in geographically dispersed locations around the world (Sidhu and Kinger, 2013). There is a paradigm shift in the way computers and computing resources are being used. Today computing is now being offered and used as a commercial resource whereby users pay the provider on a pay-as-you-use model, similar to other utilities such as electricity, water, gas, etc. (Voorsluys *et al*., 2011). Buyya *et al*. (2009a) defines a Cloud as a parallel and distributed computing system consisting of a pool of inter-connected and virtualized computers that are dynamically provisioned and presented a single computing resource to the users based on pre-agreed Service Level Agreements (SLA). These pools of computing resources are made available either at a hardware level - Infrastructure As A Service (IAAS), at the software level – Software As A Service (SAAS) or at a developer level - Platform As A Service (PAAS) and deployed either as a private, public, community or hybrid model (Mell and Grance, 2009). Cloud Computing leverages on the following technology for its operation namely: virtualization (Zhang *et al*., 2014); grid computing (Qusay, 2011) and utility computing.

As with most things in life, there is no perfect system and Cloud Computing (CC) is one such. Ajayi and Oladeji (2015) reviewed some of the challenges faced in CC, however the focus of this study is on adherence to pre-agreed Quality of Service (QoS) constraints while aligning with drives for energy conservation in Cloud data centers. The rest of this paper is organized as follows; in section 2 a literature review of various related works is presented; section 3 discusses multi-objective based resource management. In section 4, we propose a multi-objective based resource management technique; while in section 5 results of experimental.

Resources in CC include but are not limited to CPU, RAM, Storage and Network Bandwidth (Membrey *et al*., 2012). Managing, allocating or sharing these resources dynamically are major challenges in cloud computing. Various authors have approached these challenges from various perspectives some of which are as follows:

Virtualization and virtual machine migration

The ability to configure customized Virtual Machines (VMs) to utilize different partitions of resources on a physical host machine is one of the greatest benefits of VMs in resource provisioning. These VMs being independent of each other can be started and stopped dynamically by users thus able to meet the ever changing demand level of resources prominent with Cloud environments (Buyya *et al*., 2011).

Candler (2014) defines Virtual Machine Migration (VMM) as moving a VM from one host Physical Machine (PM) to another and is usually done for two main reasons - load balancing and host maintenance. A model for resource allocation and optimization using simulated annealing was proposed by Akshat and Sanchita, (2014). The proposed model was one in which VMs were migrated based on pre-set threshold value. Power consumptions levels were used as basis for VM migration. These migrations were from PMs that consumed above or below the Maximum Utilization Threshold and Minimum Utilization Threshold respectively to other PMs. In this work, the cost and effect of VM migrations were not considered, CPU utilization alone was used as the sole criteria for measuring utilization; and the authors assumed a linear relationship between CPU utilization and power consumption, an assumption which has been debunked by Principled Technologies (2011). Zhang *et al*. (2014) explored resource allocation in CC via randomized combinatorial auctioning of VM instances and proved that dynamic resource provisioning is more efficient than static resource provisioning. An online combinatorial auction for allocation of VMs in CC was proposed by (Shi *et al*., 2014), using primal-dual algorithm, randomized auction sub-framework and greedy primal-dual algorithm for load balancing. Farahnakian*et al*. (2016), presented a hybrid combination of usage prediction and Ant-Colony System for VM allocations and migrations within a data center in an energy efficient manager. The authors classified VM allocations as NP-hard problem and thus applied the meta-heuristics ACS. Obtained results show improvement in energy conservation and resource utilization versus a heuristic approach. Salfner *et al*. (2011) conducted experiments to calculate downtime and effect of live migrations on applications running on migrated virtual machines. It was concluded that the memory load and memory access pattern of the guest systems are the most important factors to be considered for VMM. Principled Technologies (2011) carried out experiments on the migration time of live VMM on two products – vSphere a product of VMWare and Microsoft

Hyper-V and reported that vSphere was about five times faster than Microsoft's Hyper-V and in comparison to vSphere, more crashes were experienced with Hyper-V. The author concluded that VMM is still not perfect and a lot of downtimes and system failures still occur in live environments.

