Journal of Network and Computer Applications 77 (2017) 87–105

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Review

Federated cloud resource management: Review and discussion

Misbah Liaqat[a,b,*], Victor Chang[c], Abdullah Gani[a,b,*], Siti Hafizah Ab Hamid[a], Muhammad Toseef[d], Umar Shoaib[d], Rana Liaqat Ali[e]

[a] Department of Computer System and Technology, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia
[b] Center for Mobile Cloud Computing Research, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia
[c] Xi'an Jiaotong Liverpool University, Suzhou, China
[d] Faculty of Computing and IT University of Gujarat, Pakistan
[e] Department of Physics, COMSATS Institute of Information Technology, Islamabad, Pakistan

ARTICLE INFO

Keywords:
Resource management
Cloud computing
Federated clouds

ABSTRACT

Federation of clouds is the future of cloud computing, mobile cloud computing, Internet of things, and big data applications. The utilization of federated resources is envisioned to increase the quality of service, cost benefits, and reliability. Resource management in the federation of multiple clouds is a pressing issue owing to the lack of cross-domain knowledge, security, trust, and administrative policies. This study classifies these resource management functions in the federated cloud environment into resource pricing, resource discovery, resource selection, resource monitoring, resource allocation, and disaster management. Each federated resource management function is discussed, and insights into state-of-the-art research are then provided. These resource management functions are also compared based on the performance metrics that are suitable for an individual function. Finally, we indicate the open challenges for further research with regard to each classified resource management function.

1. Introduction

Cloud computing (Armbrust et al., 2010; Buyya et al., 1969; Nurmi et al., 2009; Vaquero et al., 2008), a user-centric computational model, is a flexible paradigm of deploying and sharing distributed services and resources with the pay-per-use model. With virtual machine (VM) technology (Smith and Nair, 2005) and data centers (DCs), computational resources, such as memory, central processing unit (CPU), and storage, are dynamically reassembled and partitioned to meet the specific requirements of end users. The cloud not only considers flexible platform-independent access to resources and information anywhere and anytime but also changes the way of designing, deploying, building, expanding, and running applications (Cao et al., 2009). The demand's growth for cloud services is presenting considerable challenges for cloud providers to meet the requirements and satisfaction of end users (Talia, 2012).

Research by Bakshi (2009) and Bernstein et al. (2009) showed that the trend in cloud computing pattern will shift from a single provider to federated clouds, which are expected to include numerous distributed public and private cloud platforms (Lopez-Rodriguez and Hernandez-Tejera, 2011). Federated clouds enable public and private clouds to share their resources with each other to scale up their resource pools at peak times. The promises of nearly infinite computing power, concomitant storage, and economies of scale can only truly be achieved by cloud federation. The reason behind cloud federation is the finite physical resources in the resource pool of a single provider.

In the federated cloud environment, a cloud provider acts as both infrastructure provider and consumer. The egocentric and rational behavior of federation members focuses on maximizing their revenue and resource utilization by serving as many consumers as possible. Here, consumers can be either federation members who rent resources from one another or regular cloud users. The role of efficient resource management is prominent in such a scenario to guarantee the service request of both the federation members and direct consumers. The resource management functions in the federated cloud environment ensure the objective of federation members and the aggregate utility of the federation, which is necessary for the continuation of the federation.

This study focuses on the resource management functions of federated cloud in which an individual cloud provider provides and consumes Infrastructure as a Service (IaaS) to and from other federation members. The terms inter-cloud, federated cloud, and multi-cloud are interchangeably used in this article for the federation of cloud providers. Three surveys (Grozev and Buyya, 2014; Petcu,

* Corresponding authors at: Department of Computer System and Technology, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia.
E-mail addresses: mibahaseeb@siswa.um.edu.my (M. Liaqat), abdullah@um.edu.my (A. Gani).

http://dx.doi.org/10.1016/j.jnca.2016.10.008
Received 13 May 2016; Received in revised form 14 October 2016; Accepted 18 October 2016
Available online 20 October 2016
1084-8045/

2014; Toosi et al., 2014) were previously published on cloud federation. However, the targets of these surveys outlined the taxonomies, terminologies, definition, and challenges for inter-cloud systems. This survey differs from them in the essence that we focus only on the resource management aspects of inter-cloud computing and study the proposed solutions for inter-cloud resource management issues. Our contributions are as follows: (1) the federated resource management mechanism is classified into resource pricing, resource discovery, resource selection, resource monitoring, resource allocation, and disaster management. (2) Rigorous works on resource pricing, resource monitoring, resource discovery, resource selection, resource allocation, and disaster management are given. The operation and drawbacks of the mechanism are presented with a comparative analysis in terms of different performance metrics for each considered class. (3) Open challenges in each considered class of federated resource management is indicated. (4) Novice researchers are encouraged to work on the research problems.

