Journal of Network and Computer Applications 61 (2016) 81–92

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Review

A survey of fault tolerance architecture in cloud computing

Mehdi Nazari Cheraghlou [a,*], Ahmad Khadem-Zadeh [b], Majid Haghparast [c]

[a] Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
[b] Iran Telecommunication Research Center (ITRC), Tehran, Iran
[c] Department of Computer Engineering, Yadegar -e- Imam Khomeini (RAH) Branch, Islamic Azad University, Tehran, Iran

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form
21 September 2015
Accepted 18 October 2015
Available online 28 October 2015

Keywords:
Cloud computing
Fault tolerance architecture
Proactive
Reactive

ABSTRACT

Utilizing cloud computing services has had numerous advantages such as the reduction of costs, development of efficiency, central promotion of soft wares, compatibility of various formats, unlimited storage capacity, easy access to services at any time and from any location and, most importantly, the independency of these services from the hardware. It should be mentioned that the provision of various cloud computing services is faced with problems and challenges that the fault tolerance can be mentioned as the main restrictions.

In this paper, first, methods of creating the capacity of Fault Tolerance in Cloud computing are pointed out. Subsequently, policies of the implementation of these methods are stated. Next, the offered architectures for the production of such capacity in Cloud computing is delineated and, finally, they are compared in terms of the type of policy or policies employed in the architecture and the method of fault detection and fault recovery.

Contents

* Corresponding author.
E-mail addresses: ir.m.n@ieee.org (M. Nazari Cheraghlou), zadeh@itrc.ac.ir (A. Khadem-Zadeh), haghparast@iausr.ac.ir (M. Haghparast).

http://dx.doi.org/10.1016/j.jnca.2015.10.004
1084-8045/

1. Introduction

The emergence of Cloud is the biggest change in the world of IT, leading to the stimulation of all individuals and all companies. Complete definition of cloud computing which is considered a standard for cloud computing is the golden definition that is presented by NIST institute (Mell and Grance, 2011). In this definition, it is expressed that cloud computing is a demand-based and easy access model under a network to a sharing set of configurable computing resources (including servers, networks, storage devices, applications, and services). The resources are provided and used quickly and they are released with minimal effort and cost.

1.1. Cloud computing models

Cloud computing can be implemented in the forms of public, private, community and hybrid cloud (Mell and Grance, 2011). Using the public cloud is possible for all but private cloud is dedicated to a collection and only members can take advantage of cloud services. Community cloud is exploited in sharing form for individuals or organizations that have similar missions and needs. Hybrid cloud is a combination of two or more different clouds that each of them should be able to provide more combined services together while preserving their separate identities.

1.2. Cloud computing challenges

Cloud is a virtual and abstract image of a large network that neither its volume nor its size and also processing and storage resources are specified and limited. Location and time are also unknown and unlimited in the cloud. This means that resources' location is hidden from the perspective of users and applications and presentation time and completion of the services and they seem to be unlimited. Desired services can be accessed from any place and at any time. These characteristics cause removing the restrictions in using systems and traditional networks in providing service to users, but they may bring some new problems, restrictions, and challenges for users and applications. On top of these problems, the fault tolerance challenge of cloud has a special status and importance. Because if we have the best clouds with the best services but they do not have fault tolerance, they are not reliable and expectation of appropriate and desired service are futile. Therefore, fault tolerance means that if a fault occurred in a cloud, it should be able to detect and identify that fault and recovery and improve it without any damage to the final output of cloud computing. This capability causes that the cloud be able to have an optimum and acceptable performance in the presence of the faults.

1.3. Fault tolerance techniques: overview

The techniques that are used to create the fault tolerance capability in cloud computing can be divided into three general categories. The first is redundancy techniques, the second is tolerance policies against and finally, the third is load balancing fault tolerance. Redundancy techniques include hardware redundancy and software redundancy and time redundancy. Hardware redundancy is a technique of structural redundancy which masks the fault using a complete series of modules. The method is in such a way that an identical number of hardware modules perform identical operations. Hardware modules' inputs are the same and the modules' outputs are aggregated and then, a majority vote is taken. Thus, the fault effect will be removed in the output. It is worth noting that voting will be done by TMR: Triple Modular Redundancy. In the case of software redundancy, our agenda is running a program with the same input but with different implementation algorithms. In other words, the data processing method is identical but it is performed using different algorithms on the same input data. Obviously, the same outputs are expected but if different outputs be achieved, the correct output can be achieved by the majority vote of outputs. In the case of time redundancy, identical hardware and software are as the constant parameters of the problem. We pursue multiple running of a program in an identical hardware. In this case, we can completely cover the fault in the output with the majority vote of the performance results of running a program.

Fault types that have been mentioned in (Kumar et al., 2015; Saikia and Devi, 2014; Amin et al., 2015) and occur in cloud computing are different based on computing resources. Among them are Network Fault, Physical Faults, Process Faults, Processor Faults and Service Expiry Fault.

Techniques for creating and increasing the fault tolerance based on the load balancing also can be performed based on hardware and software and also based on the network (Singh and Kinger, 2013). In the case of hardware, requests from a client are sent to the hosts of a cluster. In the case of software, we have a dispatcher server that is performed on all incoming requests. The disadvantage of this method is that this dispatcher server has a high potential for bottleneck. The third method which is a software solution based on the network, does not require any additional hardware. There is also no need to have a central dispatcher to be a bottleneck because, all host receives incoming packets and redundancy occurs with respect to the number of clusters. The packet filtering algorithm is very effective in handling the packets.

Among other applications of Load Balancing Mechanism in cloud computing, we can refer to Resource Management which has been discussed in (Manvi and Krishna Shyam, 2014). Resource Management in cloud computing is accompanied by advantages including scalability, QOS, reduced overhead and increased throughput. Resources are generally divided into physical and logical groups. Logical resources provide temporary control over physical resources. Also, logical resources support the development of applications and effective communication protocols. Load Balancing mechanisms, in addition to being considered as one of the methods to increase the fault tolerance of cloud computing, provide Logical Resource Management in cloud computing. Given that physical facilities in cloud computing are placed in a distributed manner, Network Resource Management develops and improves using load balancing techniques. Further, fault tolerance increases at the same time.

The last way to create fault tolerance in the cloud computing services is creating this capability based on using a series of policies. These policies are divided into two categories of proactive and Reactive which will be studied. The method in proactive is such that the fault in cloud computing is estimated and necessary precautions should be considered. But the working style in the reactive group is different and there will not be any prediction and prevention of fault because it wastes resources and increases the response time of the system especially in the case of real time