Load balancing

Effatparvar and Garshasbi (2014) defined load balancing as the technique for spreading work between multiple computing resources for the purpose of optimizing resource utilization, improving throughput and response time. It ensures that workloads are evenly distributed across all the PMs in a data center to avoid a situation where some nodes are overworked while others are idle. Dhinesh and Krishna (2014) proposed a nature inspired load balancing technique based on the behavior of a colony of honeybees foraging for food. It was reported that this technique is best suited for scenarios of heterogeneous service request types however it can lead to starvation since it is a priority based algorithm. Mahajan *et al*. (2013) discussed on a variant of round-robin called Round-Robin with Server Affinity (RRSA), which distributes workloads using the conventional round robin algorithm; however with the introduction of a hash map and PM state list which stores information about the last allocated PM and the current state of the PMs respectively. A slightly better response and processing time was recorded in comparison with conventional Round Robin but performance was lost during the process of searching the hash map. Mishra and Jaiswal (2012) described a load balancing mechanism based on ACO, in which artificial ants traverse a network searching for overloaded and under-loaded servers, after which workload is transferred from the most over loaded PMs to idle ones. Bermejo *et al*., (2016) proposed a dual level approach to resource management. In their work, load balancing decisions are taken both within the PM and by a global controller. The PMs are autonomous and manage their local resource levels independently. Each PM then sends necessary load balancing information to the global controller which analysis all received information and makes informed decisions for the global allocation or re-allocation of workloads with a view of stabilizing the entire system. High inter-nodal control messages compete with actual workload for network bandwidth. Also the need to get updates from all PMs prior to workload allocation and migration can impact on QoS, especially if such updates arrives late. Numerous other authors have proposed various load balancing approaches some of which have been analyzed in (Ajayi *et al*., 2015).

Energy management

The issue of energy consumption in information technology equipment has in recent times been receiving increasing attention. Statistics and report have shown that if unchecked, energy consumption by IT could become a major problem in the not too distant future. Some of such reports include: a growth of 56% in electricity consumption of data centers between 2005 and 2010, as reported by Stanford University, (Koomey, 2011); an annual increase of 16% in IT energy cost has also been reported by McKinsey (2010). These reports show that serious attention must be paid to energy management in cloud computing. Baliga *et al*. (2011) compared the energy consumption of cloud services against traditional PC and reported that the largest amount of energy was consumed during transporting data between the users and the Cloud infrastructure. It was also concluded that even with energy conserving techniques Cloud Computing still consume much more energy than traditional PCs, for processor and data intensive tasks. Google (2012), reported an up to 90% and 85% reduction on energy consumption and carbon emissions respectively in its datacenters; however these were based on Google Apps Engine (GAE), which is primarily designed for light weight office tasks such as emails, calendaring, word processing and spreadsheet. NRDC (2012) shows a large carbon efficiency gains from moving server functions from on-premise to a public cloud, however the report only took office productivity applications into consideration and did not consider applications with huge computational demands.

Multi-objective based resource management models

In recent times, research works on resource management have been geared towards managing multiple objectives and the introduction of hard and soft objectives. We defined hard objectives as primary objectives that must be met or adhered to, while soft objectives are secondary objectives that are achieved or compromised in the process of achieving the primary objective.

Das *et al*. (2013) proposed an adaptive VM provisioning approach to managing QoS in cloud computing. The hard objective in this work was managing the QoS (agreed service time) of admitted user jobs while the soft objective was managing the number of VMs in the system through VM recycling. The authors recorded slight improvement with this approach when compared with the conventional analytical technique for VM provisioning; however the extra processing cycles introduced when searching for suitable VMs to re-host a workload were not considered. Also in classifying user workload, no limit was put on the number of queues that can be created, which could invariable lead to a large number of queues with only few VMs in them. In the work of (Beloglazov *et al*., 2012), the hard objective was improving the overall energy conservation of Cloud data centers while the soft objective was maintaining agreed Service Level Agreements (SLA). These objectives are contrasting in nature, thus the authors proposed an approach to achieving the hard objective with minimal violation of the equally important soft objective.

