Research Report

Data Center Selection for QoS-Aware Service Provision

CLOUDS ARE INCREASINGLY BEING USED FOR THE DELIVERY OF COMPLEX SOFTWARE SERVICES WITH STRINGENT QUALITY OF SERVICE (QOS) DEMANDS. IN ORDER TO ACHIEVE COST-EFFICIENT PROVISION THROUGH THE UNDERLYING INFRASTRUCTURE, A SUITABLE SELECTION OF CLOUD DATA CENTERS IS REQUIRED. THIS REPORT PRESENTS A CORRESPONDING OPTIMIZATION APPROACH, WHICH IS SUITABLE FOR BOTH PRIVATE AND PUBLIC CLOUD SETTINGS.

Ulrich Lampe

Ronny Hans

Ralf Steinmetz

Introduction

For the financial services sector as "heavy user" of information technology, the application of cloud computing offers multiple potential benefits, most notably cost-savings due to consolidation and centralization of resources (e.g., Creeger, 2009). The historic root and traditional purpose of cloud computing has been to provide basic infrastructure services, such as virtual machines or storage. Recently, however, more sophisticated software services and applications, such as *Desktop as a Service*, are also being delivered by the cloud.

Such applications are characterized by relatively stringent Quality of Service (QoS) requirements, e.g., with respect to latency or availability, in order to provide adequate Quality of Experience (QoE) for the end user. Due to these requirements, the existing cloud infrastructure – which is primarily driven by cost considerations and commonly makes very limited QoS assurances – appears insufficient.

In our research work, we examine how clouds can be exploited to deliver software services in a QoS-aware, yet still cost-efficient manner. In this context, we address two distinct, but similar questions:

1. How to select data centers for construction among a set of potential locations at *design time*?
2. How to assign users and services to existing data centers at *run time*?

Both questions ultimately aim at an optimal selection of (potential or existing) cloud data centers and (permanent or temporary) assignment to users with specific QoS and resource demands. Hence, the overarching research problem is coined as *Cloud Data Center Selection Problem*. It does not only concern large public cloud providers, but also companies which aim to provide a cost-efficient private cloud infrastructure to their internal customers.

The solution approach that is proposed in this report is part of a more comprehensive effort that will ultimately facilitate the cost-efficient exploitation of cloud resources as part of E-Finance processes.

Optimization Approach for Data Center Selection

Our solution approach to the Cloud Data Center Selection Problem is based on the formulation of a mathematical model, which can subsequently be solved by off-the-shelf solver frameworks in order to obtain an exact, i.e., optimal, solution.

For the initial model, we assume that the (private) cloud provider considers a preselected set of geographically distributed data centers. Each data center may provide resources within a certain predefined range, which is, e.g., given by size constraints. Furthermore, each data center is characterized by fixed costs, e.g., for construction or lease, and variable costs depending on the resource use, e.g., required server units. Using these parameters, both private and public data centers can be modeled, depending on the specific cost model.

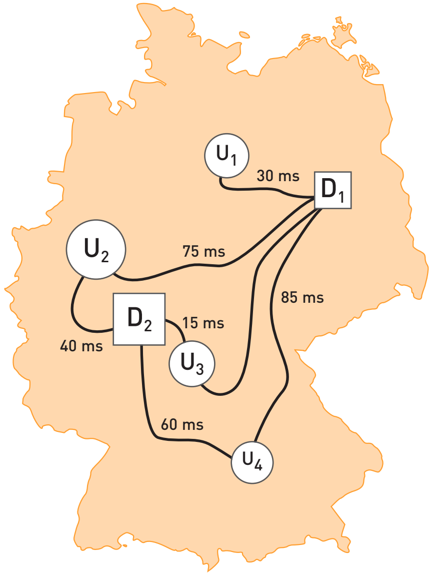

Figure 1: Simplified example of a Cloud Data Center Selection Problem

The data centers should ultimately serve a set of user clusters. Each user cluster corresponds to a more or less fine-granular representation of a set of users. These user clusters exhibit specific resource demands and QoS requirements. Accordingly, the data centers make certain QoS guarantees for each user cluster, e.g., depending on the geographical distance and network topology.

A tangible example is given in Figure 1, where a company aims to serve four major user clusters (U_1 to U_4) using two data centers (D_1 and D_2).

Each user cluster and data center exhibits different resource demands and maximum supplies, as indicated by the respective symbol sizes. Furthermore, the latencies of the different links differ depending on the network topology. Thus, some data centers may be unsuitable to serve certain users with the desired service type.

