Heterogeneous Workload Consolidation for Efficient Management of Data Centers in Cloud Computing

Deep Mann
ME (Software Engineering)
Computer Science and Engineering Department
Thapar University
Patiala-147004

Inderveer Chana
Associate Professor
Computer Science and Engineering Department
Thapar University
Patiala-147004

ABSTRACT
Cloud computing is a recent innovation, which provides various services on a usage based payment model. The rapid expansion in data centers has triggered the dramatic increase in energy used, operational cost and its effect on the environment in terms of carbon footprints. To reduce power consumption, it is necessary to consolidate the hosting workloads. In this paper, we present a Single Threshold technique for efficient consolidation of heterogeneous workloads. Our technique focuses on the energy consumption of the data center due to the heterogeneity of the workloads and also gives information about the SLA violations. The experimental results demonstrate that our technique is efficient for the data centers to consolidate the heterogeneous workloads.

General Terms
Cloud Computing, Workload Consolidation Algorithms.

Keywords
Cloud Computing, Energy Efficiency and Heterogeneous Workload Consolidation.

1. INTRODUCTION
The concept of cloud computing was introduced in 1961 by John McCarthy. Later, the computing technologies such as Utility Computing [2] combined with established standards of Web 2.0 gave rise to cloud computing. Cloud Computing rather than offering a product provides services such as infrastructure, platform and software to end-users. These services are provided by sharing resources, software and other information under a usage based payment model. There are many proposed definitions of cloud computing due to its growing popularity defining its characteristics. A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers [1].

It has been projected that the data centers consume 0.5 percent of the world's total electricity usage [3]. If recent demand continues, it is projected to be augmented by 2020. In 2005, the total energy consumption (servers, cooling units) was projected at 1.2% of the total U.S. energy consumption, amplifying every 5 years [4]. Amazon's estimations [5] about its data center are shown in Figure 1. The major part of the energy used in society nowadays, is generated from fossil fuels which results in harmful CO_2 emissions. So there arises a need to reduce the power consumption of the data center. The power consumption of the data center can be reduced by consolidating the various parameters like resource utilization, performance and energy consumed.

Fig 1: Distribution of energy in data center [5]

The objective of this paper is to demonstrate an efficient technique to consolidate the heterogeneous workloads in Cloud Computing. In this technique, we have taken the heterogeneous applications having varied length parameter. The applications are placed on the virtual machines accordingly on the host having less utilization than upper utilization. On the basis of size, we have categorized the workloads into six categories and studied the variation of energy consumption by varying the heterogeneity of the workloads. Section 2 focuses related work on existing homogeneous and heterogeneous workload consolidation techniques. Section 3 discusses about the new heterogeneous workload consolidation technique in Cloud Computing and the framework used to implement the proposed technique and Section 4 concludes the paper.

2. RELATED WORK
Workload Consolidation Techniques can be categorized into two forms:

2.1 Homogeneous Workload Consolidation Techniques
These techniques include the similar kind of workloads which is having fixed number of parameters such as length, number of CPU's required and buffer size of input and output files, etc. Workload consolidation in data centers is achieved with the help of virtualization, which helps in saving the energy as unused nodes can be put to sleep mode or by shut down of the machines. Lefevre et al. [6] has designed the Green Open Cloud Architecture, which is an energy aware framework for cloud computing and shown the life of single virtual

machine during its different faces like boot of a VM, running of VM and shutting down of a VM and the power consumption during these phases. They have given the technique to reduce the energy consumption of virtualized data centers, which saves the 25% electric consumption of the infrastructure. They have validated the technique using different scheduling algorithms like round robin scheduling algorithm and unbalanced scheduling algorithm, on a multicore platform. The application type used in this technique has varied length parameter. This technique gives information about the energy consumption and VM migration but does not provide any information about the SLA violations.

Lee et al. [7] have presented two workload consolidation algorithms (ECTC and MaxUtil) on the basis of resource usage patterns like low, high and random. They have taken the arbitrary workload for the experimental evaluation and shown the energy saving results using VM migrations and without VM migrations and found that energy saving without using VM migrations results more efficient. This technique maximizes the resource utilization without the performance degradation of the workload.

