International Journal of Grid and High Performance Computing
Volume 12 • Issue 1 • January-March 2020

Improving Service Performance in Oversubscribed IaaS Cloud

Bouaita Riad, ENSET SKIKDA, Lire Laboratory, Abdelhamid Mehri Constantine 2 University Ali Mendjli, Constanine, Algeria
Zitouni Abdelhafid, Lire Laboratory, Abdelhamid Mehri Constantine 2 University Ali Mendjli, Constanine, Algeria
Maamri Ramdane, Lire Laboratory, University of Constantine 2 - Abdelhamid Mehri, Constanine, Algeria

ABSTRACT

Cloud customers tend always to overestimate their resource requirements and thus, they utilize only a portion of the allocated resource which gives an opportunity for cloud providers to oversubscribe their resources. Oversubscription is a powerful technique that leverages unused resources which improves the profit of cloud providers while minimizing cost for customers. However, the benefits of this technique are without inherent risks: it increases the possibility of overload. This article proposes an autonomous architecture that uses memory oversubscription to maximize resources utilization rate. This architecture uses live migration of VMs as well as network memory as two strategies to mitigate overload generated by oversubscription.

KEYWORDS

IaaS Cloud, Live Migration, Network Memory, Oversubscription, Service Performance

INTRODUCTION

Resource capacity planning is becoming a big challenge for cloud providers when modern applications require a huge amount of resources. To adapt with the environments where resources are dynamically requested and released, conventional approaches that overprovision VMs with a large amount of resources to meet these requirements can induce several problems such as underutilization of these resources in most cases. Oversubscription is an adequate technique that deploys more VMs as much as possible. This technique becomes feasible since customers tend to overestimate their resource requirements. It can alleviate the problem of resource underutilization which allows minimizing resource provisioning costs while meeting customers' needs. This technique has inevitably some risks: it can increase the possibility of overload especially when the oversubscriptions ratios are not managed carefully.

In this paper, we present an autonomous architecture that improves the resource utilization rate through the technique of memory oversubscription. The overload induced by this technique is mitigated by two techniques: Live Migration and Network Memory. The first strategy consists of live migrating some VMs from an overloaded host to an underloaded one without interrupting any one. The latter consists on using the global memory of the cluster as a remote paging replacing the disk swapping as it has a best performance (nano second vs. micro second) (Baset, Wang, & Tang, 2012).

DOI: 10.4018/IJGHPC.2020010103

Live migration is a technique largely used to improve service performances in an IaaS Cloud. It provides significant benefits, such as minimizing the Service Level Agreement (SLA) violations and insuring the quality of service (QoS). Although several live migration algorithms had been implemented, they do not address the migration cost. Here, our proposal aims to optimize the migration cost which is an important concern.

On the other side, the protocol RDMA (Remote Direct Memory Access) is used to mitigate the overload due to oversubscription if the network memory strategy is decided. This protocol allows to transfer data without involving the CPU or cache for both source and target hosts.

For the oversubscription ratio, unlike traditional strategies using, either static ratios or depending only of the local host capacity, we adopt a dynamic oversubscription one, taking into account the free memory of the local host as well as that available on remote hosts of the cluster. This ratio is dynamically managed by reporting the amount of oversubscribed memory in real time.

Simulation results show the effectiveness of our proposal -in term of improving service performance- compared with the use of live migration alone as a technique to mitigate overload.

Our contribution consists essentially on:

- Maximizing resource utilization rate in the cluster with the technique of oversubscription by adopting a real time and dynamic oversubscription ratio;
- Improving service performance by leveraging the global memory of the cluster trough the network as a technique to mitigate the overload at any host via the remote paging technique.

The remainder of this paper is organized as follows: In Section 2, a brief description of oversubscription is introduced in the context of Cloud Computing. Section 3 describes some related work in the field of oversubscription and overload mitigation strategies. In section 4 we propose an autonomous architecture that mitigates memory overload by the two techniques described above. To validate our proposed approach, we present our simulation study in section 5. Finally, in Section 6 we conclude our paper by specifying some future work.

OVERSUBSCRIPTION IN CLOUD COMPUTING

Oversubscription (Also called overcommitment or overbooking) is a technique offering more resources than actually available assuming that most customers would not consume their entire allocated portion, which increases the resource utilization rate and subsequently increases the profits of providers (Householder, Arnold, & Green, 2014). This technique is used in several fields. For instance, air line companies sell more tickets than the real number of seats in order to maximize the filling rate in case of discontinuance of certain customers. If the number of customers appears more than expected, they will be moved to another flight or transmitted to another company. Another example is the healthcare industry, when clinics overbook the patient appointment scheduling considering that patients have different now-show probabilities for minimizing the patient's waiting time on the one hand, and the doctor's idle time and overtime on the other hand. Noting that this technique may result in a clinic overcrowding with high number of waiting patients and doctor's overtime (Zacharias & Pinedo, 2013).

In Cloud Computing Oversubscription is a resource management technique where the sum of allocated resources (CPU, memory, storage, bandwidth, etc.) exceeds the real capacity of the physical host, assuming that customers will not use the entire of their allocated resources or they would not use them at the same time.

This technique devotes more resources than is actually available on physical machines hosting a set of applications (Caglar & Gokhale, 2014). Oversubscription can be both on the provider side and on the customer side. In this paper, we focus on oversubscription at provider's level, who shares