Journal of King Saud University – Computer and Information Sciences (2017) **29**, 19–28

King Saud University

**Journal of King Saud University –
Computer and Information Sciences**

www.ksu.edu.sa
www.sciencedirect.com

A novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing

Aarti Singh [a,*], **Dimple Juneja** [b], **Manisha Malhotra** [a]

[a] *MMICT & BM, M.M. University, Haryana 133207, India*
[b] *DIMT, Kurukshetra, Haryana, India*

Received 9 April 2015; revised 3 August 2015; accepted 12 September 2015
Available online 17 November 2015

KEYWORDS

Cloud computing;
Cloud mobile agent;
Cost optimization;
Resource allocation;
Virtual machine

Abstract A cloud computing environment offers a simplified, centralized platform or resources for use when needed at a low cost. One of the key functionalities of this type of computing is to allocate the resources on an individual demand. However, with the expanding requirements of cloud user, the need of efficient resource allocation is also emerging. The main role of service provider is to effectively distribute and share the resources which otherwise would result into resource wastage. In addition to the user getting the appropriate service according to request, the cost of respective resource is also optimized. In order to surmount the mentioned shortcomings and perform optimized resource allocation, this research proposes a new Agent based Automated Service Composition (A2SC) algorithm comprising of request processing and automated service composition phases and is not only responsible for searching comprehensive services but also considers reducing the cost of virtual machines which are consumed by on-demand services only.

1. Introduction

Cloud computing is a business model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be elastically provisioned on demand through world wide web and released with minimal management effort or service provider interaction. From the vast number of resources available on the cloud, end user is required to pay only for services provided by a concerned service provider. There are a number of virtual machines (Ezugwu et al., 2013) present at cloud datacenter and each virtual machine handles one resource with many instances, respectively and since the resources are used at the request of end user, therefore the cost of usage increases automatically which becomes a major bottleneck in the deployment of too many virtual machines. With unlimited number of resources at cloud data center, allocating and discovering active and most suitable service resource is another major challenge. Besides various pros and cons associated with this

[*] Corresponding author. Tel.: +91 9896710680.
E-mail address: singh2208@gmail.com (A. Singh).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2015.09.001
1319-1578 © 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

technology, enterprises are willing to execute their businesses by shifting either to public, private or hybrid cloud where public cloud offers the services and infrastructure off-site over the Internet, private cloud maintains the services and infrastructure on a private network and hybrid cloud includes a variety of public and private options with multiple providers adding to the cost of multiple security providers also. In contrast to public clouds which is known to be the most efficient in sharing resources, private clouds are more efficient in terms of security adding toward a high cost (Su et al., 2013) of maintaining the software and infrastructure. In order to manage the aforementioned cloud centers, there exist stringent requirements and the management of the same becomes complicated. Hence the need of a novel strategy for resource management is quite apparent. This work argues that mobile agents can potentially manage resource allocation, especially in distributed applications while considering request processing, automated service composition and cost optimization of virtual machines as primary factors. The current work proposes an autonomous service composition framework relaxing the management requirements to a large extent and hence overcoming the abovementioned cons. Next subsection provides a brief overview of networking architecture associated with typical data centers.

1.1. Data center networking architecture overview

A cloud data center mainly comprises of servers, infrastructure, power draw or power supporting devices (UPS, generators etc.), cooling devices and networking components (Lenk et al., 2009). Each of these components has some cost involved with them. Cost involved with server, infrastructure (Marrone et al., 2013) and power supporting devices and cooling devices in not of direct concern for the user, however networking components are the backbone of accessing cloud services for the user. Thus, networking is of importance for end users and service providers. When user's task is computation intensive and requires resources from more than one server then the speed of computation may drop because of an increase in propagation delay between servers. Further, cost of communication and cost of service would increase with increase in distance between data centers. This is due to the fact that cost of WAN is significantly more than the cost of LANs. Thus the overall cost of providing service to the user will increase if the service involves use of physically distributed servers. Use of micro data centers has been suggested to improve efficiency of services to the users. Further agility (Greenberg et al., 2009) is the key to reduce cost of cloud services for users which is the ability to grow and shrink resources to meet demand of the user and to draw those resources from optimal locations.

The existing physical structure of data centers causes hindrance in optimal utilization of available resources. Fig. 1 given below illustrates physical architecture of a data center, it is taken from CISCO Systems (2004).

In a data center requests arriving from the Internet are at layer 3 and they are identified using an IP address (layer 3), they are routed through border and access routers to a layer 2 domain based on the destination virtual IP address (VIP). The VIP is configured onto the two load balancers connected to the top switches, and complex mechanisms are used to ensure that if one load balancer fails, the other picks up the traffic. For each VIP, the load balancers are configured with a list of direct IP addresses (DIPs), which are the private and internal addresses of physical servers in the racks below the load balancers. This list of DIPs defines the pool of servers that can handle requests to that VIP, and the load balancer spreads requests across the DIPs in the pool.

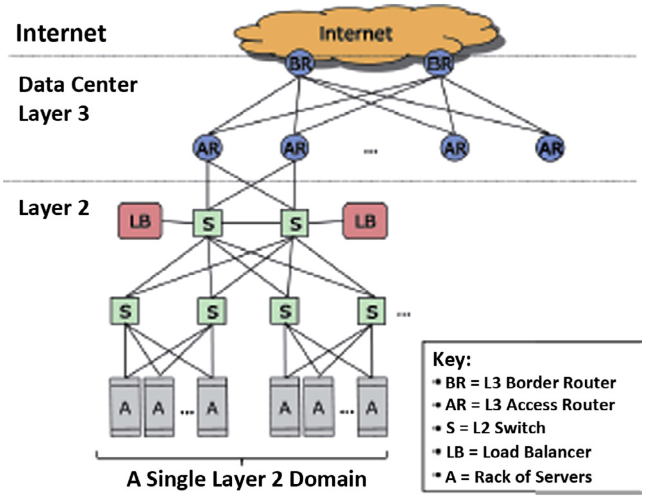

Figure 1 Cloud data center networking architecture as suggested by CISCO (CISCO Systems, 2004).

For efficient service provisioning in clouds, user requests should be allocated to any server in one data center or other data centers, however present network architecture does not support agility to that extent and leads to issues such as fragmentation of resources and poor server to server connectivity. Researchers are making efforts to improve these networking hurdles (Greenberg et al., 2008) however these problems still prevail. This work aims to propose an intelligent service composition and provision mechanism so that prior to allocation of resources, optimal hardware resources may be allocated to user and some of networking problems may be avoided.

Rest of the paper is structured as follows. Section 2 discusses the related work in this field. Section 3 describes the proposed technique, algorithms and flow diagram based on it. Section 4 elaborates on the results and comparisons with the existing techniques. Finally conclusion is given in Section 5.

2. Related work

The section throws light on the work of some renowned researchers who had been pillars and founders of the current research work.

Research on resource management strategies in different fields (Chia-Ming et al., 2014) of distributed computing with different policies is not new. However in CC, dynamic resource provisioning (Quang-Hung et al., 2014) without delay or any compromise on delay is of utmost concern. Since, ubiquity and cost-effectiveness are two keywords describing CC, cost effectiveness centers on optimal resource allocation. Literature has been reviewed to explore existing strategies of resource allocation and scope of improvement. Buyya et al. (2002, 2003) presented resource allocation frameworks which could optimize the objective function for users and resource providers. Li et al. (2009) offered scheduling and optimization