In this work, CPU utilization was the only yardstick used for measurement compliance. The work relied heavily on PM probes during allocation which theoretically can lead to an increased overall response time. In the work proposed by Hieu *et al*. (2015), the authors focused on maintenance of Quality of Service (QoS) as the hard objective, while energy conservation and reduction in SLA violations were soft objectives. A usage prediction algorithm was introduced that used local regression to determine the short-term future utilization levels hence the authors were able to proactively take actions to prevent SLA violations from occurring. Increased response time is also a potential drawback of this approach. Mosa and Paton (2016) in their work presented a utility function based VM allocation approach for profit maximization through efficient resource utilization (hard objective) and energy conservation with SLA adherence as soft objectives. The work identifies optimal allocation of VMs to PMs as a NP-hard problem and thus used a meta-heuristic genetic algorithm to achieve this goal in the most rewarding (profitable) way. The authors employed a utility factor which was based on expected income less estimated energy, violation and performance degradation costs. The approach recorded improvements in terms of QoS adherence and energy conservation but did not pay attention to resource utilization. Hu *et al*. (2016), presented a Service-Oriented Resource management scheme, in which workloads were classified into various groups and VMs were dynamically adjusted based on requirements. The trust of the work was on improving SLA adherence and resource utilization.

Proposed Method

We propose a multi-objective class-based approach to resource management in Cloud Computing, with energy conservation as hard objective and efficient PM utilization via load balancing of workloads as the soft objective.

A. Allocation of Workload

In our approach, like that of Beloglazov *et al*. (2012) and Hieu *et al*. (2015), workloads are allocated to servers using a variant of Best Fit Descending (BFD) algorithm. In allocating these workloads, we introduce a concept of workload classification similar to that of Das *et al.* (2013). However our approach does not support VM reuse. We classify workloads into three classes viz. Gold, Silver and Bronze based on agreed SLA criteria, with Gold having highest priority, followed by silver and bronze as best effort.

B. Load Balancing

Once the user workloads have been successfully allocated, the process of load balancing the workloads across PMs is initiated. Load balancing of workload ensures that certain servers are not underworked at the detriment of others that are overworked. To this end, we propose a VM migration scheme that is based on the class the workload belongs. That is, Gold, Silver and Bronze, such that:

$$\text{Bronze} < \text{Silver} < \text{Gold} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Workloads on a PM must belong to one of the three classes and Eq. (1) implies that on selecting a workload for migration all bronze class workloads must be selected first before any silver class can be selected and all silver class must be selected for migration before any gold class workload can be selected. Once a PM has been identified as overworked, the VM migration is activated and a VM has to be selected for migration. This selection can be modeled as follows:

Let B, S and G represent Bronze, Silver and Gold workload classes, respectively. Let N, B_T, S_T and G_T respectively represent the total number of user workloads in the entire system, total number of B, total number of S and total number of G such that N = (B_T + S_T + G_T). Let P_T represent the total number of VMs allocated to a given PM p, such that P_B, P_S and P_G are the numbers of B, S and G in p. Let X be a VM selected for migration (without replacement), the probability of it being B is given by the hyper-geometric distribution:

$$P(X = B) = \frac{\binom{B_T}{P_B} * \binom{N-B_T}{P_T-P_B}}{\binom{N}{P_T}} \quad \ldots\ldots\ldots\ldots(2)$$

Adapted from Weisstein (2003)

The probability of selecting a silver, that is P(X = S), is a conditional probability which can occur if and only if all Bs have previously been selected. This means that the probability of selecting S is dependent on the previous selection being a B or another S (if all Bs had previously been selected) and is modeled using Bayesian model of two elements but with an added condition and described below:

$$P(X = S) = \begin{cases} P(S|B) = \dfrac{P(B \cap S)}{P(B)}\,, & giventhatP(B) > 0 \\ P(S), & giventhatP(B) = 0 \end{cases}$$

…………..(3) (Adapted from Ghahramani, 2013)

The probability of the selection being a G can be modeled using a Bayesian model for three elements and given by:

$$\mathrm{P(X = G)} = \begin{cases} \mathrm{P(G|B \cap S)} = \dfrac{\mathrm{P(B \cap S \cap G)}}{\mathrm{P(B) * P(S|B)}}, & giventhatP(B)andP(S) > 0 \\ P(G), & giventhatP(B)andP(S) = 0 \end{cases}$$

……………(4) (Adapted from Ghahramani, 2013)

In Eq. (3), the probability of any S being selected is dependent on all previous selections being B or another S (if no more B exists). While in Eq. (4), the probability of the selected workload being a G is dependent on the probability of all previous selections being S or G (given that only Gs are left in the server).