2.2 Heterogeneous Workload Consolidation Techniques

These techniques include different kind of workload which is having varying parameters like length; number of CPU's required and buffer size of input and output files. Zhan et al. [8] have studied the consolidation of heterogeneous workloads for large organizations for which they have proposed the phoenix cloud architecture. The proposed approach uses the varying workloads and simulation results show the increased number of completed jobs. This approach does not give any information about the energy consumption of the data center and SLA.

Tian et al. [9] have proposed a Map Reduce technique which uses the triple queue scheduler for the consolidation of heterogeneous workload. They have categorized the workload into three types namely CPU bound jobs, input/output bound jobs and CPU bound without shuffle jobs on the basis of their CPU and I/O utilization. By this technique hadoop throughput can be increased by 30%. This improves the overall performance of the system but do not take into the consideration of the energy consumption of the system.

There are various other techniques [10], [11], [12] which manage the heterogeneous workloads but are not energy efficient for the cloud computing platform. Our aim is to consolidate the heterogeneous workloads in an efficient way so that resource utilization should be maximized and energy consumption of the data center should be minimized which will result in the less number of carbon footprints.

3. HETEROGENEOUS CONSOLIDATION TECHNIQUE

In comparison to the techniques mentioned above, we propose a technique which aims at the efficient consolidation of heterogeneous workloads by using the current utilization of the resources and places the workload accordingly. We have extended the work of [14] by modifying the single threshold technique according to the length of the workload.

3.1 Green Cloud Architecture

For this technique we are using the Green Cloud Architecture [2], the high level of architecture is shown in the Figure 2. The proposed technique works on the PaaS layer as allocation policies are implemented on PaaS layer which are followed by the IaaS layer.

3.2 VM Allocation Policy

The VM allocation Policy used is Single Threshold. In this policy, the upper utilization of the host is fixed and virtual machines are placed according to the current CPU utilization. For the execution of the applications, we have used the First Come First Serve scheduling algorithm. On arrival of the heterogeneous applications, the virtual machine is assigned to the application. Later, when the utilization of the host reaches up to the set threshold, virtual machines are migrated to the hosts whose threshold is below the fixed threshold.

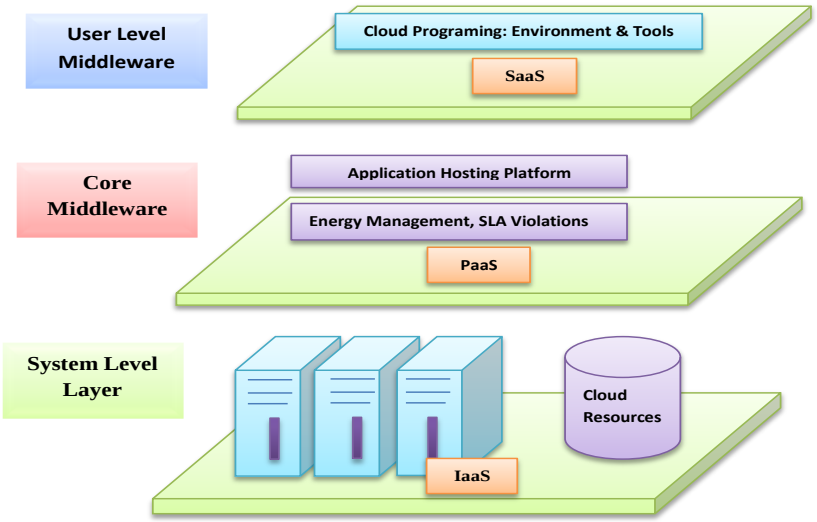

Fig 2: High Level Architecture

3.3 Performance Metrics

In order to get the optimal points where energy consumption in KWh of the data center is minimum we have chosen two metrics. First performance metric gives the total energy consumption of the data center by the resources due to the execution of heterogeneous workloads and the second performance metric gives the percentage SLA violations. The SLA violations occur when the virtual machine does not get the required amount of Million Instructions per Second (MIPS).

3.4 Experimental Setup and Discussions

Reference analysis has been done based on CloudSim Toolkit [13]. CloudSim provides an efficient way of performing the experiments with heterogeneous workloads and helps in calculating the energy consumption of the data center, VM migrations, SLA violations, etc.