The remainder of this article is divided into three sections. Section 2 presents a general background of cloud computing, federated cloud computing environments, and resource management in federated cloud environments. Section 3 provides the classification of resource management functions along with the discussion of the state-of-the-art literature. Section 4 concludes the paper.

2. Background

2.1. Cloud computing

The fundamental idea of cloud computing is to deliver computational resources as services over the Internet. Consumers are not required to invest in a large computer system to conduct their business; instead, they can acquire cloud computing services on the basis of their demands (Liaqat et al., 2016). The underlying hardware is commonly hosted in large DCs using sophisticated virtualization practices to realize high-level agility, scalability, and availability. According to Armbrust, the resource's elasticity, which does not require a premium to be paid for the pool of resources, is unprecedented in the IT history (Armbrust et al., 2010). Therefore, the establishment of enterprises systems using the cloud platform has gradually become popular in recent decades. In addition, in conventional computing system s, users are required to invest in dedicated software and hardware whereas cloud computing delivers the software and hardware resources in a pay-per-use manner (Foster et al., 2008). This feature considerably reduces the cost of maintenance and deployment. Cloud computing is offered using three service delivery models, i.e., Platform as a Service (PaaS), Software as a Service (SaaS), and IaaS (Aversa et al., 2010).

An impressive increase in the demand for and availability of cloud systems, which are best represented by Amazon's system, has been observed in the last few years. Amazon's EC2[1] is one of the most extensively used cloud services and offers a variety of VMs with different capacities where customers use the VM instances to run their application based on utility computing. Moreover, Rackspace[2] and Joyent[3] provide similar services to their consumers. The public availability of these services is often referred to as public cloud. By contrast, the infrastructure used for the particular organization is known as a private cloud (Tolba and Ghoneim, 2015). For example, EUCALYPTUS is a software environment that is used for the deployment of the private cloud and has compatibility-related concerns with Amazon's EC2. Furthermore, it represents the extensible and modularized policy for resource allocation. Currently, EUCALYPTUS supports two simple types of policies, namely, round robin and greedy (Sotomayor et al., 2009).

[1] http://aws.amazon.com/ec2/.
[2] http://www.rackspace.com/cloud/.
[3] http://www.joyent.com/.

2.2. Federated clouds

Current cloud technologies are designed according to the needs of service providers. The main expectation of the providers is that one size fits all needs. Therefore, the service consumers are forced to adapt their applications to the available stack of software. In the standard cloud computing model, in which a customer employs a single cloud service provider (CSP) and pays for resources, service disruption can affect customers that depend exclusively on it without having access to it. Depending on a single cloud provider makes enforcing sufficient usability and responsiveness to customers throughout the globe difficult. The utilization of resources and services from multiple clouds is motivated by the requirements of their customers or their providers. The usage of services from numerous clouds is driven by the reasons stated in Table 1.

Several scenarios can be considered to accomplish cloud federations, e.g., hybrid cloud, multi-cloud, and aggregated clouds (Ferrer et al., 2012; Grozev and Buyya, 2014; Petcu, 2014) for the dynamic cooperation and balancing of workload among a set of cloud DCs and providers. Cloud federation is not yet mature mainly because of the diversity of the approaches to the concept implementation. The desired semi-automated management through the selection of offers, based on the monitoring tools for the quality of services (QoS), is not yet technically possible. The differences among the current application programming interfaces (APIs) hinder the easy composition or configuration of services to be consumed from multiple clouds. Several technical barriers should be overcome, such as interoperability and portability, data and service mobility, and middleware openness, to make the usage of services from multiple clouds a reality. Many researchers currently focus on these barriers, and the literature has presented considerable innovative solutions for particular problems. However, middleware prototypes that can support a large number of scenarios of using services from multiple clouds remain lacking.

2.3. Resource management

A cloud computing infrastructure, whether single or federated cloud, is a complex distributed system composed of a multitude of computational resources. These resources handle the unpredictable client requests and the effects of external events beyond user and system administrator control. Cloud resource management significantly affects the performance, functionality, and cost factors of system evaluation. Cloud resource management also involves complex decisions and policies for multi-objective optimization. This task is challenging because of the complexity, geographical span, and unceasing and unpredictable interactions with the system, thereby making a precise global information state impossible.

Cloud resource management strategies related to the three delivery models of cloud, namely, PaaS, IaaS and SaaS, differ from one another. In all cases, the CSPs are faced with fluctuating, large workloads that challenge the claim of cloud elasticity. In some cases, when they can predict a workload spike, they can provide resources through advance reservation, e.g., seasonal web application may be subject to spikes.

For an unplanned spike, the situation is complicated. Auto scaling can be used for unplanned spike loads, provided that a monitoring system that justifies the decision to allocate, reallocate, or release resources on demand in real time exists. Auto scaling services are given by PaaS providers, such as Google App Engine.[4] Auto scaling for IaaS is complex because of the lack of and the deficiencies in the available standards.

In cloud computing, whether single or federated, variation is

[4] https://cloud.google.com/appengine/.