C. Proposed Load Balancing Algorithm

Algorithm 1: Load Balancing of Allocated Workload

```
1. Get CPU utilization of PM (p) to determine status
2. Set Upper and Lower Thresholds
3. If CPU Utilization > Upper Threshold value
           Status = OVERUTILIZED
   Elseif CPU Utilization < Lower Threshold
           Status= UNDERUTILIZED
4. If h is UNDERUTILIZED
   a. Perform VM_Mig (p, all VMs in p)
   b. If step a is successfully completed, put h to sleep
5. If h is OVERUTILIZED
   a. Check VM_types of VMs in p
   b. If any VM_type = BRONZE
         Select it for migration
      Else if any VM_type = SILVER
         Select it for migration
      Else if any VM_type = GOLD
         Select it for migration
   c. Perform VM_Mig(p, selected VM in p)
VM_Mig (p, v): Migrate v to other PMs, where possible
```

Implementation and Discussion of Results

Experimental implementations were done using CloudSim framework (Buyya *et al*., 2009b). The simulation environment was made up of a data center containing 800 heterogeneous servers, of 1,860 MIPS and 2,660 MIPS processing capabilities (Table 1). Workload traces from PlanetLab (Park and Pai, 2006) and Google Cluster Dataset (Wilkes and Reiss, 2011) were used to simulate the workload utilization requirements. This is similar to that used in PABFD (Beloglazov and Buyya, 2012) and VMCUP (Hieu *et al*., 2015).

Table 1: Specifications of the PMs used for simulation

Category	Make	CPU	Cores	Memory
1	HP ProLiant ML110 G4	1,860 MHz	Intel Xeon 3040, 2 cores	4GB
2	HP ProLiant ML110 G5	2,600 MHz	Intel Xeon 3075, 2 cores	4GB

Table 2: Summary of results using PlanetLab Dataset

METRIC	PABFD	VMCUP	Proposed
No. of PMs in DC	800	800	800
No. of workloads submitted	1,078	1,078	1,078
Total energy consumption (KWh)	175.43	151.42	105.62
Avg. No. of power state changes	6.82	1.04	1.02
Avg. No. of PM shutdowns during load balancing	5,456	831	819

Table 3: Summary of results using Google Cluster Dataset (GCD)

METRIC	PABFD	VMCUP	Proposed
No. of PMs in DC	800	800	800
No. of workloads submitted	168	168	168
Total Energy Consumption (KWh)	11.1	10.28	6.33
Avg. No. of Power State Changes	1.33	1.00	1.00
Avg. No. of PM shutdowns during load balancing	1067	447	441

Obtained results for PlanetLab dataset and GCD are summarized in Tables 1 and 2, respectively. The energy consumption levels of PABFD and VMCUP were compared against our proposed model using upper threshold of 80% utilization. Figs. 1 and 2 show that our proposed approach has the lowest energy value compared to other approaches with energy consumption of 105 KWh versus PABFD and VMCUP at 175 and 151KWh, respectively for PlanetLab dataset. Same trend is observed with GCD with our approach consuming 6.33 KWh as against 11.1 and 10.28 KWh for PABFD and VMCUP, respectively. In terms number of Power State Change (PSC) per Physical Machine; Figs. 3 and 4 show that our model outperformed PABFD in terms of the. PSC is a measure of how often PMs are switched off or on with respect to workload migration. A high PSC value implies an indiscriminate migration and; thus lower values are desirable. Fig. 3 shows PABFD having 6.82, VMCUP having 1.04 while our model has 1.02. Similarly, in Fig. 4, PABFD 1.33 while VMCUP and our model both have 1.00. These results imply that our proposed model was able to consolidate VMs efficiently onto PMs and thus reducing the number of times PMs had to be switched on or off. In comparison to VMCUP, our approach is at par.

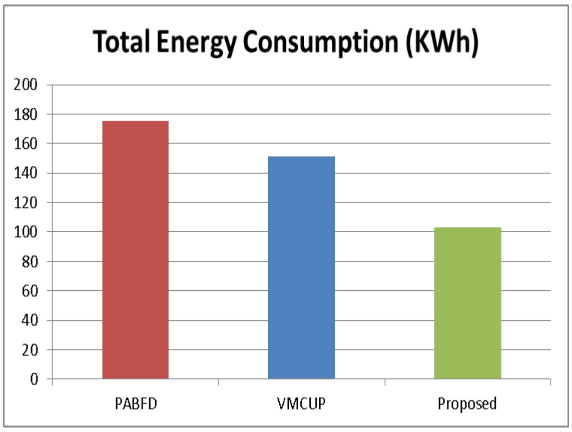

Fig. 1: Energy consumption: Proposed model vs. PABFD and VMCUP (PlanetLab dataset)