We simulated data center with 100 physical nodes, each having one CPU core performance with 1000, 2000, or 3000 Million Instructions per Second, 1TB of storage and 8GB of RAM. All VM requires one CPU core with 250, 500, 750 or 1000 MIPS, 1 GB of storage and128 MB of RAM. Each VM executes a heterogeneous workload. Total number of VMs taken is 100.

Each application is submitted to the virtual machine for its execution as per the requirement. We have taken the SPEC power benchmark, stated in the fourth quarter of 2010, that the average power consumption at 100% utilization for servers consuming less than 1000 Watt was approximately 259 W[1] [14], to compare the energy efficiency at different threshold utilizations. The heterogeneity of the applications is based on the size of the applications. We have divided the applications into six cases as defined below.

Case 1: The heterogeneity between the applications is 1000 MIPS. The results of the simulation as shown in Figure 3, 4 indicate that the energy consumption as well as the SLA violations increases or decreases at different threshold utilizations. The optimal utilization of the host should be 90% as shown in the figures indicated. At this point energy consumption is 1.32 KWh and SLA violations are minimum i.e. 5.37%, giving the minimum energy consumption with respect to the SLA violations at this threshold.

Case 2: The heterogeneity between the applications is 2500 MIPS, as shown in Figure 5, 6. In this case optimal utilization of the host should be 80% as at this point energy consumption is minimum, 1.4 KWh and SLA violations are also minimum 5.56%.

Case 3: The heterogeneity between the applications is 5000 MIPS, as shown in Figure 7, 8. The energy consumption is increasing with the increase of heterogeneity of workloads. In this case optimal utilization should be 70% as at this point energy consumption is 1.64 KWh and SLA violations are 4.82%.

Case 4: The heterogeneity between the applications is 7000 MIPS, as shown in Figure 9, 10. In this case energy consumption decreases with the increase of utilization and the optimal utilization should be 80% as at this point both energy consumption and SLA violations are minimum, i.e. 1.91 KWh and 4.64% respectively.

Case 5: The heterogeneity between the applications is 20000 MIPS, as shown in Figure 11, 12. In this case optimal utilization should be 70% as at this point energy consumption is minimum, i.e. 3.03 KWh and respective SLA violations are 4.02%.

Case 6: The heterogeneity between the applications is 50000 MIPS, as shown in Figure 13, 14. In this case optimal utilization should be 70% as at this point energy consumption is minimum, i.e. 5.95 KWh as shown in graph and respective SLA violations are 3.44 %.

