Article can be accessed online at http://www.publishingindia.com

SLA Aware Approach for Virtual Machine Placement in Cloud Datacenter

Surinder Singh Khurana*, Nirmaljeet Kaur**

Abstract

The massive adoption of cloud computing by IT industry has caused adramatic increase in energy consumption and its impact on the environment in terms of carbon footprints. The high cost of energy and link between energy consumption and carbon emission have raised an issue of energy management. Recently, a number of approaches have been proposed to address this issue. Most of these approaches are based upon the proper utilisation of the resources. On the other side, to avoid Service Level Agreement (SLA) violations, the servers deployed in data centres can be utilised up to a maximum threshold value. To avoid SLA violation, some virtual machines (VMs) will be migrated from over utilised hosts. Although there are many techniques to select such VMs that can be migrated, but no technique considers Service Level Agreement parameters while selecting such machines. In this paper, SLA based approach to select VMs to migrate from overloaded host has been purposed. The purposed approach will work on the basis of VM categories that are negotiated in service level agreement. The simulation results demonstrate that the proposed approach demands slightly more energy than the other efficient technique (Minimization of Migration). However, it reduces the cost paid in lieu of SLA violation as it reduces SLA violations belonging to VM categories for which SLA violation penalties are higher.

Keywords: Cloud Computing, Energy Consumption, Data Centre, SLA, Virtual Machine

Introduction

With the advancements in internet technology, computing has now become the utility for every task in the real world. The vast growth of internet also demands huge data storage and computing power capabilities. Hence every IT company needs to set up a large data centre. Setting up and maintaining the data centre becomesa problem for every medium scale IT Company, not only due to the initial cost but also due to maintenance and operating cost. These companies have to pay large amount of money for electricity bills. The low utilisation of resources in data centre lead to the invention of sharing based computing model. In 2006, the concept of cloud computing (Bogatin, 2006) comes in to the picture as utility based model and in recent years it has captured a large share of IT market.

It defines the use of computing resources like hardware and software which are available in a remote location and accessible over Internet using pay per use model. By using cloud computing, the computing services have been shifted from the local computer to remote servers. This makes the enterprise to use the resources (network, server, storage, application, service etc.), without huge investment on its purchase, implementation, and maintenance. Cloud computing is not only cost efficient but also saves the environment by reducing the energy requirements. Even the overall energy consumption has been reduced by sharing of resources of cloud by different

* Assistant Professor, Centre for Computer Science & Technology, Central University of Punjab, Bathinda, Punjab, India. E-mail: surinder.seeker@gmail.com

** Student, M.Tech in CST, Centre for Computer Science & Technology, Central University of Punjab, Bathinda, Punjab, India. E-mail: nirmaljeet107@gmail.com

companies but still a huge amount of electrical energy has been consumed by data centres hosting cloud applications. A study (Venkatraman, 2012) found, on an average, 63% of growth in data centre energy demands. The increasing demand and cost of energy is a big difficulty for cloud vendors because it increases total cost of ownership and decreases return on investment. A very small portion of total electrical consumption has been utilised in the cloud data centre. The reason behind this energy wastage is to provide some spare computing facilities to reduce the SLA violation in dynamic computing service. A lot of work has been done to make the cloud data centre more energy efficient and to reduce the SLA violations. But still datacentres consume a large amount of energy, so there is need to improve the components of cloud computing to make it energy efficient and SLA responsive.

Related Work

To make cloud data centre energy efficient, many techniques have been proposed. These techniques can broadly be classified as:

Dynamic Voltage and Frequency Scaling based Technique

Dynamic voltage and frequency scaling (DVFS) is a commonly used power-management technique where the clock frequency of a processor will decrease during low system utilisation to reduce power consumption (Dharwar *et al.*, 2012).

Von Laszewski *et al.* (2009). have proposed a scheduling algorithm for DFVS enabled clusters for executing multiple virtual machines. This scheduling algorithm is successful to reduce the power consumption of a DVFS-enabled cluster. In contrast with a fully utilized cluster using the highest voltages, this technique suggests to use processors with lower voltages with DVFS- enabled clustering.

