An Availability-aware Approach to Resource Placement of Dynamic Scaling in Clouds

Wenting Wang, Haopeng Chen, Xi Chen
REINS Group, School of Software, Shanghai Jiao Tong University
Shanghai, P.R.China
{wwtvanessa, chen-hp, chenxi_sjtu}@sjtu.edu.cn

***Abstract*- The availability of Web applications influenced by Virtual Machine (VM)-based physical locations during resource scaling is a crucial concern for customers and cloud providers. In this paper, we present a novel computing model to describe availability attribute of one application in hierarchical structured cloud. Meanwhile, we propose an availability-aware approach to explore how and where to allocate computing resource via vertical and horizontal scaling. Partial experimental results in simulation environment are also presented.**

Keywords-Cloud-computing; VM placement; Scaling up and down; Availability

I. INTRODUCTION

Dynamic scaling is one of the key abilities provided by clouds to add or remove computational resource to applications in response to workload fluctuation. Compared with intensively discussed issues of performance and profit requirements in scaling process [1][4], the availability requirement of one application is equivalently crucial especially for enterprise-level services. Scaling resolution need to consider both the locations of VM instances and geographical features of cloud to satisfy the availability demand of customers. Generally, the availability of one application is affected by three factors as follows:

- Relative Locations: how components (i.e. VM instances) of one application are placed relatively to one another [3] in distributed virtual environment mainly influence its availability. The scaling process may change the situation of the original relative locations of VMs, consequently affecting the availability of the application.
- Cloud Environment: the structure of clouds is another significant concern. For instance, Amazon provides a relatively transparent structure of EC2, namely regions and availability zones, in which allows customers to launch VM instances from a failure of a single location to protect their applications [2].
- Scaling Policy: VMs-based application scalability can be implemented by either changing the partition of resources (e.g. CPU, memory, storage, etc.) inside a VM or adjusting the amount of VM instances. These two kinds of scalability (as shown in Fig1.) behave differently on changing instance number, relative locations and communication cost, consequently affecting satisfactions for availability and performance of the application.

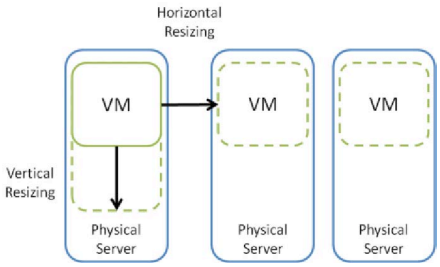

Figure1. Vertical Resizing and Horizontal Resizing

Therefore, this paper addressed the problems of: 1) Modeling the availability of applications deployed in cloud without loss of generality; 2) Resource placement of dynamic scaling in an availability-aware and communication cost-efficient way integrated with vertical and horizontal resizing policies.

II. MODELING AND APPROACH

A. Availability Modeling

Regarding to different geographical network topology and logical groupings of datacenters from multiple cloud providers, modeling the cloud structure and the availability of applications must be done in effective and scalable way to support heterogeneous environment.

We model the cloud infrastructure as a tree structure with arbitrary depth (see Fig. 2). In the tree structure of the cloud, top levels have been already taken into commercial providers' account such as Amazon [2], e.g. Regions and Availability Zones. The nodes of bottom level are Physical Hosts. Without loss of generality, the structure of middle levels (e.g. server racks) can be defined in a scalable way as needed. We assume that all the nodes in the same level are at equal height and all VMs should be placed in leaves of the tree namely physical hosts.

The failures of nodes of different levels are caused by various reasons. For instance, a region failure may be caused by natural disasters such as earthquake or flood whereas the failure in availability zone may be caused by blackout. The failure probability in multiple VMs host on different physical nodes ($U=\{u_1,u_2...u_n\}$) cannot be simply calculated as $\prod \overline{P_u(VM)}$ where $u \in U$ because the hierarchical paths of any u_i and u_j node might be not completely independent, in which they may share common parent nodes. Hence, calculating the failure probability in multiple VMs involves two parts: (i) probability of the

978-0-7695-4755-8/12 $26.00 © 2012 IEEE
DOI 10.1109/CLOUD.2012.82

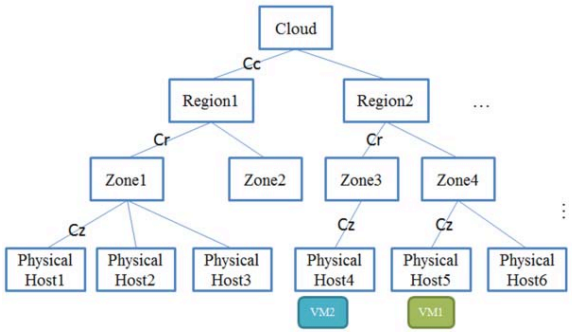

Figure2. Availability Model in clouds

failure happens in any of their common parents and (ii) the probability of failure happens in their own private hierarchic path under their common parent when all their common parents function well.