The objective of the optimization approach consists in cost-minimization under the constraint that all user service demands are satisfied and that the corresponding QoS requirements are fulfilled. While this formulation only targets costs, non-monetary objectives, such as average QoS properties, may also be considered due to the flexible design of the approach.

For further details, we refer the interested reader to our recent publication (Hans et al., 2013), which contains additional formal specifications as well as the complete optimization model.

Evaluation Results

In order to assess the practical applicability of our proposed solution approach, we have conducted an extensive quantitative evaluation. The primary goal of this evaluation was to assess whether the computational complexity of the algorithm permits an application to problem instances of practically relevant size. For that purpose, we created 12 test cases, each involving a different number of data centers and user clusters. For each test case, 50 problem instances were created and solved with our approach. Subsequently, we computed the mean computation time for each test case across all problem instances.

In order to realistically model the problem instances, we used data from the 2010 United States Census as a basis (U.S. Census Bureau, 2013). Data center and user cluster specifications were then generated based on properties such as median incomes, population densities, and geographical distances.

The results of the quantitative evaluation are provided in Figure 2. As can be seen, the mean computation times quickly increase with the problem size, namely the number of considered data centers and user clusters. For the largest test cases, which involve 40 data centers and up to 600 user clusters, average computation times in the order of magnitude of minutes are observed.

This indicates that our proposed solution approach is well-applicable for long-term data center selection at design time, e.g., in early project stages, where sufficient time is available for the decision process. However, for short-term decisions during runtime (e.g., to handle peak service demands), the use of more efficient heuristic approaches appears favorable.

Conclusions and Outlook

Cloud computing offers the potential for cost savings, whether applied internally in the form of a private cloud or through the lease of external resources from a public cloud. However, with more and increasingly complex applications being provided by clouds, complex QoS requirements have to be satisfied.

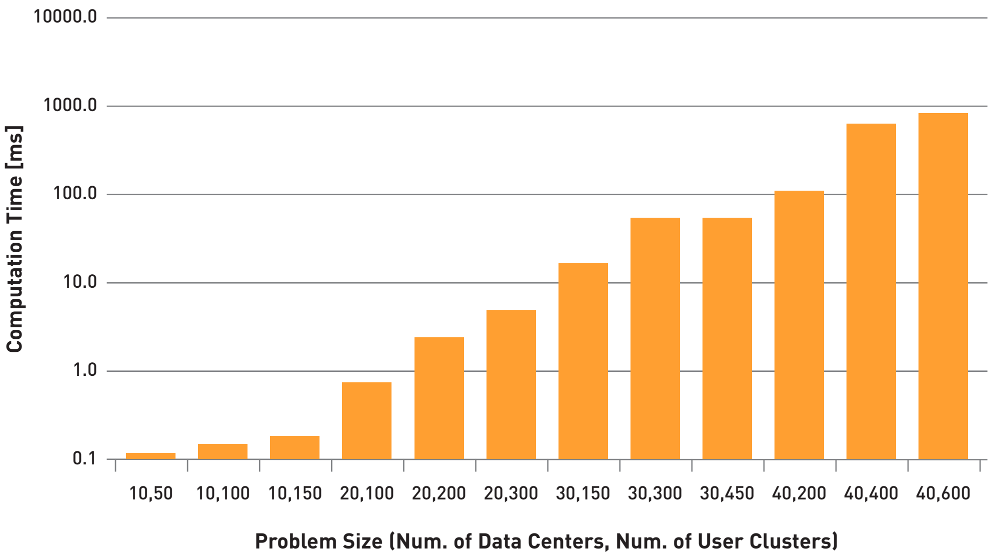

Figure 2: Mean computation times for optimally solving Cloud Data Center Selection Problems of different sizes. Sample size n = 50 per problem size

In this report, we introduced an initial optimization approach for cost-efficient, QoS-aware data center selection, which can be applied in various application scenarios. As our evaluation has shown, the approach is well suited for the long-term selection of cloud infrastructures.

Given the relatively high computational complexity, we are currently working on heuristic solution approaches that are also applicable under stringent time constraints at run time, and thus permit to dynamically assign end users to data centers.

References

Creeger, M.:
Cloud Computing: An Overview.
In: ACM Queue, 7 (2009) 5, pp. 1-5.

Hans, R.; Lampe, U.; Steinmetz, R.:
QoS-Aware, Cost-Efficient Selection of Cloud Data Centers.
In: Proceedings of the 6th International Conference on Cloud Computing, Santa Clara, CA, United States, 2013.

U.S. Census Bureau:
Gazetteer Files – Geography – U.S. Census Bureau.
http://www.census.gov/geo/maps-data/data/gazetteer.html, 2013.