Rizvandi *et al.*(2011) have considered energy issue with task scheduling in clusters and presented the MVFS-DVFS algorithm. They considermore than one voltage-frequency to reduce energy consumption on processors. Provided results advocate that the optimal energy in a discrete set of voltage-frequencies for each task is achieved by a combination of two voltage-frequencies. The MVFS-DVFS consumes the least amount of energy among other DVFS based alternatives. Only DVFS enabled processors have been considered by this technique.

DVFS technique has been used in large scale computing communities, for example, cluster or grid computing and supercomputing, to minimise power consumption and achieve high performance levels.

Recently, Wu *et al.* (2014) introduced a scheduling algorithm which works based on the use of priority job scheduling for cloud computing. To control the supply voltage and frequency for servers they use DVFS controller. DVFS controller minimises the power consumption by giving applicable amount of frequencies for each server by adjusting to the preferred gear. They also have presented frequency-voltage pairs with corresponding gears. The DVFS controller sets the appropriate gear to supply the proper frequency and voltage to each server. This technique can reduce the energy consumption of a server only when it is in the idle mode or in light workload. Even though the use of DVFS may reduce the energy requirements but still many issues (like performance degradation and predicting future load etc.) associated with this technique remains unresolved and need attention.

Server Consolidation based technique

Server consolidation is process of combining services running on different servers into a lesser number of powerful server machines. Many organisations are moving towards server consolidation to reduce infrastructure complexity, improve system availability and save money. In case of change in user's demand, VMs can be resized and migrated to other physical servers if necessary (Ferreto *et al.*, 2011). In this work they used sever consolidation technique. Their approach prioritizes virtual machines with steady capacity. Avoidance of VM migrations with steady capacity reduces the number of migrations with minimal penalty in the number of physical servers. Algorithm focus on reducing the number of physical servers required. Therefore, when there is server with a single VM that can be accommodated in another server, VM will be migrated and server will be put in off mode. This technique will only enable the migration if the VM requires a change in its capacity; otherwise it will remain in the same server. Another approach, V-MAN (Marzolla, Babaoglu&Panzieri 2011), a gossip-based algorithm for consolidating Virtual Machines using server

consolidation technique has been used to optimise cloud energy requirements . This technique has successful to put hosts in power-saving mode, resulting in a large reduction in the total energy consumption of the cloud infrastructure. V-MAN is fully decentralised and does not require any global knowledge but some limitations are also present in such decentralised systems. Each server can exchange messages with a limited number of peers. Also, they assumed that all VMs are identical, but in real world it could not be possible.

Beloglazov *et al*.(2012) define an architecture and major principles for energy-efficient Cloud computing. Authors present open research challenges, allocation algorithms and resource provisioning for energy-efficient resource management of Cloud computing environments. They proposed an approach for placement of virtual machine in such a manner that reduces energy requirements and number of migrations. Their approach outperformed the others but they did not considered SLA parameters while selecting virtual machines for migration.

Setzer & Bichler (2013) have suggested to minimise the number of servers dynamically, but at the same time giving adequate computing resources to each point in time. They used singular value decomposition (SVD) to obtain significant attributes from a large constraint matrix and give a new geometric interpretation of these attributes, which permits the allocation of huge collection of applications to physical servers or other hardware with this new formulation. Large volumes of data would typically render the server consolidation problem intractable for all but small instances.

Server consolidation makes it possible to put unused servers into the sleep mode, which can typically save more power than DVFS. However, server consolidation may incur a higher overhead. Due to variations in workload, the power optimizer may require the migration of a VM from one server to another, or require a server to awaken. These operations are time consuming, especially when the processors of the related servers are busy, or when the network bandwidth is limited. Thus, the optimiser should not be invoked too frequently (Wang & Wang, 2014).

