Computer Networks 68 (2014) 138–148

Contents lists available at ScienceDirect

Computer Networks

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

Optimal application allocation on multiple public clouds

Simon S. Woo *, Jelena Mirkovic

Computer Science Department, University of Southern California, Los Angeles, CA, USA
Information Science Institute, Marina Del Rey, CA, USA

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form 5 November 2013
Accepted 4 December 2013
Available online 5 February 2014

Keywords:
Cloud computing
Resource allocation
Hybrid cloud computing architecture
Federated cloud computing
Multiple public cloud computing

ABSTRACT

Cloud computing customers currently host all of their application components at a single cloud provider. Single-provider hosting eases maintenance tasks, but reduces resilience to failures. Recent research (Li et al., 2010) also shows that providers' offers differ greatly in performance and price, and no single provider is the best in all service categories. In this paper we investigate the benefits of allocating components of a distributed application on multiple public clouds (multi-cloud). We propose a resource allocation algorithm that minimizes the overall cloud operation cost, while satisfying required service-level agreements (SLAs). In spite of the additional delays for inter-cloud communication and the additional costs for inter-cloud data transfer, our simulation study, using real cloud performance and cost data, demonstrates that multi-cloud allocation outperforms single-cloud allocations in a variety of realistic scenarios.

1. Introduction

Cloud providers vie for customers by offering differentiated, reliable, specialized, unique, or cheaper services than their competitors. The performance of generic services, such as Web server hosting and computing, can vary greatly between providers given the same price range [1]. Such an environment makes *multi-cloud resource allocation* appealing. If providers differ greatly in their offers, hosting components of a distributed application with different providers should lead to a better price/performance trade-off than hosting them all at any single provider. In addition to cost and performance considerations, single-provider hosting lowers the reliability of an application in case of a provider-wide outage [2]. Although providers try to guarantee 99.9% availability through vertical and horizontal computing resource scaling, scalability can also be a critical issue for a single-provider approach as the number of applications and demands increase in the future. Furthermore, recent research shows that multi-cloud hosting has lower latency, as seen by end-users, than single-cloud hosting [3].

On the other hand, multi-cloud hosting adds communication delays and extra operational cost for inter-cloud data transfer. Also, the multi-cloud approach adds more complexity to application maintenance due to the lack of standard application programming interfaces (APIs) for application deployment on different clouds.

The main goal of this work is to evaluate performance and cost of multi-cloud resource allocation, and to compare these to the performance and cost of the single-provider approach for a variety of realistic cloud use scenarios. We do not assume anything about an end user's platform (e.g., fixed vs. mobile), nor how a user's application tasks are divided between their end platform and the clouds. The end-user observed delay consists of the time spent to perform tasks on one or more public clouds, and the time to transfer the data between the user's device

* Corresponding author at: Computer Science Department, University of Southern California, Los Angeles, CA, USA. Tel.: +1 310 448 9170; fax: +1 310 448 9300.

E-mail addresses: simonwoo@usc.edu (S.S. Woo), sunshine@isi.edu (J. Mirkovic).

http://dx.doi.org/10.1016/j.comnet.2013.12.001
1389-1286/© 2014 Elsevier B.V. All rights reserved.

and the clouds. Our study focuses on optimizing the first factor in this sum – the time spent to perform tasks on one or more public clouds. This factor remains constant for any end user's platform, but the division of application tasks between the user's device and the cloud, as well as delay requirements, may change. We investigate a wide range of delay requirements in our work and show that for many of them multi-cloud allocation outperforms most of single-cloud allocations. Thus, we believe our results apply to a wide range of end-user's platforms. The delay between end user and the cloud changes based on a cloud's location, thus some allocations that benefit from multi-cloud allocation approach may not be acceptable to users that are physically far away from one or more clouds. We leave consideration of this second factor in end-user observed delay for our future work. Also, mobile users are bandwidth-constrained and their applications may be distributed between their platform and the clouds in a way that minimizes data transfer between the user's device and the cloud. We believe an application developer would decide which application tasks are best performed on clouds. Our algorithm can then be used to find an optimal allocation for this subset of tasks. Also, an end user may have security requirements about which application components can be hosted on which public clouds. We leave consideration of this to our future work.

While others have proposed hybrid or federated cloud computing paradigms [4,5] and have evaluated the benefits of private/public cloud allocations [6,7], ours is the first work that extensively evaluates the benefits of application allocation on multiple public clouds. A recent publication [3] shows that multi-cloud allocation reduces the delay as seen by the end user. That research focuses only on Web service hosting and evaluates only the delay aspect of resource allocation. We, on the other hand, focus on several popular types of cloud applications and evaluate both the delay and the cost of multi-cloud and single-cloud allocations.

Our contributions are:

1. We propose a novel cloud resource allocation algorithm that determines the best allocation of application components over multiple public cloud providers, under a given performance constraint.
2. We show through extensive evaluation, which relies on realistic benchmark data about cloud performance and price, as described in [1], that multi-cloud allocation outperforms single-cloud allocation in a variety of realistic cloud use scenarios. We also investigate the cloud and application features that best bring out the benefits of multi-cloud resource allocation.

2. Related work

A hybrid or a federated cloud is composed of two or more private, community, and/or public clouds which work together to achieve the application objective, while each cloud remains as a unique entity [8]. Typically, in a hybrid cloud, the organization manages and uses in-house computing resources as well as external cloud resources. A multi-cloud is a specific form of the hybrid cloud where all clouds are public clouds.

Migrating parts of the applications to a cloud has been addressed in [9]. Also, [10] provides an approach to dynamically split transactions in different data-centers. However, [9,10] assume allocation on a single cloud provider. Publications [4,5] point out the potential scaling problem of single-provider hosting as the number of users and applications increase, and, thus, champion federation among different cloud providers. These publications, however, focus only on reliability and do not evaluate the potential performance and cost-saving benefits of federated allocation.

Publications [11,12] similarly focus on overcoming inter-cloud interoperability, workload distribution, and inter-cloud policy issues of hybrid or federated cloud hosting. Publications [7,13] considered the application of the public/private clouds to improve the cost and response time of applications. However, they only consider the use of compute machines and private/public cloud allocation. On the other hand, we consider more cloud services (compute, storage and DB resources) and allocation on multiple public clouds. Also, we focus on the evaluation of performance and cost benefits, and we provide a side-by-side comparison between a single-cloud and multiple public clouds approach.

A recent publication [3] shows that multi-cloud allocation reduces the delay as seen by the end user, but focuses only on Web service hosting and delay benefits. We, on the other hand, investigate more cloud applications, and we evaluate both the performance and the cost of each allocation scenario.

3. Multi-cloud resource allocation

While cloud providers try to differentiate themselves from their competitors by offering some unique, specialized services and APIs, most clouds offer generic compute, database (DB), and storage services and have a similar price structure for them (i.e., some combination of flat rate and per usage cost). In [1], the authors benchmarked these generic services over four popular cloud providers and had two major findings:

1. No single provider offered the best performance in all three service categories. For example, one provider offered very fast computing resources but slower storage and no DB service. Another offered fast DB service and fast binary large object (blob) storage (for some blob sizes), but its compute machines were much slower than its competitors'.
2. There was a large difference in price between providers for the same level of performance for a given service.

If a cloud application needs resources from several of these generic services, the differences in performance and price across providers make a multi-cloud resource allocation an attractive choice. Depending on the exact resources being allocated and the application workload, such multi-cloud allocation has the potential to outperform