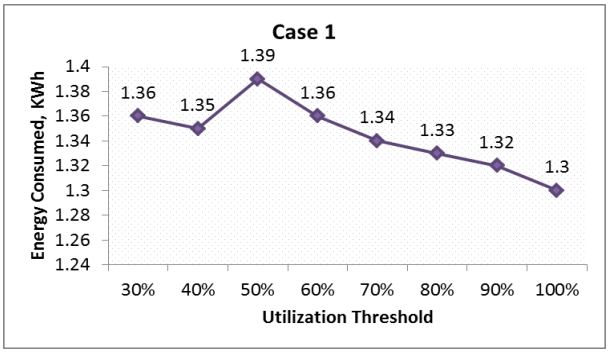

Fig 3: Energy Consumption in Case 1

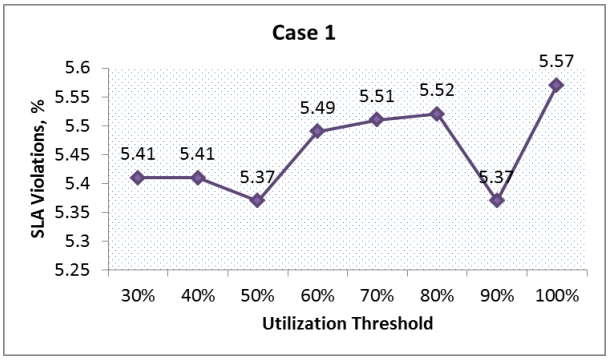

Fig 4: SLA violation in Case 1

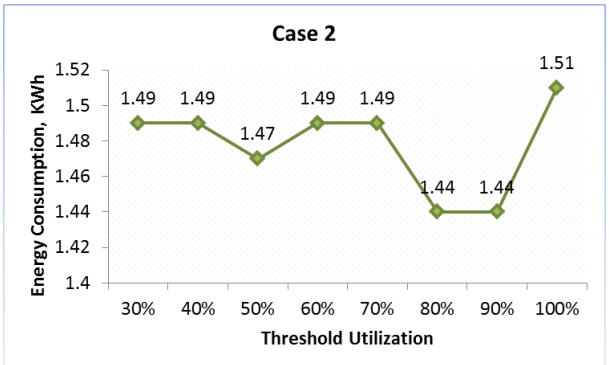

Fig 5: Energy Consumption in Case 2

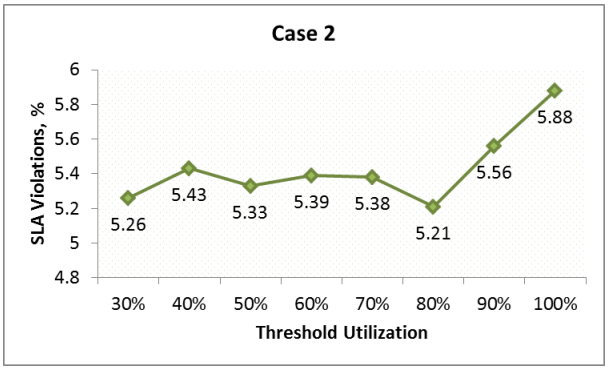

Fig 6: SLA violation in Case 2

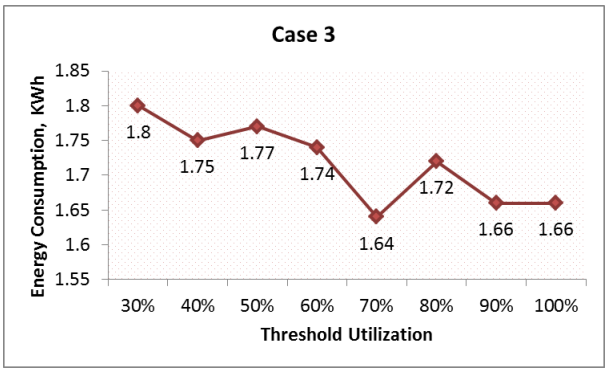

Fig 7: Energy Consumption in Case 3

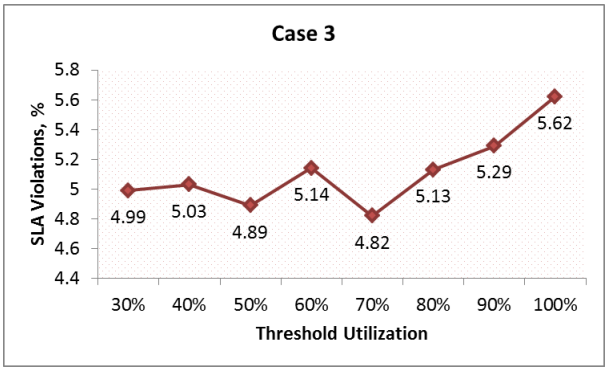

Fig 8: SLA violation in Case 3

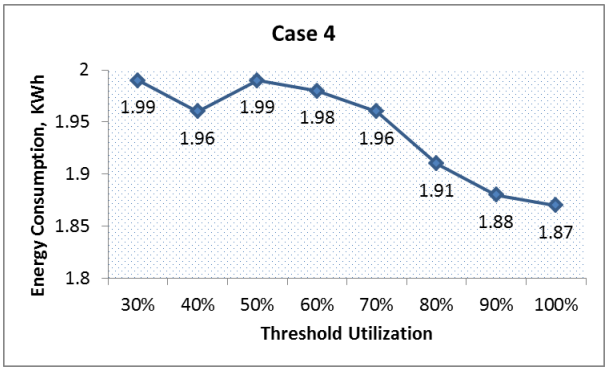

Fig 9: Energy Consumption in Case 4

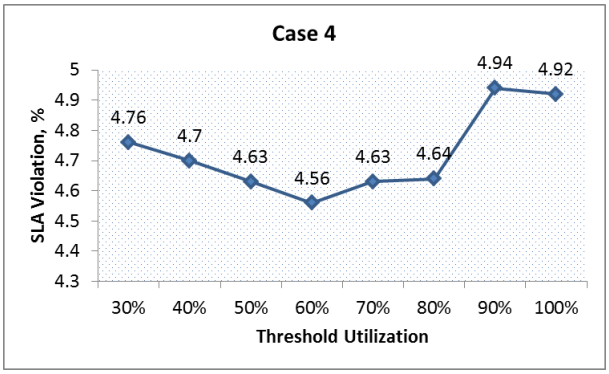

Fig 10: SLA violation in Case 4

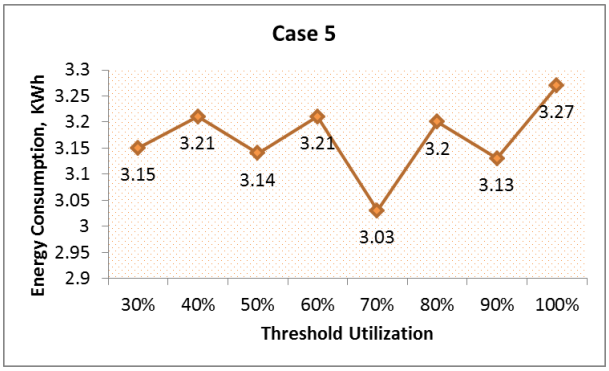

Fig 11: Energy Consumption in Case 5

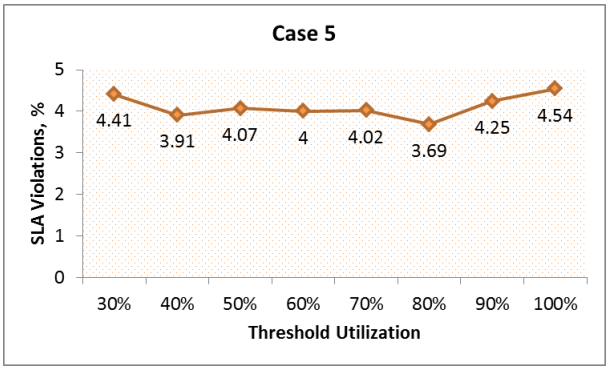

