Contents lists available at ScienceDirect

Information Sciences

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

The placement method of resources and applications based on request prediction in cloud data center

Liang Quan [a,b,*], Zhang Jing [a], Zhang Yong-hui [a], Liang Jiu-mei [c]

[a] *School of Information Sciences and Engineering, Fujian University of Technology, Fuzhou 350108, China*
[b] *School of Information Engineering, Beijing University of Science and Technology, Beijing 100083, China*
[c] *School of Chemistry and Chemical Engineering, Hunan Institute of Engineering, Beijing 411101, China*

ARTICLE INFO

Article history:
Received 16 May 2013
Received in revised form 9 April 2014
Accepted 16 April 2014
Available online xxxx

Keywords:
Cloud computing
Cloud data center
Request prediction
Utility ratio matrix
Resource reconfiguration

ABSTRACT

The cloud computing data center has numerous hosts as well as application requests. It needs to allocate resources dynamically according to the varied demands of users. It should not only provide QoS such as short response time and high throughput but also achieve efficient power consumption. This paper first puts forward a reconfiguration framework based on a request prediction, which anticipates the application request volume in advance. To determine the objective of relatively optimal configuration, it can work out the allocation scheme which can improve the resource utilization ratio as well as lower energy consumption. In addition, a concept of Utility Ratio Matrix (URM) is put forward to represent allocations of hosts and Virtual Machines (VMs), and a reconfiguration algorithm based on request prediction is also presented. The algorithm will predict the application requests so as to work out the allocation scheme in advance. The algorithm can separate the reconfiguration computing from the real allocation so that it can avoid a time delay between the reconfiguration result and the varied demands, and can also reduce the energy consumption in data center. The corresponding analysis and experimental results indicate the feasibility of the reconfiguration algorithm in this paper.

1. Introduction

Cloud computing delivers infrastructure, platform, and software that are made available as subscription-based services in a pay-as-you-go model to consumers. These services are referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) respectively [8,13]. Clouds aim to power the next generation data centers as the enabling platform for dynamic and flexible application provisioning. This is facilitated by exposing data center's capabilities as a network of virtual services (e.g. hardware, database, user-interface, and application logic) so that users are able to access and deploy applications from anywhere in the Internet driven by the demand and QoS (Quality of Service) requirements [7]. Therefore, Clouds need to run many user applications simultaneously, which appear to the users as if they could use all the available resources in the Cloud.

How to meet the requests of a huge number of applications while providing the QoS guarantees is one of the main challenges for cloud data centers. Virtualization is a key technology adopted by data centers to cope with this challenge,

[*] Corresponding author at: School of Information Sciences and Engineering, Fujian University of Technology, Fuzhou 350108, China.
E-mail address: liangquanlq@gmail.com (Q. Liang).

http://dx.doi.org/10.1016/j.ins.2014.04.026
0020-0255/© 2014 Elsevier Inc. All rights reserved.

which provides the necessary abstraction so that the underlying fabric (compute, storage, and network resources) can be unified as a pool of resources to build resource overlays (e.g. data storage services, Web hosting environments).

However, to leverage the virtualization, cloud data centers must have intelligent resource allocation mechanisms. Dynamic resource configurations in accordance with changing requirements and resource status are required. Nevertheless, demands may change in real time, which may lead to a consequence that the derivation of new configurations lags behind the request variations. Considering the time consumed in the adjustment of VMs, nodes, and other resources, it will aggravate the time delay and may lead to QoS violation. Among many application scenarios such as web applications and cluster systems, similar dynamic resource reconfiguration policies have been adopted, which have invariably shown the ubiquitous time delay [16,23,24]. In addition, a data center is often equipped with a large storage system and many computing servers, leading to a large power consumption. Therefore, the power cost is a critical factor that limits the scale and efficiency of cloud data centers. The adoption of an efficient and reliable deployment policy of VMs and other resources to improve the resource utilization ratio while lowering the power consumption is another problem faced by cloud data centers, which is significant for building an energy-efficient green network environment [20].

