SLA-based Virtual Machine Management for Mixed Workloads of Interactive Jobs in a Cloud Datacenter

Vivek H. Bharad
PG Scholar
Computer Engineering Department
School of Engineering
RK University
Rajkot, Gujarat, India

Hitesh A. Bheda
Assistant Professor,
Computer Engineering Department
School of Engineering
RK University
Rajkot, Gujarat, India

ABSTRACT
Efficient provisioning of resources is a challenging problem in cloud computing environments due to its dynamic nature and the need for supporting heterogeneous applications. Even though VM (Virtual Machine) technology allows several workloads to run concurrently and to use a shared infrastructure, still it does not guarantee application performance. Thus, currently cloud datacenter providers either do not offer any performance guarantee or prefer static VM allocation over dynamic, which leads to inefficient utilization of resources. Also, the workload may have different QoS (Quality Of Service) requirements due to the execution of various types of applications such as HPC and web, which makes resource provisioning much harder. Earlier work either concentrate on single type of SLAs (Service Level Agreements) or resource usage patterns of applications, such as web applications, leading to inefficient utilization of datacenter resources. In this paper, we tackle the resource allocation problem within a datacenter that runs different types of application workloads, particularly non-interactive and transactional applications. We propose an admission control and scheduling mechanism which not only maximizes the resource utilization and profit, but also ensures that the QoS requirements of users are met as specified in SLAs. In our study, we find that it is important to take care of various types of SLAs along with applicable penalties and the mix of workloads for better resource allocation and utilization of datacenters. The proposed mechanism provides substantial improvement over static server consolidation and reduces SLA violations.

Index Terms
SLA, VM, HPC (High Performance Computing)

1. INTRODUCTION
Cloud computing has led to a paradigm shift where enterprises, rather than maintaining their own infrastructure, started to outsource their IT and computational needs to third party service providers [2]. The clouds are large scale outsourcing datacenters that host thousands of servers which can run multiple virtual machines (VMs) at a same time. Therefore, they host a huge amount of applications and provide users with an abstraction of unlimited computing resources on a pay-as-you- go basis.

While there are several advantages of these virtualized infrastructures such as on-demand scalability of resources, there are still issues which prevent their widespread adoption [3]. In particular, for a commercial success of this computing paradigm, cloud datacenters need to provide better and strict Quality of Service (QoS) guarantees. These guarantees, which are documented in the form of Service Level Agreements (SLAs), are crucial as they give confidence to customers in outsourcing their applications to clouds [4]. However, current cloud providers give only limited performance or QoS guarantees. For instance, Amazon EC2 offers only guarantees on availability of resources, not on performance of VMs [5] [6].

Resource provisioning plays a key role in ensuring that cloud providers adequately accomplish their obligations to customers while maximizing the utilization of the underlying infrastructure. An efficient resource management scheme would require dynamically allocating each service request the minimal resources that are needed for acceptable fulfillment of SLAs, leaving the surplus resources free to deploy more virtual machines. The provisioning choices must adapt to changes in load as they occur, and respond gracefully to unanticipated demand surges. For these reasons, partitioning the datacenter resources among the various hosted applications is a challenging task. Furthermore, current cloud datacenters host a wider range of applications with different SLA requirements [6] [7] [8]. For instance, transactional applications require response time and throughput guarantees, while non-interactive batch jobs1 are concerned with performance (e.g., completion time). Resource demand of transactional applications such as web applications tend to be highly unpredictable and bursty in nature [9] , while demand of batch jobs can be predicted to a higher degree [10]. Hence, the satisfaction of complex and different requirements of competing applications make the goal of a cloud provider to maximize utilization while meeting different types of SLAs far from trivial.

Traditionally, to meet SLA requirements, over-provisioning of resources to meet worst case demand (i.e., peak) is used. However, servers operate most of the time at very low utilization level which leads to waste resources at non-peak times [11]. This over-provisioning of resources results in extra maintenance costs including server cooling and administration [12]. Some companies such as Amazon (Schneider,) are trying to utilize such slack of resources in the form of spot "instances" by renting them out at much lower rate but with low performance guarantees. Similarly, many researchers tried to address these issues by dynamic provisioning of resources using virtualization, but they focused mainly on scheduling based on one specific type of SLA or application type such as transactional workload. Although computationally intensive applications are increasingly becoming part of enterprise datacenters and cloud workloads, still research considering such applications is in infancy. Today, most of the datacenters run different types of applications on separate VMs without any awareness

of their different SLA requirements such as deadline, which may result in resource underutilization and management complexity.

