Accepted Manuscript

Model-based sensitivity analysis of IaaS cloud availability

Bo Liu, Xiaolin Chang, Zhen Han, Kishor Trivedi, Ricardo J. Rodríguez

PII: S0167-739X(17)30179-6
DOI: https://doi.org/10.1016/j.future.2017.12.062
Reference: FUTURE 3900

To appear in: *Future Generation Computer Systems*

Received date : 8 March 2017
Revised date : 17 November 2017
Accepted date : 29 December 2017

Please cite this article as: B. Liu, X. Chang, Z. Han, K. Trivedi, R.J. Rodríguez, Model-based sensitivity analysis of IaaS cloud availability, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2017.12.062

Model-based Sensitivity Analysis of IaaS Cloud Availability

Bo Liu[a,b], Xiaolin Chang[a], Zhen Han[a], Kishor Trivedi[c], Ricardo J. Rodríguez[d]
[a]Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, P. R. China
[b]Beijing Institute of Information Application Technology, P. R. China
[c]Department of Electrical and Computer Engineering, Duke University, USA
[d]Centro Universitario de la Defensa, Academia General Militar, Zaragoza, Spain

Abstract— The increasing shift of various critical services towards Infrastructure-as-a-Service (IaaS) cloud data centers (CDCs) creates a need for analyzing CDCs' availability, which is affected by various factors including repair policy and system parameters. This paper aims to apply analytical modeling and sensitivity analysis techniques to investigate the impact of these factors on the availability of a large-scale IaaS CDC, which (1) consists of active and two kinds of standby physical machines (PMs), (2) allows PM moving among active and two kinds of standby PM pools, and (3) allows active and two kinds of standby PMs to have different mean repair times. Two repair policies are considered: (P1) all pools share a repair station and (P2) each pool uses its own repair station. We develop monolithic availability models for each repair policy by using Stochastic Reward Nets and also develop the corresponding scalable two-level models in order to overcome the monolithic model's limitations, caused by the large-scale feature of a CDC and the complicated interactions among CDC components. We also explore how to apply differential sensitivity analysis technique to conduct parametric sensitivity analysis in the case of interacting sub-models. Numerical results of monolithic models and simulation results are used to verify the approximate accuracy of interacting sub-models, which are further applied to examine the sensitivity of the large-scale CDC availability with respect to repair policy and system parameters.

Index Terms—Availability; Sensitivity analysis; Markov chain; IaaS; Cloud computing; Stochastic Reward Nets

1 Introduction

The past few years have witnessed fundamental changes caused by cloud computing to business computing models. Infrastructure as a Service (IaaS) is one of the basic cloud services. This cloud service is provisioned to customers in the form of virtual machines (VMs), which are deployed on physical machines (PMs). Each VM has specific characteristics in terms of number of CPU cores, amount of memory and amount of storage. It was reported that global spending on IaaS cloud services is expected to reach 56 Billion USD by 2020 [1]. The ever-increasing demands for IaaS cloud services have created the need for cloud service providers (CSPs) to analyze cloud infrastructure availability in order to maintain high cloud service availability [2] while reducing various costs. Service availability is commonly specified via Service Level Agreements (SLAs) [3]-[5]. Any availability violation may cause the loss of revenue. In addition, some common IaaS cloud management tools, such as OpenStack [6], have allowed configuring standby PMs for high availability. However, there is no suggestion about how to configure.

System availability is affected by various factors, such as system parameters and repair policy. The latter one determines how quickly PMs get repaired upon their failure. Repair policy analysis is significant to the CDC design with respect to CDC availability. State-space models are popular and found effective for system availability analysis [7]. They also allow the derivation of sensitivity functions of the measures of interest with respect to various system parameters, which are assigned in a continuous domain. These functions could be applied to assess the impact of each of these parameters on system quality of service (QoS) and then to identify the QoS bottlenecks for systems of interest.

This paper aims to explore analytical modeling and sensitivity analysis techniques to improve the availability of a large-scale IaaS cloud data center (CDC). Following Ghosh *et al*. [12], we assume that there are three PM pools, namely hot (running PMs), warm (turned on, but not ready PMs) and cold (turned off PMs). Thus, there are two kinds of standby PMs, warm-standby and cold-standby. A small provisioning delay is needed for deploying default VM images on hot PMs. Additional provisioning time (to make the PM ready) is required for the VM deployment on a warm PM. Further delay is added when PMs in the cold pool are used, since they need to be turned on before being used. PMs can move among pools due to failure/repair events. PM repair times of different PM pools may be different. The main reason of considering the CDC with three PM pools in this paper is that the modeling approach of this scenario could be applied directly to scenarios with arbitrary number of pools. Note that although IaaS CSPs in production CDCs have offered standby PMs for disaster recovery [8], there is no published information about the number of PM pools.

Large scale is a feature of CDCs, leading to the well-known largeness problem [8] associated with a monolithic or one-level Markov chain for the availability analysis of an IaaS CDC. Moreover, complex interactions among CDC components and different failure/repair behaviors further exacerbate the largeness problem. Our numerical results show that the