Proposed Approach

Due to the dynamic nature of cloud computing, sometimes servers will be overloaded. The situation causes performance degradation due to the lack of resources at the guest operating systems. To handle the situation few of the VMs need to migrated from such overloaded hosts. In this section we purpose a new approach that selects the VM based upon the CPU utilisation and categories of VMs, for migration. The proposed approach prefers to migrate the VMs that belongs to the higher category so that that the resources may be allocated to these VMs to avoid costly penalties for SLA violations. Our approach reduces the cost by optimising the energy consumption and amount of penalties paid for SLA violations.

Categorisation of Virtual Machines

As the cloud computing is utility based computing, we proposed to categorize VM based on the VM usage cost. We divide the VMs into five categories: category 1 to 5 with following assumptions:

- The usage cost of category-1 VMs is lowest while the cost of category-5 VMs is highest i.e. higher the category, higher the cost.
- In case of SLA violation, cloud service provider have to pay higher penalty amount for the category 5 VMs as compared to other categories i.e. higher the category, higher the penalty.

Cloud users will select the VM category they want to use and mention in the Service Level Agreement.

Proposed Algorithm

1. Prepare the list OH of Hosts having CPU utilisation > Upper-Threshold
2. For each Host H in list OH do

   2.1 Until CPU utilisation of H > Upper-Threshold

   2.1.1. Find the list UCL of VMs that belong to the uppermost category among VMs running on H.

   2.1.2. Find the VM V from list UCL such that Absolute ((CPU utilisation of H-Upper-Threshold)-CPU utilisation by V) is minimum.

   2.1.3. Add V to list of machines to be migrated.

   2.1.4 Set CPU-Utilisation of H = CPU-Utilisation of H - CPU utilisation by V

   [End of step 2.1].

   [End of step 2].
3. Migrate the VMs selected in step 2.

   [End of the Algorithm]

Firstly, this algorithm prepares the list of overloaded hosts. Then for each overloaded host, the algorithm selects VMs to migrate by considering the category of VM. The algorithm stops when new utilisation of host is lower than the upper utilisation threshold. Flowchart given in Figure 1depicts the working of this algorithm.

Simulation Parameters

The performance of the proposed approach has been evaluated by simulating it in CloudSim-3.0.3 (Calheiros *et al*., 2011) with the parameters mentioned in Table1. In order to evaluate the performance of purposed approach other scheme (Random Selection Policy, Minimum Migration Time Policy and Minimization of Migration Policy) used for the same purpose have also been simulated and performance of the purposed approach has been compared with the performance of these schemes.