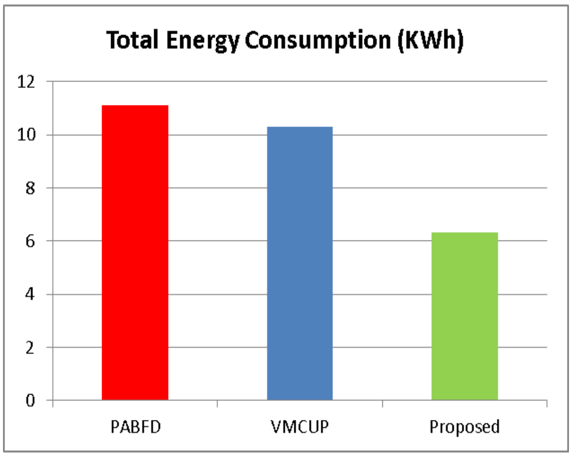

Fig. 2: Energy consumption: Proposed model vs. PABFD and VMCUP (GCD)

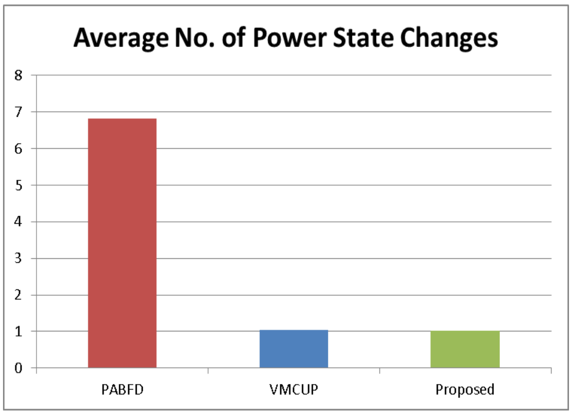

Fig. 3: PSC: Proposed model vs. PABFD and VMCUP (PlanetLab dataset)

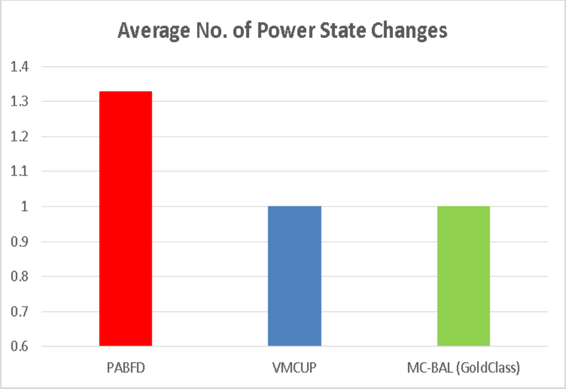

Fig. 4: PSC: Proposed model vs. PABFD and VMCUP (GCD)

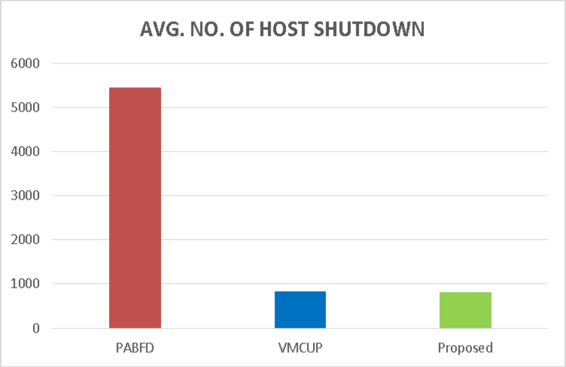

Fig. 5: Average No. of host shutdown: Proposed model vs. PABFD and VMCUP (PlanetLab dataset)

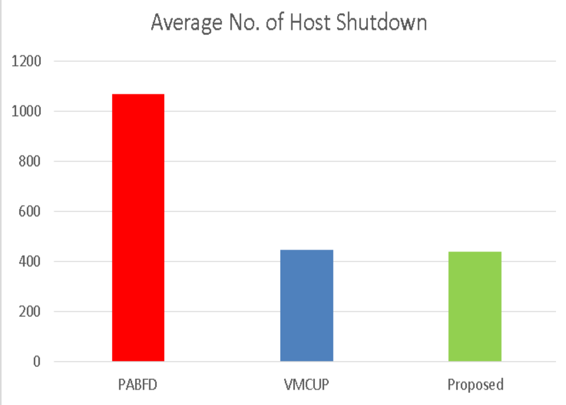

Fig. 6: Average no. of host shutdown: proposed model vs. PABFD and VMCUP (GCD)

