Journal of Network and Computer Applications 40 (2014) 206–215

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Optimization on content service with local search in cloud of clouds

Lingfang Zeng [a], Yang Wang [b,*]

[a] Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576, Singapore
[b] IBM CAS Atlantic, Faculty of Computer Science, University of New Brunswick, Canada E3B 5A3

ARTICLE INFO

Article history:
Received 16 November 2012
Received in revised form
5 August 2013
Accepted 18 September 2013
Available online 27 September 2013

Keywords:
Cloud of Clouds
Mobile service
Content distribution
Data movement
Local search

ABSTRACT

This paper studies the problem of distributing a content service in Cloud of Clouds to satisfy a sequence of mobile request demands with minimum monetary costs. A content may have single or multiple replicas, each being stored in one or more virtual machines (VMs) to facilitate the accesses via downloading to or replicating at (i.e., migrating) the VM sites where the mobile requests are made. As the origins of the mobile access patterns are frequently changed over time, this problem is particularly important for the users to achieve improved QoS for those time-bounded services and also beneficial to the service providers to minimize the expense on using the cloud infrastructure as well. However, these benefits do not come without compromise. The content distribution comes at cost of bulk-data transfer and service disrupts, which may increase the costs of the services. In this paper, to reap the content migration benefits while minimizing the service costs, we propose an online heuristic algorithm based on the local-search techniques to migrate the content replicas in adaptation of the mobile access patterns. For comparison purposes, we also study this problem in its off-line form, and propose a heuristic to evaluate the algorithm via a YouTube trace-based simulation. The simulation results show that our algorithms can effectively adapt to the changes of mobile access patterns to efficiently satisfy the user requests in a cost effective way.

1. Introduction

A Cloud of Clouds is generally viewed as the next revolution in the Cloud computing paradigm wherein the computational and storage infrastructure for handling scientific, business and enterprise applications could span across multiple Clouds and data-centers. However, as the complexity, heterogeneity and scale of applications grow, it will be increasingly important to be able to compose federated "Cloud of Clouds" that can address requirements for heterogeneous capabilities and large scales. For example, with an anticipated growth of mobile users of cloud services in the near future, issues related to the interoperability between cloud service providers (CSPs) and users are becoming challenging.

A content provider may reduce its expense by taking advantage of the geographically distributed datacenters and the competitive prices offered by different CSPs. Also, the content provider may use the on-demand scaling feature of the Clouds and easily adjust its requirement for the cost-effective computation powers and storage capacities. Since the costs of computation, storage and bandwidth resources are highly sensitive to the varying access models, to efficiently distribute the contents, optimizations on the content placement strategies are usually required. Although the content distribution network (CDN) problem has been widely studied in the literature (DiPalantino and Johari, 2009; Li et al., 2010; Borst et al., 2010; Jiang et al., 2009), there are some unique challenges when considering this problem in a Cloud of Clouds environment. On one hand, a content provider can construct an arbitrary network based on the demands to facilitate the accesses. This overlay network may have different topologies with respect to the underlying physical networks provisioned from infrastructure service provider (ISP). As a consequence, the content service is becoming a joint problem, requesting for both access routing and content distributing in a Cloud of Clouds. On the other hand, in Clouds the charge models for uploading and downloading the content replicas are often asymmetric with different prices, which implies that the content replication directions are usually needed to take into account in the distribution decisions (Chen et al., 2012).

To maximize the benefits of the content migration while minimizing the monetary cost of the computation, storage and transferring, in this paper, we study this content service problem by proposing an online heuristic algorithm to effectively distribute the shared content replicas in a Cloud of Clouds environment for mobile applications via content service migration and replication to adapt to the mobile access patterns. Our primary objective is to reduce the monetary cost which could exemplify the features of

* Corresponding author.
E-mail addresses: elezengl@nus.edu.sg (L. Zeng), lfzeng@hust.edu.cn, ywang8@unb.ca (Y. Wang).

1084-8045/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2013.09.006

system qualities such as bandwidth, virtual machine (VM) and storage utilization. A content provider always considers the budget constraints since the resource provisioning of cloud systems is typically based on *pay-as-you-go* model.

Although the benefits are obvious, content service migration and replication could incure high cost of bulk-data transfer and service disruption. However, several researchers have demonstrated that it is feasible to migrate virtual services in a wide-area network (Hirofuchi et al., 2009; Voorsluys et al., 2009; Wood et al., 2011). Furthermore, regarding service migration, some preliminary results on single server migration have already been achieved with respect to virtual networks (Bienkowski et al., 2010; Oikonomou and Stavrakakis, 2010) and autonomic networks (Mortier and Kiciman, 2006; Dobson et al., 2006). On the other hand, the time delay of accessing the contents during service migration can also be hidden if we properly implement the algorithms in reality,[1] rendering our proposed algorithms to be practical.

