International Journal of Computational Intelligence and Informatics, Vol. 3: No. 2, July - September 2013

An Enhanced Technique on Workload Management for Increasing the Return on Investment in Cloud

T Ambika
Computer Science and Engineering
SNS College of Technology
Coimbatore, Tamil Nadu, India
ambika.av@gmail.com

S Karthik
Computer science and Engineering
SNS College of Technology
Coimbatore, Tamil Nadu, India
profskarthik@gmail.com

***Abstract*-With the advent of cloud computing remote and wide spread cloud users are provided various services on demand. Workloads with cloud providers comprises of both transactional applications and long-running batch jobs. With an existing methodology the heterogenic nature of workloads is dealt with a technique of consolidating the varied workload types within the same machine. The infrastructure cost and power consumption will be reduced when compared with the techniques where a separate machine is allocated for each workload type. For each workload a scheduling mechanism chosen and performance goals of each workload are to be satisfied. Traditionally, for the purpose of resource provisioning with virtualization the Virtual Machine (VM) size is estimated for individual machines i.e., one VM at a time. Today the need for energy and resource consumption is drastically increasing for a better return on investment for cloud providers. In this paper, we present the concept of resource provisioning with consolidating multiple VMs could be combined along with the workload consolidation. This paper reduces the underutilization of resources as the unused portion of every virtual machine is utilized. The proposed consolidating technique is more effective in energy consumption and resource utilization when compared with other techniques as consolidation is done at two stages. The heterogeneous workload consolidation depending on the performance goal of each workload type is done at one stage for collocating them at the same machine and the consolidation is also done at the resource provisioning time where the unused portion of the VMs is done at the other stage.**

Keywords-Workload, resource provisioning, virtualization, cloud computing

I. INTRODUCTION

In Cloud, dynamic resource provisioning plays a vital role with fluctuations in the arrival of applications [11]. Workloads which are the abstraction of the actual work to be performed which are mapped to a set of resources, may it be a computing storage or network resources which may be offered by the cloud. Workloads are of different types with varying performance goals depending on the nature of the workload. Transactional workloads are to be executed in real time where as long running batch jobs can be executed at any time in batch mode. Management of workloads requires different parameters to be considered. Transactional workloads are interactive in nature they are managed using flow control, load balancing etc., while the non-interactive workloads require to be managed with appropriate scheduling mechanisms and the resources required.

A. Heterogeneous workloads

The different workloads are consolidated with the single machine to have a cost effective infrastructure. This has certain other challenges to be managed with. The performance goals vary from one workload type to the other. On considering the nature of transactional workloads [8] they are to be carried out within a short time period and hence their performance goal deals with response time within that period. Whereas with the long running batch jobs the performance goals is based on the finishing time of individual jobs are to be considered and scheduling is to be carried out accordingly. On managing with the heterogeneous nature of the workloads the Service Level Agreements (SLAs) can be met and satisfied only when the performance goals of each workload type are concentrated separately.

Hence to deal with such complexities on consolidating the workloads the comparative performance functions are used for both workload types long running jobs and transactional workloads as in [4]. The actual performance of the application is compared with the goal of that application. The aim of the comparative performance function is to meet the SLAs with the minimum possible violation and there by providing a guaranteed Quality of Service (QoS) as along with the utility computing approach with the cloud providers [9] the cloud customers must also be made satisfied for the compute clouds to run under a win-win situation. For the autonomic placement of the

workloads the placement controller technique is used as in [4]. The placement controller provides a decision making job for an effective placement of the heterogeneous workloads.

B. Virtualization

Virtualization techniques adopted by the cloud providers provide them with more economical way of handling the available resources where in multiple VMs are created over a single virtual machine. Power consumption is comparatively low as with this approach the idle nodes could be switched off. The VM creation itself has its own overhead to be created and initialized to execute the tasks [7]. The next concern considered for resource provisioning with virtualization is that the capacity of the VM size is to be calculated. This plays a vital role as the concept of effective resource provisioning is to be maintained.