Fig 12: SLA violation in Case 5

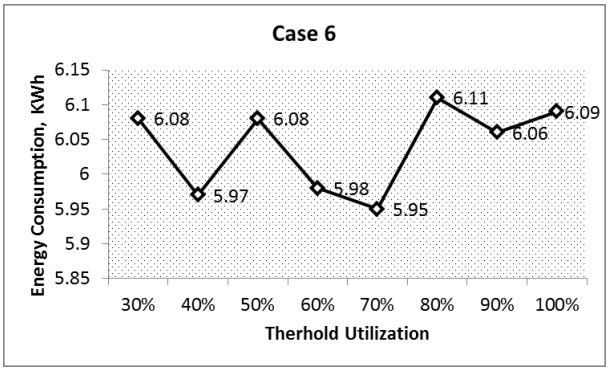

Fig 13: Energy Consumption in Case 6

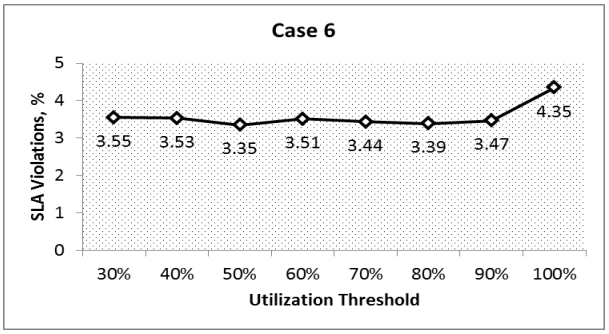

Fig 14: SLA violation in Case 6

It can be concluded from the results that energy consumption of the data center increases with the increasing heterogeneity of the workloads. The SLA violations too increase with the decrease in the energy consumption. The priorities assigned to the virtual machines help to save the energy consumption as the workload will be allocated to the virtual machine having high priority and less utilization. So we can set the threshold utilization of the node according to the variability of the workloads as shown in the table 1.