To overcome these limitations, we present a novel dynamic resource management action that not only maximizes resource usage by sharing resources among multiple concurrent applications owned by different users, but also considers SLAs of different types. We handle scheduling of two types of applications, namely, compute intensive non-interactive jobs and transactional applications such as Web server, each having different types of SLA requirements and specifications. Our strategy makes dynamic placement decisions to respond to changes in transactional work- load, and also considers SLA penalties for making future decisions. To schedule batch jobs, our proposed resource provisioning mechanism predicts the future resource availability and schedules jobs by stealing CPU cycles, which are under-utilized by transac- tional applications during off-peak times.

2. LITERATURE REVIEW

There are several works that relate to our research focus particularly in the area of constant change or dynamic resource provisioning and allowing mixed/heterogeneous workloads within a cloud datacenter. We broadly classify the works with respect to dynamic resource provisioning such as scheduling mixed workloads, SLAs, and auto-scaling of applications. The comparison of the proposed work with most important existing ones, with respect to various parameters, is summarized in Table 1. The details of the related works are discussed below.

Joint-VM provisioning approach by exploiting statistical multiplexing among the workload patterns of multiple VMs, so that the unused resources of a low-utilized VM is borrowed by other co-located VMs with high utilization [14]. It also quickly reutilize the resources for a virtualized utility computing platform using "ghost" virtual machines (VMs), which take a part in application clusters, but do not handle client requests until required. These works concentrate on fixed number of VMs, while we consider variable amount of incoming workload [14].

It is require to analysis and resource provisioning for workloads management with considerable network and disk I/O requirements [16]. The management workloads scale with an increase in compute power in the datacenter. The workload for Internet applications is non-stationary; consider the workload mix received by a Web application for their mix-aware dynamic provisioning technique. Our paper also considers non-interactive applications define a unique resource-level metric (i.e., SLA) for specifying finer level guarantees on CPU performance [6]. This metric allows resource providers to dynamically allocate their resources among the running services depending on their demand. In contrast to the proposed work, they do not handle multiple types of SLAs and SLA penalty-related issues.

To take advantage of virtualization features, reallocate mix workloads on one server machine, thus reducing the granularity of resource allocation [9]. Preliminary working model of a framework for facilitating resource management in service providers, which allows both cost reducing and achieve the QoS based on SLAs. In contrast, our work concentrates on handling multiple types of SLAs both for High Performance Computing (HPC) and Web based workloads with a new admission control policy. A decentralized and robust online clustering approach for a dynamic mix of heterogeneous applications on clouds, such as long running computationally intensive jobs, bursty and response-time sensitive requests, and data and IO-intensive analytics tasks. When compared to our approach, the SLA penalties are not considered [7]. A lease management architecture called Haizea, that implements leases as VMs, leveraging their ability to suspend, migrate, and resume computations and to provide leased resources with customized application environments. Again, this paper does not consider the issues of SLAs and QoS [8].

The overhead of a dynamic allocation scheme in both system capacity and application-level performance relative to static allotment. It also provided implications and guidelines for a proper feedback controller design in dynamic allocation systems. In our work, the idea of dynamic allocation is extended for multiple types of workloads including HPC and Web [17]. In contrast, we propose architecture for specifying and monitoring SLAs to achieve the above. It also consider SLA-aware virtual resource management for cloud infrastructures, where an automatic resource manager controls the virtual environment which decouples the provisioning of resources from the dynamic placement of virtual machines. Even though the paper fulfills the SLA and operating costs, it does not deal with SLA penalty related issues. Many researchers developed a technique that enables existing middleware to fairly manage mixed workloads both in terms of batch jobs and transactional applications. The aim of this paper is towards a fairness goal while also trying to maximize individual workload performance. But our aim is to efficiently utilize the datacenter resources while meeting the different types of SLA requirements of the applications [8].

3. OPEN ISSUES

The aim of cloud service providers is to maximize the utilization of their datacenters by efficiently executing the user applications using minimal physical machines. It is also a problem to distinguish application that is either HPC or batch jobs and as per the demand of the job cloud data center need to give resource for the computation and require to maintain the SLA between user cloud providers. So it is required to manage resources such a way that unutilized resource on cloud data centers is utilize. Also another issue of executing batch jobs is parallelism of their jobs. So it is also hard to maintain or achieve parallel computation on cloud data center.

4. CONCLUSION

In this paper, we discussed how current cloud datacenters are facing the problem of underutilization and incurring extra cost. They are being used to run different types of applications from Web to HPC, which have different QoS requirements. This makes the problem harder, since it is not easy to predict how much capacity of a server should be allocated to each VM. Therefore, in this paper, we proposed a novel technique that maximizes the utilization of datacenter and allows the execution of heterogeneous application workloads, particularly, transactional and non- interactive jobs, with different SLA requirements. For designing more effective dynamic resource provisioning mechanisms, it is a must to consider different types of SLAs along with their penalties and the mix of workloads for better resource provisioning and utilization of datacenters; otherwise, it will not only incur unnecessary penalty to cloud providers but can also lead to under utilization of resources.