Both over provisioning and under provisioning will not result in effective utilization of resources and energy as expected. Traditionally the VM size estimation is done for each VM individually. The concept of VM consolidation for effective resource provisioning is studied [12]. With an economic pressure to reduce the energy consumption to a greater extent virtualization is to be improved in a more diversified manner. With this context, in this paper we advocate of the existing idea of consolidating the heterogeneous workloads to the same physical machine along with the idea of consolidating the VMs such that the VM capacity can be fully utilized.

This proposed method has a twofold effect on increasing the return on investment with the cloud providers. Because this method aims to consolidate the VMs at one end which increases the resource utilization factor and decreases the energy consumption where by the unused physical machines are switched off and the related VMs are combined to the other machines.

The rest of the paper is organized as follows: Section 2 deals with the study of workload management and resource provisioning techniques. Section 3 discusses about the management and collocation of heterogeneous workloads and Section 4 proposes consolidation of VM along with workload consolidation. Section 5 gives the conclusions of the proposed work.

II. RELATED STUDY

The prominent challenges with the cloud providers are to manage the workloads with heterogeneous nature and make effective resource provisioning out of it. This must be done in order to meet the SLA requirements and to provide guaranteed QoS to its customers. Study has been made with the management of homogeneous workloads where the statistical analysis of workloads on data-intensive clusters which consists of the workloads of parallel single CPU tasks are studied in [6]. The new patterns of job arrivals are discussed, which includes Pseudo-periodicity, long range dependence and grid based bag-of-tasks. They are much necessary for making workload modeling and performance predictions. But dealing with different workload types dynamically is essential to deal with compute clouds. The performance study of network I/O workloads is discussed in [13]. The different workloads which demand for the CPU or I/O resources are studied here. Further analysis is made on various factors that affect the performance throughput and resource sharing process.

The performance on co-location of applications requiring CPU and network resources i.e., the heterogeneous applications are measured along with performance measurement of co-locating homogeneous applications. Comparison was also made with different allocation and scheduling strategies. A detailed study in [8] is done for consolidation of heterogeneous workloads in the same machine. The performance goals for different workloads vary, say for long running batch jobs it is the completion time of each job and in case of transactional workload the goals are defined in terms of response time and throughput. For the SLA based resource provisioning with heterogeneous workloads the authors in [10], have proposed the penalty model for transactional and non-interactive batch jobs. Scheduling and rescheduling scenarios for VMs are discussed for enforcement of SLAs.

Autonomic workload execution based on the utility is studied in [9], where the workloads considered are the queries and work flows. Workload evaluation strategies with utility functions are and for the autonomic query execution, the admission controller, a query scheduler and execution controller are used. Utility based optimization is adapted in [3] also where the concentration was jointly on capacity allocation, load balancing and energy saving policies while meeting availability constraints.

A comparison with static and dynamic allocation is made where the overhead of dynamic allocation scheme is compared with static allocation in both system capacity and application performance in with different virtualization technologies. Comparison was done with different combinations of workloads and virtualization techniques but with homogeneous environment. In [14] a solution for dynamic multi-tier applications with varying resource requirements is proposed tested by the authors. Their architecture comprises of the application and the node controller to manage the application components and the VMs. The traditional queuing model is extended to suit the purpose.

VM consolidation for reduced energy consumption and cost is dealt in [1]. Anton Beloglazov and Rajkumar Buyya have given the cost of live migration. The author makes VM selection with fixed and dynamic thresholds and VM allocation is made with an effective algorithm. Joint VM provisioning in [12] provides the joint VM size estimation technique. For that purpose Xiaoqiao meng et al. forecasts the workload arrival for a period of time and the errors on that accord are also considered. They do the VM selection for joint sizing by considering VM pairs with negative correlations to suit the changing behavior of VM demand.