This paper makes the following main contributions:

- We study the problem of content service in a Cloud of Clouds for the ease of mobile accesses with minimum costs. We model the content placements in the Cloud of Clouds as an optimization problem.
- By exploiting the local-search techniques, we develop an efficient heuristic solution to this problem.
- We evaluate the heuristic via YouTube trace-based simulations, and show that our algorithms can adapt to the changes of mobile access patterns to efficiently satisfy the user request sequences in a cost effective way.

The remainder of the paper is organized as follows. In Section 2, we discuss the motivation with comparisons to related work. In Section 3, we describe the system model which forms the basis for our discussion. The two local-search based online algorithms are proposed in Section 4 while a heuristic for the off-line setting is presented in Section 5. All algorithms are then evaluated in Section 6, and finally, Section 7 concludes the paper.

2. Related work and motivations

2.1. Related work

To mitigate above mentioned problem, cooperation among computations, storage and networks to adapt to the access patterns could be an effective way to minimize the user access latency, and reduce the operation cost of content providers. This adaptation may include dynamically shipping the requests as well as the related content replicas to some vantage locations in the clouds that are close to the users. Since fast provisioning of contents and VM instances has significant influence on the overall system performance, a considerable amount of research has been conducted for content placements with minimum costs in cloud-based CDNs.

The monetary cost models have evolved to include network (download/upload), storage, computation and power costs to facilitate the data and service managements in Clouds (Jiang et al., 2009; Dán, 2011). Wang et al. (2011, 2013) studied the data staging problem based on a homogeneous cost model in cloud-based CDNs for efficient mobile accesses. In their work, a variety of practical constraints with respects to the storage and communication overhead are investigated in the migration of the requested content replicas with minimum costs, which bears some similarities to our case. MetaCDN by Broberg et al. (2009) is a low cost CDN model using cloud-based storage resources. The system provides mechanisms to place contents in different cloud-based storage networks and route user requests to appropriate replicas. Peng et al. (2012) took advantage of the hierarchical network topologies of data centers to reduce VM instance provisioning time and in the meantime minimized the overhead of maintaining chunk location information. Zheng et al. (2011) presented storage migration scheduling algorithm that can greatly improve storage I/O performance during wide-area migrations. Most recently, SpotCloud by Wang et al. (2012) is a customer-provided cloud platform which enables the customers to find the best trade-off between the benefits and the costs. By exploiting social influences among users, Wu et al. (2012) proposed efficient proactive algorithms for dynamic, optimal scaling of a social media application in a geo-distributed cloud. Dai et al. (2012) discussed the collaborative hierarchical caching with dynamic request routing for massive content distributions. Chen et al. (2012) presented and evaluated a suite of online heuristic algorithms based on Integer Programming (IP). Essentially, their scheme focuses on “Cloud CDN” based on multiple storage clouds.

In most of existing work, requests are simply forwarded to the upper-layer parent servers when the content is not locally available. Or request-redirection occurs over distributed sets of servers, to minimize redirection latency (Pathan, 2009). Our work is similar to Dai et al. (2012) and Chen et al. (2012) – dynamic request routing is designed jointly with content placement strategies – and focuses on evaluating the new features of adaptation to the changes of mobile access patterns to efficiently satisfy the user requests in a cost effective way.

Because the core of cloud application is virtualization, our work in content service is necessarily built upon work in a number of other related areas: Many VM placement (suk Kee and Kesselman, 2008; Celesti et al., 2010) and data placement algorithms address distinct problems, such as replica placements and server consolidations (Vogels, 2008). Users' requests are pre-known in terms of their positions and access frequencies in the networks, while the location and the minimum number of the VMs to serve the requests with minimum total access costs are to be determined (Benoit et al., 2008; Gupta and Tang, 2006; Kalpakis et al., 2001). Liu and Datta (2011) discussed the data placements given the intermediate data for workflow applications. Mills et al. (2011) proposed an objective method for comparing VM placement algorithms in large clouds. Recently, Calcavecchia et al. (2012) proposed and evaluated a novel technique called *Backward Speculative Placement* (BSP), which analyzes the past demand behavior of a VM to a candidate target host. Biran et al. (2012) presented a VM placement algorithm, namely *Min-Cut Ratio-aware VM Placement* (MCRVMP), which satisfies the predicted communication demands and time-variations. They claimed that the general MCRVMP problem is NP-Hard and introduced several heuristics to solve it in reasonable time.

To offer both cloud service infrastructure and content delivery, a content provider is faced with the coupled application server (e.g., VM) selection, content placement and request routing problems. These problems interact with each other because the server selection and the content placement affect the operation costs while the request routing influences the offered load visible to the servers. Actually, all these factors, the server selections, content placements as well as the request routing, have great impact on our algorithm design since they not only influence the operation costs of CSPs but also affect their profitability.

[1] Given the read-only access patterns in CDN, there could be two basic strategies to achieve this goal. First, we can simply allow the source service replica continue to serve the incoming requests while migrating to the target site. When the target is ready, the source replica can be informed of being closed or removed. After the migration, the subsequent requests will be served by the target replica. Second, we can redirect the incoming requests to other service replicas when the selected content is replicated/migrated to other sites.