Figures 5 and 6 depict the cumulative number of times PMs were shut down in the datacenter. Fig. 5 shows that PABFD resulted in 5,456 PM shutdowns while testing load balancing with PlanetLab dataset, VMCUP resulted in 825 shutdowns while our approach had 819 shutdowns. A similar trend is observed in Fig. 6, where our proposed approach resulted in 441 PM shutdowns versus 1,067 for PABFD and 447 for VMCUP. Lower values are desirable and imply a more efficient load balancing scheme.

These results imply that our proposed class-based model is able to consolidate workloads efficiently onto PMs. Resulting in a reduction in the number of actively running PMs and consequently the overall energy consumption of the data center.

Conclusion

CC is not entirely new, only recently has it started gaining popularity both at a personal and enterprise/commercial level.

Despite its bells and whistles however, the issue of resource management in CC is still a major source of concern. Resource management in Cloud environment is concerned with efficiency of resource usage without violating pre-set service level objectives and constraints. Of concern also is balancing the overall performance levels with power consumption; especially with the clamor for green computing and the need for data centers to reduce their carbon emission footprints, in a bid to save the earth. In this paper, an efficient resource management scheme that uses workload classes to balance PMs in manner to conserves energy was proposed. Results from simulations show that the proposed approach was able to do achieve the set objective while competing effectively with the state of the art approaches. In the future, we seek to determine how our approach deals with adherence to Service Level Agreements and the effect it would have if applied to the workload allocation phase.

References

Ajayi O & Oladeji F 2015. An overview of resource management challenges in Cloud computing. Book of Proceedings, 10th Unilag Annual Research Conference & Fair, 2: 554-560.

Ajayi O, Oladeji F &Uwadia C 2015. Analysis of two-phased approaches to load balancing in Cloud computing.*J. Computer Sci.&Its Applic.*, 22(2): 123-131.

Akshat D & Sanchita P 2014. Green Cloud: Smart resource allocation and optimization using simulated annealing technique. *Indian J. Computer Sci.&Engr. (IJCSE),* 5(2): 0975-5166.

Baliga J, Ayre R, Hinton K & Tucker R 2011. Green Cloud computing: Balancing energy in processing, storage, and transport. *Proceedings of IEEE,* 99(1): 149-167.

Beloglazov A, Abawajy J & Buyya R 2012. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing.*Future Generation Computing Systems,* 28(5): 755-768.

Beloglazov A& Buyya R 2012. 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. pp. 1397–1420.

Bermejo B, Guerrero C, Lera I&Juiz C 2016. Cloud Resource Management to Improve Energy Efficiency Based on Local Node Optimization. 6th Intl. Conference on Sustainable Energy Information Technology (SEIT, 2016), *Procedia Computer Science,* 83: 878-885.

Buyya R, Garg S & Calheiros R 2011. SLA-Oriented Resource Provisioning for Cloud Computing. *International Conference on Cloud and Service Computing* (CSC), IEEE, pp.1-10.

Buyya R, Yeo C, Venugopal S, Broberg J& Brandic I 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems. *J. Future Generation Computer Sci*., 25(6): 599-616.

Buyya R, Ranjan R& Calheiros R 2009. Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: Challenges and opportunities, Proc. of the 7th High Performance Computing and Simulation Conference (HPCS'09), IEEE Press, pp. 1-11.

Candler B 2014. Virtual Machine Migration, Network Startup Resource Center. [Online] Available at www.nsrc.org/workshops.

Das AK, Adhikary T, Razzaque A & Hong C 2013. An Intelligent Approach for Virtual Machine and QoS Provisioning in Cloud Computing. *Information Networking (ICOIN) IEEE,* pp. 462-467.

Dhinesh, B. & Krishna P. 2014. Honey bee behavior inspired load balancing of tasks in Cloud computing environment. *Applied Soft Computing*, 13(5) 2292-2303

Effatparvar M & Garshasbi M 2014. A genetic algorithm for static load balancing in parallel heterogeneous systems. *Int. Conference on Innov. Mgt&Techn. Res.*, 129:358-364.