5. REFERENCES

[1] Antonescu A-F, Robinson P, Braun T. Dynamic sla management with forecasting using multi-objective optimization. In: Proceeding of 2013 IFIP/IEEE international symposium on integrated network management (IM 2013). Ghent, Belgium; 2013.

[2] Buyya R, Yeo C, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: vision, hype and reality for delivering computing as the 5th utility. Future Generat Comput Syst 2009;25(6):599–616.

[3] Ostermann S, Iosup A, Yigitbasi N, Prodan R, Fahringer T, Epema D. An early performance analysis of cloud computing services for scientific computing. Delft University of Technology, PDS-2008-006.

[4] Yeo C, Buyya R. Service level agreement based allocation of cluster resources: handling penalty to enhance utility. In: Proceedings of the 7th IEEE international conference on cluster computing. Boston, USA; 2005.

[5] Nathuji R, Kansal A, Ghaffarkhah A. Q-clouds: managing performance interference effects for qos-aware clouds. In: Proceedings of the 5th European conference on Computer systems (EuroSys 2010). Paris, France; 2010.

[6] Goiri I, Julià F, Fitó JO, Macías M, Guitart J. Resource-level QOS metric for CPU-based guarantees in cloud providers. In: Proceedings of 7th international workshop on economics of grids, clouds, systems, and services. Naples, Italy; 2010.

[7] Quiroz A, Kim H, Parashar M, Gnanasambandam N, Sharma N. Towards autonomic workload provisioning for enterprise grids and clouds. In: Proceedings of 10th IEEE/ACM international conference on grid computing. Melbourne, Australia; 2009.

[8] Sotomayor B, Keahey K, Foster IT. Combining batch execution and leasing using virtual machines. In: Proceedings of the 17th international ACM symposium on high-performance parallel and distributed computing. Boston, USA; 2008.

[9] Carrera D, Steinder M, Whalley I, Torres J, Ayguadé E. Enabling resource sharing between transactional and batch workloads using dynamic application place- ment. In: Proceedings of the ACM/IFIP/USENIX 9th international middleware conference, Leuven, Belgium; 2008.

[10] Smith M, Schmidt M, Fallenbeck N, Doernemann T, Schridde C, Freisleben B. Secure on-demand grid computing. Future Gener Comput Syst 2009;25(3):315–25.

[11] Barroso L, Holzle U. The case for energy-proportional computing. Computer 2007;40(12):33–7.

[12] Kim J-K, Siegel HJ, Maciejewski AA, Eigenmann R. Dynamic resource management in energy constrained heterogeneous computing systems using voltage scaling. IEEE Trans Parallel Distrib Syst 2008;19(11):1445–57.

[13] Kim J, Ruggiero M, Atienza D, Lederberger M. Correlation-aware virtual machine allocation for energy-efficient datacenters. In: Proceedings of the conference on design, automation and test in Europe. Ghent, Belgium; 2013.

[14] Meng X, Isci C, Kephart J, Zhang L, Bouillet E, Pendarakis D. Efficient resource provisioning in compute clouds via VM multiplexing. In: Proceedings of the 7th international conference on autonomic computing, Washington, USA; 2010.

[15] Zhang W, Qian H, Wills C, Rabinovich M. Agile resource management in a virtualized data center. In: Proceedings of Ist joint WOSP/SIPEW international conference on performance engineering. California, USA; 2010.

[16] Soundararajan V, Anderson J. The impact of MNGT. Operations on the virtualized datacenter. In: Proceedings of the 37th annual international symposium on computer architecture. France; 2010.

[17] Wang Z, Zhu X, Padala P, Singhal S. Capacity and performance overhead in dynamic resource allocation to virtual containers. In: Proceedings of the 10th IFIP/IEEE international symposium on integrated network management. Munich, Germany; 2007.

[18] Minarolli D, Freisleben B. Distributed resource allocation to virtual machines via artificial neural networks. In: Proceedings of 22nd Euromicro international conference on parallel, distributed and network-based processing (PDP), Turin, Italy; 2014.

[19] Casalicchio E, Menascé DA, Aldhalaan A. Autonomic resource provisioning in cloud systems with availability goals. In: Proceedings of the 2013 ACM cloud and autonomic computing conference, Miami, FL, USA; 2013.

[20] Hu Y, Wong J, Iszlai G, Litoiu M. Resource provisioning for cloud computing. In: CASCON '09: Proceedings of the 2009 conference of the Center for Advanced Studies on Collaborative Research, Ontario, Canada; 2009.