B. Communication Cost Modeling

Performance affected by communication cost is another significant issue in virtual resource scaling and attracted attention recently. In our model, communication cost between two VMs in the same host is assumed as zero, and the communication costs remain equal in the same hierarchy. Regions may be more geographically insulated than Availability Zones, consequently incurring higher communication cost which requires lower latency or more bandwidth. The communication cost between two VMs can be formalized in measuring the end-to-end communication cost between two hosts [3] via geographical distance which can be specialized by the distance between two nodes.

Thus, the problem is to find the correct placement solution of satisfying the availability demand while minimizing the cost from communication to improve the overall performance.

C. Availability-Aware Placement Approach

The increased or decreased amount of VMs predicted provisioning component offered has been studied in the literature [4]. Therefore, in our work, we only focus on how and where to allocate resource to satisfy availability demands.

Generally, the placement plan which performs both vertical and horizontal methods will be made according to the availability gap between the existing application and customer's demand by single step. When the availability requirement has been fulfilled, we vertically resize the existing resource to maintain the communication and software cost. Otherwise, horizontal resize up is prior. A new PM host is chosen by dynamically measuring the distance to the current VMs group, namely the distance threshold. The higher threshold implies the greater availability enhancement to the overall application.

When scaling down, current resource tends to be vertically adjusted to reduce the influence on overall availability. When all VMs in the application have been reduced to unit size, horizontal resizing down has to be

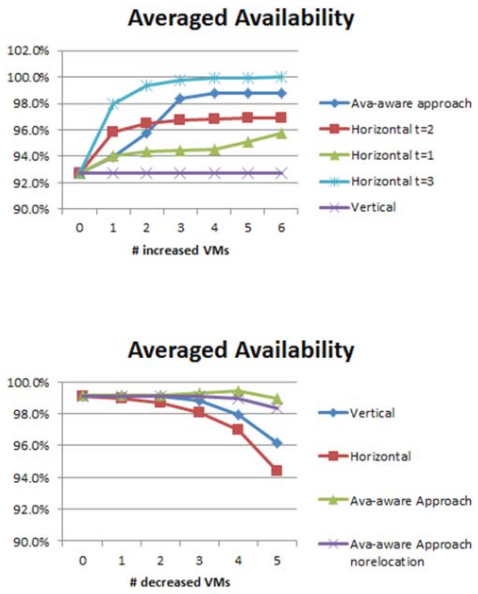

Figure3. Averaged Availability in scaling up and down

launched. The scheduling process terminates a VM whose distance to others is smallest, which maintain the availability to the largest extent.

Finally, a mechanism called "relocation" is performed to ensure the availability satisfied. Specifically, relocation focuses on enabling the application to generate a more distributed placement in limited times via relocating one of its VMs on a new host with a larger distance threshold.

III. CONCLUSION AND FUTURE WORK

This paper modeled the availability of one application deployed in cloud platform and designed an availability -aware approach by taking advantages of both vertical and horizontal resizing to dictate where and how to add or remove resource to the application. In addition, we have implemented our proposed approach in simulation platform imitated homogenous cloud environment, and have carried on a set of experiments compared with both horizontal and vertical approach respectively. The results have been briefly illustrated in Fig.3. Future work covers improving the implementation of our approach in a real cloud platform and using real applications data to evaluate the effectiveness.

REFERENCE

[1] B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitizer, A. Tantawi, "An Analytical Model for Multi-tier Internet services and its applications," In Proceedings of ACM SIGMETRICS, 2005 .

[2] Amazon Elastic Computing Cloud, https://aws.amazon.com/ec2/

[3] D. Jayasinghe, ,C. Pu,T. Eilam, M. Steinder, I.Whally, E.Snible, Improving Performance and Availability of Services Hosted on IaaS Clouds with Structural Constraint-aware Virtual Machine Placement, Services Computing (SCC), 2011 IEEE International Conference, July 2011

[4] X. Chen, H. Chen, Q. Zheng, W. Wang and G. Liu;, "Characterizing Web Application Performance for Maximizing Service Providers' Profits in Clouds ", In Proceedings of IEEE International Conference on Cloud and Service Computing (CSC), 2011.