III. MANAGING HETEROGENEOUS WORKLOADS

The performance functions are used to measure the actual performance of the application and is compared with the goal of that application as in [4]. For resource allocation based on the applications the comparative performance functions are used. The performance representation for the transactional workloads as stated in [4,5] is based on the response time goal with every application. For a particular placement and the load distribution the CPU power requirement can be obtained to arrive at the comparative performance.

For the performance representation of the long running batch jobs as in [4] each job is associated with the resource utilization logs and the performance goals of the jobs. Along with it the job status and the comparative goal of the job is also assessed. The comparative performance functions are to be calculated for the placement controller for making effective placement decisions for the batch jobs. The comparative performance functions for the batch jobs vary with that of the transactional workloads because it is obvious that the performance goals for the workloads vary. Hence here another version of the comparative performance function is used as per the batch jobs requirements and the placement decisions are made accordingly. Figure 1 gives the result on using the comparative or relative performance function.

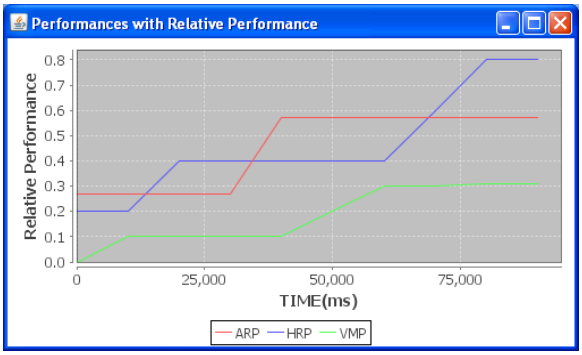

Figure 1. Relative Performance

The scheduling techniques chosen for the jobs must leverage the execution and completion of the jobs with effective CPU utilization. The scheduling algorithms such as shortest job first, priority based scheduling, round robin etc., are available. The most prominent scheduling techniques such as first come first serve algorithm, earliest deadline first are considered in [4] for comparison with the varying job profiles and SLA goals.

IV. CONSOLIDATION OF VMS AND WORKLOADS

Virtualization being part of the trend in utility computing [5] provides the basis for Infrastructure as a Service (IaaS). Resource provisioning is done with effective resource utilization and energy consumption in [12]. In [12], Xiaoqiao Meng et al. proposes the technique of VM consolidation as the joint-VM provisioning approach. The technique for estimating the capacity for provisioning with the VMs is done. A probable workload is forecasted to serve the purpose. The VM capacity is determined to meet the SLA requirements. The adopted VM selection method here uses the negative correlations to facilitate the varying behavior of VM demand as shown in figure 2.

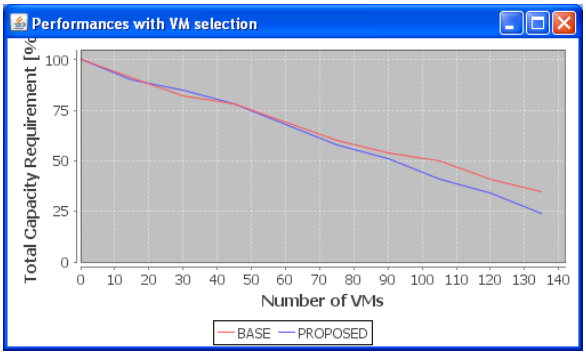

Figure 2. VM Selection

Alternatively, Anton Beloglazov and Rajkumar Buyya in the paper [1] proposes an energy efficient adaptive threshold- based technique for dynamic consolidation of VMs where in the SLA violation is also kept at the lowest possible percentage. VM selection is done using the fixed utilization thresholds [2] in their previous work and upon which the dynamic utilization threshold is proposed. Reallocation algorithm here uses the minimum migration technique [1] for dynamic thresholds.

Thus with the discussed VM consolidation techniques the workload consolidation technique discussed earlier in Section 3 could be done such that the effective workload management and the related resource provisioning will result in an increased return on investment for the cloud providers.