Figure 1: Flow chart Representing Proposed Algorithm

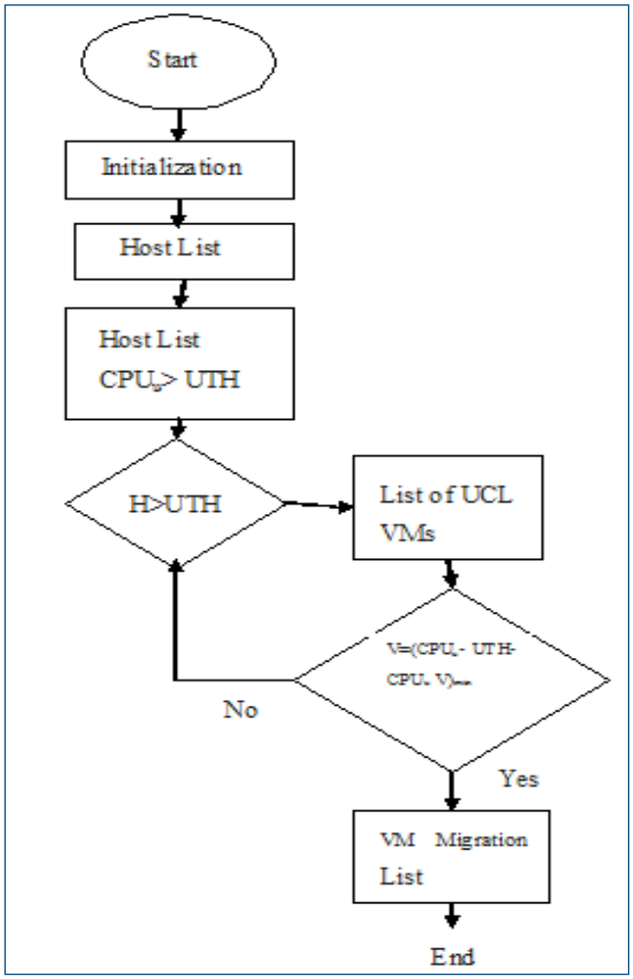


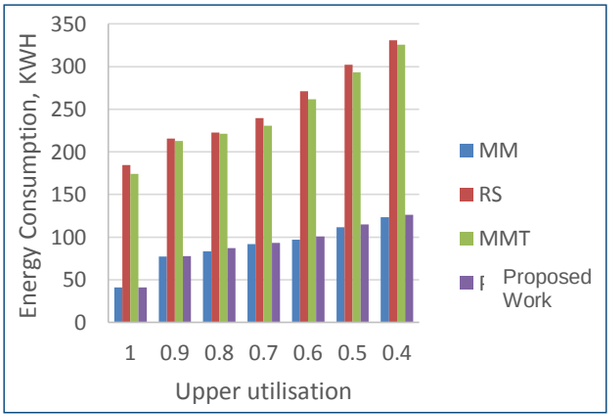
Table 1: Simulation Parameters

Parameters	*Value/s*
Physical Node	200
Host Type	2
MIPS	1860, 2660
Host Storage	1GB
Host RAM	4 GB
Host CPU Core	1
Number of VMs	300
VM Type(based on configuration)	4
MIPS(VM)	2500, 2000, 1000, 500
VM Size	2.5GB
RAM(VM)	870, 1740, 613MB
CPU Core (VM)	1
Simulation Length	10*60*60 seconds (10 hours)

Results & Discussions

To evaluate the performance three parameters have been observed: Energy Consumption, Number of total SLA Violations and VM category-wise number of SLA violations. The total energy consumed by datacenter while operated with various approaches has been represented in graph given in figure2.

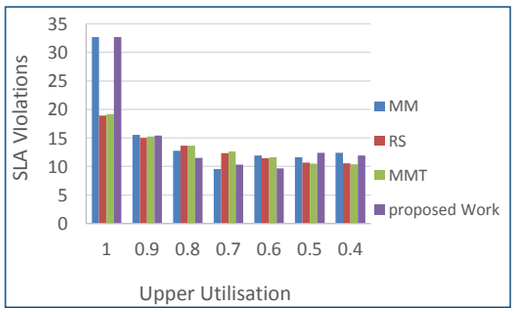
Figure 2: Energy Consumption by MM, MMT, RS and Proposed Approach



As represented by graph given in Figure 2, the proposed scheme consumes least energy among allthe other approaches for all values of upper utilisation except for Minimization of Migration approach (Beloglazov *et al*., 2012).

The observed number of SLA violations occurred has been mentioned in graph given in Figure 3.

Figure 3: SLA Violations Occurred with MM, MMT, RS and Proposed Approach



As shown in graph (Figure 3), in most of the cases, number of SLA violations occurred with Minimization of Migration (MM) technique is minimum. The number of SLA violations occurred with the proposed approach are marginally additional than MM technique but less than other approaches except in case of 0.6 upper utilisation threshold value. In this case the purposed approach has minimum number of SLA violations.

The proposed approach demands slightly more energy as well as lead to slightly more SLA violation than Minimization of Migration approach. However, the cost paid in lieu of SLA violations can be reduced by using the proposed approach as it reduces the SLA violations belonging to VM categories for which penalties are higher. The category wise number of SLA violations occurred with our approach has been mentioned in Table 2.