Table 1. Value of Threshold according to the heterogeneity between the applications

Heterogeneity between Applications (MIPS)	Threshold Utilization (%)	Energy Consumption (KWh)	SLA Violations (%)
1000	90	1.32	5.37
2500	80	1.4	5.56
5000	70	1.64	4.82
7000	80	1.91	4.64
20000	70	3.03	4.02
50000	70	5.95	3.44

4. CONCLUSION

Cloud computing technology has been widely adopted by the industry. The recent study of the data centers shows large amount of energy consumption and emission of CO_2. This paper presents a technique which consolidates the virtual machines dynamically on the basis of the length of the heterogeneous workloads, which results in the less energy consumption of the data center. We can set the threshold of the host according to the variability of the workload as shown in the experimental results.

5. REFERENCES

[1] Goiri, I., Fito, J., Julia, F., Nou, R., Berral, J. L., and Guitart, J. 2010. Multifaceted resource management for dealing with heterogeneous workloads in virtualized data centers. In 11th ACM/IEEE international conference on grid computing (Grid 2010). pp. 25–32. Brussels. Belgium.

[2] Buyya, R., Yeo, C., and Venugopal, S. 2008. Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-OB, IEEE CS Press. Los Alamitos, CA. USA).

[3] Foster, I., Zhao, Y., Raicu, I., and Lu, S. 2008. Cloud Computing and Grid Computing 360-Degree Compared in Grid Computing Environments Workshop. 200B. GCE'OB. pp. 1-10.

[4] Koomey, J. 2007. Estimating total power consumption by servers in the US and the world. Final report. vol. 15.

[5] Berl, A., Gelenbe, E., Di Girolamo M., Giuliani, G., Meer, H. D., Dang, M. Q., and Pentikousis, K. 2009. Energy-Efficient Cloud Computing. The Computer Journal. vol. 53. No. 7. pp. 1045-1051.

[6] Lefevre, L., and Orgerie, A. C. 2010. Designing and evaluating an energy efficient Cloud. The Journal of Supercomputing. Springer. vol. 51. No. 3. pp. 352-373.

[7] Lee, Y. C., and Zomaya, A. Y. 2010.Energy efficient utilization of resources in cloud computing systems. Journal of Supercomputing. pp. 1-13.

[8] Zhan, J., Wang, L., Tu, B., Li, Y., Wang, P., Zhou W., and Meng, D. 2008. Phoenix Cloud: Consolidating Different Computing Loads on Shared Cluster System for Large Organization. Proceeding Workshop Cloud Computing and Its Application (CCA '08).

[9] Tian, C., Zhou, H., He, Y., and Zha, L. 2009. A dynamic MapReduce scheduler for heterogeneous workloads. In Proceedings of the 2009 Eighth International Conference on Grid and Cooperative Computing. pp. 218-224. IEEE Computer Society.

[10] Nou, R., Julià, F., Guitart, J., and Torres. J. 2007. Dynamic resource provisioning for self-adaptive heterogeneous workloads in smp hosting platforms. International Conference on E-business (2nd) ICE-B. Barcelona. Spain.

[11] Steinder, M., Whalley, I., Carrera, D., Gaweda, I., and Chess, D. M. 2007. Server virtualization in autonomic management of heterogeneous workloads. In Integrated Network Management. pp. 139-148.

[12] Carrera, D., Steinder, M., Whalley, I., Torres J., and Ayguade, E. 2008. Managing SLAs of heterogeneous workloads using dynamic application placement. HPDC '08 Proceedings of the 17th international symposium on High performance distributed computing. New York. USA.

[13] Calheiros, R. N., Ranjan, R., Beloglazov, A., Rose, C. A. F. D., and Buyya, R. 2010. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience. Wiley Press. New York. USA.

[14] Beloglazov, A., Abawajy J., and Buyya, R. 2011. Energy-Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing. Future Generation Computer Systems. ISSN: 0167-739X. Elsevier Science. Amsterdam. The Netherlands. (In press, accepted on April 28).