Aim to solve the aforementioned two problems, the main contributions of this paper are as follows. First, a reconfiguration framework based on request prediction is provided so as to cope with the dynamic re-deployment of VMs and resources in cloud data centers. Second, to avoid the delayed reconfiguration, a prediction method based on Modifying Index Curve Model is presented. Third, an algorithm, called AVMR, is developed to reconfigure VMs and other resources. Based on the prediction of application request volumes, AVMR algorithm separates the configuration computation from the actual reconfiguration. Namely, it derives a specific reconfiguration in advance, avoiding the potential delay between the varied requirement and the reconfiguration.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 shows a virtualization reconfiguration framework for cloud data centers. Section 4 presents the reconfiguration algorithm of VMs and resources based on request prediction. The experimental data and analysis is described in Section 5. Section 6 concludes the paper.

2. Related works

Ref. [17] gives a comprehensive illustration of the evolution, technical problems and existing challenges of data centers. It also defines a layered model for data centers and provides a detailed description of the state of the art and emerging challenges in managing storage, networking, and compute resources and in doing power/thermal control.

To traditional digital data center, a resource on-demand approach is proposed for Web applications, which can efficiently online reconfigure clusters in response to time-varying resource requirements. It can also dynamically decide the number of running nodes and virtual machines deployed on them [22]. For dynamic resource provisioning in large-scale enterprise data centers, researchers proposed a scalable algorithm that can produce within 30 s high-quality solutions for hard placement problems with thousands of machines and thousands of applications [27]. Another Ref. [16] also introduces and evaluates a middle ware clustering technology capable of allocating resources to web applications through dynamic application instance placement. It defines application instance placement as the problem of placing application instances on a given set of server machines to adjust the amount of resources which available to applications in response to varying resource demands of application clusters. Ref. [25] proposes a resource on-demand approach for Web applications, which can efficiently online reconfigure clusters in response to time-varying resource requirements. It can also dynamically decide the number of running nodes and virtual machines deployed on them. It first predicts the future workloads of the applications with Brown's quadratic exponential smoothing method to make reconfiguration catch up with demands.

Bobroff and et al. put forward a dynamic VMs migrating method, in which the unnecessary nodes will be shut down. In this method, Linear Time Series Prediction is applied to predict the VM's demands on resources, and the VMs are listed in a descending order according to their demands. Then, apply First-fit Knapsack Algorithm to deploy VMs on proper nodes [5]. Kusic and et al. put forward a dynamic resource allocation framework based on Limited Control Prediction. The framework through a two-layer control architecture can work out the number of VM duplicates that should be set on, the position of the node where the duplicate VMs lay as well as the resource amount allocated to VMs on a same node. Although the method can improve the resource utilization ratio by shutting down the unnecessary nodes, yet its computation complexity is extremely exponential [19]. Based on the analysis on topology characteristics and traffic patterns of data centers, Ref. [11] presents a novel approach called VM Planner for network power reduction in the virtualization based data centers. The basic idea of VM Planner is to optimize both virtual machine placement and traffic flow routing so as to turn off as many unneeded network elements as possible for power saving. There are also other researches on resource scheduling and configuration in data center [28,26,12,21], however, all of which are related but different with works in this paper.

Ref. [2] proposes a coordinated cooling-aware job placement and cooling management algorithm which is Highest Thermostat Setting (HTS). HTS is aware of dynamic behavior of the Computer Room Air Conditioner (CRAC) units and places jobs to reduce cooling demands from the CRACs. HTS also dynamically updates the CRAC thermostat set point to reduce cooling energy consumption. Buyya et al. also made continuously deeper researches on energy-consumption of cloud data center and put forward some good ideas and methods [3,6,18,4]. And Dougherty et al. put forward methods of green cloud computing infrastructure to facilitate to obtain a lower energy consumption [14,10,15]. In addition, there are many researches on how to reduce the energy consumption in data center, we are not going to repeat them.

Please cite this article in press as: Q. Liang et al., The placement method of resources and applications based on request prediction in cloud data center, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.04.026