Farahnakian F, Pahikkala T, Liljeberg P, Plosila J, Hieu N& Tenhunen H 2016. Energy-Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model. IEEE Transactions on Cloud Computing.

Ghahramani Z 2013. Bayesian non-parametrics and the probabilistic approach to modelling, Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, vol. 371 no. 1984, 20110553. http://doi.org/10.1098/rsta.2011.0553

Google 2012. Google Apps: Energy Efficiency in the Cloud. [Online] Available at www.google.com/green/pdf/google-apps.pdf

Hieu N, Francesco M &Yla-Jaaski A 2015. 'Virtual Machine Consolidation with Usage Prediction for Energy-Efficient Cloud Data Centers*'*,Proc. of 8th International Conference on Cloud Computing, IEEE, pp.750-757

Hu X, Zhang R &Wang Q 2016. Service-Oriented Resource Management in Cloud Platforms. Intl. Conf. on Service Computing (SCC), IEEE, pp. 435-442.

Koomey J 2011. Growth of Data Center Electricity use 2005 – 2010. A report by Analytics Press, completed at the request of The New York Times. [Online] Available at http://www.analyticspress.com/datacenters.html.

Lu Y, Xie G, Kliot G, Geller Larus J &Greenberg R 2011. Join-Idle-Queue: A Novel Load Balancing Algorithm for Dynamically Scalable Web Services. *ACM J. Performance Evaluation*, 68(11): 1056-1071.

Mahajan, K., Makroo, & Dahiya, D. 2013. Round Robin with Server Affinity: A VM Load Balancing Algorithm for Cloud Based Infrastructure. *J. Inf. Process Sys*., 9(3): 379-394.

McKinsey MN 2010. Energy Efficiency: A Compelling Global Resource. [Online] Available at http://mckinseyonsociety.com/energy-efficiency-a-compelling-global-resource/

Membrey P, Plugge E & Hows D 2012. Practical Load Balancing Ride the Performance Tiger, Apress.

Mell P & Grance T 2009. The NIST Definition of Cloud Computing, National Institute of Standards and Technology, Information Technology Laboratory, Technical Report v 15, 2.

Mishra R &Jaiswal A 2012. Ant colony optimization: A solution of load balancing in Cloud.*Intl. J. Web & Semantic Techn.*, 3(2): 33-50.

Mosa, A. and Paton, N. 2016. Optimizing virtual machine placement for energy and SLA in Clouds using utilization. *J. Cloud Computing: Adva. Sys.&Applic.*, 5(1): 67.

NRDC 2012. Sustainability and an Ethical Imperative, NRDC Sustainability Report.

Park K &Pai V 2006. CoMon: A mostly-scalable monitoring system for PlanetLab. *ACM SIGOPS Operating Systems Review*, 40(1): 65–74.

Principled Technologies 2011. Virtual Machine Migration Comparison: VMWare VSphere vs Microsoft Hyper-v. [Online] Available at www.vmware.com/files/pdf/vmw-vmotion-verus-live-migration.pdf

Qusay F 2011. Demystifying Cloud Computing. [Online]. Available at www.crosstalkonline.org/storage/issue-archives/.../201101-Hassan.pdf

Salfner F, Troger P & Polze 2011. Downtime Analysis of Virtual Machine Live Migration. 4th Intl. Conf. on Dependability. DEPEND.

Shi W, Zhang L, Wu C, Li Z, & Lau F 2014. An Online Auction Framework for Dynamic Resource Provisioning in Cloud Computing, SIGMETRICS, Austin Texas, USA.

Sidhu P& Kinger S 2013. Analysis of load balancing techniques in Cloud computing. *Intl. J. Computers & Techn*,. 4(2): 737-741.

Voorsluys W, Broberg J & Buyya R 2011. Cloud Computing: Principles and Paradigms John Wiley & Sons, Inc.

Weisstein E2003. 'Hypergeometric Distribution', Sigma, 37, pp.38. [Online] Available at http://mathworld.wolfram.com/HypergeometricDistribution.[Accessed 25 November, 2015].

Wilkes J & Reiss C 2011. Google Cluster Usage Traces: Format + Schema of Google Workloads. [Online] Available at http://code.google.com/p/googleclusterdata/ [Accessed 25 November, 2015].

Zhang L, Li Z & Wu C 2014. Dynamic Resource Provisioning in Cloud Computing: A Randomized Auction Approach. INFOCOM, Proceedings IEEE, 433 – 441.