V. CONCLUSION

Compute clouds has its effective functioning depending on their workload management and resource provisioning policies. This paper proposes the idea of performing the consolidation at two stages. The existing workload consolidation technique dealing with the heterogeneous workloads are carried out along with the resource provisioning technique of VM consolidation. Combined adaptation of these techniques will facilitate the cloud providers to have profitable conditions at a larger extent. Return on investment with the cloud providers is increased as energy efficiency and resource utilization is achieved.

REFERENCES

[1] Anton Beloglazov and Rajkumar Buyya, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers", MGC 2010, 2010.

[2] A. Beloglazov and R. Buyya, " Energy efficient allocation of virtual machines in cloud data centers", In Proc. of the 10th IEEE/ACM Intl. Symp. on Cluster, Cloud and Grid Computing (CCGrid 2010), 2010.

[3] Bernardetta addis, Danilo Ardagna, et.al., "Autonomic management of cloud service centres with availability guarantees", IEEE 3rd Int. Conf. on cloud computing, 2010.

[4] David Carrera, Malgorzata Steinder, Ian Whalley, Jordi Torres, Eduard Ayguade, "Autonomic Placement of Mixed Batch and Transactional Workloads", IEEE Transactions on Parallel and Distributed Systems, Vol. 23, No. 2, February 2012.

[5] G Pacifici, M Spreitzer, A Tantawi, and A Youssef, "Performance Management for Cluster-Based Web Services", IEEE J. Selected Areas in Comm., Vol. 23, No. 12, pp. 2333-2343, Dec 2005.

[6] Hui Li, Lex Wolters, "Towards a Better Understanding of Workload Dynamics on Data-Intensive Clusters and Grids", IEEE, 2007.

[7] Inigo Goiri, Josep Ll.Berral, et.al., "Energy-efficient and Multi faceted resource management for profit-driven virtualized data centers", Elsevier, Future Generation Computer Systems 28(2012), pp. 718-731, 2012.

[8] Malgorzata Steinder, Ian Whalley, David Carrera, Ilona Gaweda, David Chess, "Server virtualization in autonomic management of heterogeneous workloads", Proc. IEEE/IFIP 10th symp Integrated Management (IM '07), 2007.

[9] Norman W. Paton, Marcelo A. T. de Arag˜ao, Kevin Lee, Alvaro A. A. Fernandes, Rizos Sakellariou, "Optimizing Utility in Cloud Computing through Autonomic Workload Execution", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2009.

[10] Saurabh Kumar Garg, Srinivasa K.Gopalaiyengar and Rajkumar Buyya, "SLA-Based Resource Provisioning for Heterogeneous Workloads in a Virtualized Cloud Datacenter", Springer-Verlag Berlin Heidelberg ICA3PP 2011, part I,LNCS 7016, pp.371-384, 2011.

[11] Qi Zhang, Lu Cheng, Raouf Boutaba, "Cloud computing: state-of-the-art and research challenges", J Internet Serv Appl (2010) 1: 7–18.

[12] Xiaoqiao Meng, Canturk Isci, Jeffrey Kephart, Li Zhang, Eric Bouillet, Dimitrios Pendarakis, "Efficient Resource Provisioning in Compute Clouds via VM Multiplexing", ICAC 2010, ACM, pp. 7-11. June 2010.

[13] Yiduo Mei, Ling Liu, Xing Pu, Sankaran Sivathanu, and Xiaoshe Dong, "Performance Analysis of Network I/O Workloads in Virtualized Data Centers", IEEE Transactions on Service Computing, 2011.

[14] Zhikui Wang, Yuan Chen, Daniel Gmach, Sharad Singhal, Brian J. Watson, Wilson Rivera, Xiaoyun Zhu and Chris D. Hyser, "AppRAISE: Application-Level Performance Management in Virtualized Server Environments", IEEE Transactions on Network and Service Management, Vol. 6, No. 4, December 2009.