Journal of Network and Computer Applications 60 (2016) 54–67

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

Availability in the cloud: State of the art

Mina Nabi [a], Maria Toeroe [b], Ferhat Khendek [a]

[a] Electrical and Computer Engineering, Concordia University, Montreal, Canada
[b] Ericsson Inc., Montreal, Canada

ARTICLE INFO

Article history:
Received 10 August 2015
Received in revised form
18 November 2015
Accepted 24 November 2015
Available online 14 December 2015

Keywords:
Availability
Cloud computing
Evaluation
State-of-the-Art
Taxonomy

ABSTRACT

Availability is a non-functional requirement defined as the percentage of time a system or a service is accessible. This percentage determines the acceptable total outage time for any given period. High-Availability (HA) is a stringent requirement which allows for a maximum of approximately five minutes downtime in a year including outage due to scheduled maintenance and upgrades. It is agreed that availability is among the main challenges of the cloud. There has been a lot of work on availability in cloud computing, but cloud providers and researchers from the academia and the industry have used different definitions for availability and the related concepts. Thus, it is difficult to evaluate and compare different solutions. In this paper, we present a survey of availability solutions proposed for the cloud by the main cloud providers and by 21 most relevant conference and journal papers out of the 100 papers collected initially. To conduct this survey we defined a taxonomy, which captured the main concepts, mechanisms and metrics for availability. We use this taxonomy to evaluate and classify the solutions of cloud providers as well as solutions proposed in research papers, their strengths and weaknesses. We point out potential future research directions.

Contents

E-mail addresses: mi_nabi@encs.concordia.ca (M. Nabi), Maria.Toeroe@ericsson.com (M. Toeroe), ferhat.khendek@concordia.ca (F. Khendek).

http://dx.doi.org/10.1016/j.jnca.2015.11.014
1084-8045/© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Availability is a non-functional requirement specified in terms of the percentage of time a system or a service is accessible. This percentage determines the allowed outage time for a given period. High availability (HA) is a strict requirement and refers to an availability of at least 99.999% of the time, which permits for approximately five minutes of downtime per year including scheduled and unscheduled maintenance.

Availability is considered as one of the main challenges of cloud computing as more critical services are shifting towards this paradigm. Availability may hinder the adoption of cloud computing (Armbrust et al., 2010; Subashini and Kavitha, 2011). Problems may arise not only from an individual service, but from the interactions between multiple components and automated services over distributed networks and data centers.

Work has been done to enhance cloud availability (Chan and Chieu, 2012; Cully et al., 2008; Jayasinghe et al., 2011; Singh et al., 2012; Yang et al., 2011; Yang et al., 2013; Liu et al., 2012; Zhao et al., 2010). However each cloud provider defines availability from its own technological perspective and provides architectural solutions and services based on specific definitions. These different definitions use similar terms but with different interpretations and measures. Although cloud providers claim to provide high resiliency (Naldi, 2013), their actual average outage time is longer than what it is allowed for high availability. A rough assessment of cloud provider's actual outage based on cloud providers' reported outage times is given in Naldi (2013). Furthermore, many research papers relate availability to other characteristics of the cloud, such as performance, scalability, elasticity and security; they do not propose a comprehensive solution that would guarantee availability in the cloud. Given the diversity of definitions and related mechanisms, it is difficult to assess and compare these different solutions, identify the promising ones and the important research issues.

This paper surveys current availability solutions in cloud providers' practices, and those proposed in research papers from the academia and the industry. For this purpose, we first define a taxonomy inspired by the Service Availability Forum (SA Forum) (Toeroe and Tam, 2012) concepts and mechanisms to set the baseline for the evaluation. This taxonomy represents a well-defined structure that includes basic availability concepts, mechanisms, and metrics through which we can evaluate and categorize the proposed solutions. It captures the differences and similarities between various solutions in terms of availability.

We looked into the cloud providers' practices and collected 100 journal articles and conference papers on the subject of availability in the cloud. Out of these 100 papers 55 were closely related to availability in the cloud, however only 21 proposed some solution. The remaining papers deal with availability analysis or evaluation techniques like Machida et al. (2011), Bruneo (2014), Khazaei et al. (2012), Benz and Bohnert (2013), Huang et al. (1995), Longo et al. (2011), Juels and Oprea (2013) and Siebenhaar et al. (2013). We evaluated these different solutions using our taxonomy and draw some conclusions. Furthermore, we discuss potential research directions.

The organization of this paper is as follows: In Section 2, we provide examples and introduce the various definitions of availability used in research papers and by cloud providers. To cover best the key aspects revealed from these definitions, in Section 3 we propose a taxonomy that structures multiple views of availability including mechanisms, type of failures protected against, and metrics. In Section 4, we provide our evaluation and classifications of the 21 selected research papers and cloud offerings using the proposed taxonomy. In Section 5, we discuss further and summarize the weaknesses of current solutions before elaborating on future research directions. We conclude in Section 6.

2. Definitions

Understanding what availability means in the cloud is a prerequisite for the evaluation and categorization of currently deployed and research solutions, related mechanisms and technologies for the cloud. Availability and HA have been defined in many different ways. The TL 9000 Quality Management System from the QuEST Forum (QuEST Forum, 2010) for the information and communication industry defines availability as “The probability that the system is operational when required” and it is calculated by system uptime over the total time required to be operational.

Aligned with that in Toeroe and Tam (2012) availability is defined as “*the degree to which a system is functioning and is accessible to deliver its services during a given time interval.*” It is the percentage of time a system is ready to perform its functions and is calculated as:

$$Avaiability = MTTF/MTBF = MTTF/(MTTF + MTTR) \quad (1)$$

where *MTTF* (Mean Time To Failure) is the mean time it takes for the system to fail; *MTBF* (Mean Time Between Failures) is the mean time between two failures and represents the sum of MTTF and *MTTR* (Mean Time To Repair) (Toeroe and Tam, 2012).

One can view the availability of a system through the availability of its services. Service availability can be defined as:

$$Service\ Availability = Service\ Uptime/(Service\ Uptime + Service\ Outage) \quad (2)$$

where service uptime is the duration during which the system delivers the given service, while service outage (or also referred as downtime) is the period during which the service is not delivered (Toeroe and Tam, 2012).

We talk about HA if a system or a service is available at least 99.999% (a.k.a. five nines) of the time, which allows for a maximum of five minutes and 15seconds downtime per year (Toeroe and Tam, 2012).

The term availability has been extensively used in research publications related to the cloud and by cloud providers in white papers and technical documentations. Availability is considered as one of the key characteristics of the cloud (Armbrust et al., 2010). At the same time it is one of its main challenges. Therefore, a lot of research effort has been dedicated to guarantee the availability in the cloud. Taking a closer look at these efforts one realizes that the interpretation of availability may differ from paper to paper and from provider to provider as we discuss it in the following subsections.