Table 2: VM Category-wise Average SLA Violation Occurred with the Proposed Approach

CPU Utiliz--ation	*Cat--egory 1*	*Cat--egory 2*	*Cat--egory 3*	*Cat--egory 4*	*Cat--egory 5*
1	32.27	30.06	29.77	25.31	23.93
0.9	20.04	19.27	18.34	13.88	11.98
0.8	16.68	15.79	16.55	9.85	7.44
0.7	13.01	13.23	11.39	8.65	7.30
0.6	12.76	12.54	12.86	6.64	6.50
0.5	18.93	18.26	15.17	10.24	8.98
0.4	17.64	19.78	16.93	5.93	7.03

Conclusion and Future Work

In this work a new approach has been purposed to select virtual machines from overloaded hosts of cloud data centre, for migration. The proposed approach prefers to migrate the VMs that belong to the higher category, so that the resources may be allocated to these VMs to avoid costly penalties for SLA violations. Our approach sets a trade-off between the cost paid for consumed energy and the cost paid in lieu of SLA violations. This approach has been simulated in Cloud Sim. During the simulation it has been observed that the proposed approach demands slightly more power (as 8.3% additional power in case of 70% upper utilisation) and lead to slightly more SLA violations than Minimization of Migration approach. Although performance of proposed approach is comparatively degraded than Minimization of Migration approach, yet the cost paid in lieu of SLA violations can be reduced by using the purposed approach as it reduces the SLA violations belonging to VM categories for which SLA penalty rates are high.

In this work, we have only considered SLA violations in terms of processing requirements (Million Instruction Per Second). In future new approaches can also be developed by considering SLA violations on other parameters such as RAM and UP-Time etc.

References

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

Bogatin, D. (2006). Google CEO's new paradigm:'cloud computing and advertising go hand-in-hand. ZD Net. [Online]. Retrieved from http://blogs.zdnet. com/micromarkets. (accessed January 12, 2014)

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., & Buyya, R. (2011). Cloud Sim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience*, 41(1), 23-50.

Dharwar, D., Bhat, S. S., Srinivasan, V., Sarma, D., & Banerjee, P. K. (2012). *Approaches towards Energy-Efficiency in the Cloud for Emerging Markets.* Paper presented at Cloud Computing in Emerging Markets

(CCEM), 2012 IEEE International Conference on (pp. 1-6).

Ferreto, T. C., Netto, M. A., Calheiros, R. N., & De Rose, C. A. (2011). Server consolidation with migration control for virtualized data centers. *Future Generation Computer Systems*, 27(8), 1027-1034.

Marzolla, M., Babaoglu, O., & Panzieri, F. (2011). *Server Consolidation in Clouds through Gossiping*. Paper presented at World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium (pp. 1-6).

Rizvandi, N. B., Taheri, J., & Zomaya, A. Y. (2011). Some observations on optimal frequency selection in DVFS-based energy consumption minimization. *Parallel and Distributed Computing*, 71(8), 1154-1164.

Setzer, T., & Bichler, M. (2013). Using matrix approximation for high-dimensional discrete optimization problems: Server consolidation based on cyclic time-series data. *European Journal of Operational Research*, 227(1), 62-75.

Venkatraman, A. (2012). Global census shows datacentre power demand grew 63% in 2012, Retrieved fromhttp://www.computerweekly.com(accessed January 12, 2014).

Von Laszewski, G., Wang, L., Younge, A. J., & He, X. (2009). *Power-aware Scheduling of Virtual Machines in DVFS-enabled Clusters*. Paper presented in Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on (pp. 1-10).

Wang, Y., & Wang, X. (2014). Performance-controlled server consolidation for virtualized data centers with multi-tier applications. *Sustainable Computing: Informatics and Systems*, 4(1), 52-65.

Wu, C. M., Chang, R. S., & Chan, H. Y. (2014). A green energy-efficient scheduling algorithm using the DVFS technique for cloud data-centers. *Future Generation Computer Systems*, 37, 141-147.