Ronny Hans, Ulrich Lampe, Ralf Steinmetz: QoS-Aware, Cost-Efficient Selection of Cloud Data Centers. In: Proceedings of the 6th International Conference on Cloud Computing (CLOUD 2013), p. 946-947, Institute of Electrical and Electronics Engineers (IEEE), June 2013. ISBN 978-0-7685-5028-2.

QoS-Aware, Cost-Efficient Selection of Cloud Data Centers

Ronny Hans, Ulrich Lampe, and Ralf Steinmetz
Multimedia Communications Lab (KOM), Technische Universität Darmstadt, Darmstadt, Germany
Email: {firstName.lastName}@KOM.tu-darmstadt.de

***Abstract*—Cloud computing is increasingly used to deliver multimedia services with stringent Quality of Service (QoS) requirements, rather than simple infrastructure. Due to these requirements, the QoS-aware, cost-efficient selection of data centers arises as a new research challenge. In our ongoing work, we examine the corresponding *Cloud Data Center Selection Problem* and propose an exact optimization approach. A brief evaluation indicates that the proposed approach is primarily suited for small problem instances due to its high computational complexity, and hence highlights the need for the development of heuristic optimization approaches.**

***Keywords*-cloud computing; multimedia; quality of service; cost; data center; location; selection**

I. Introduction

Cloud computing is increasingly used for the delivery of sophisticated multimedia services today, rather than just simple infrastructure services. A prominent example of such multimedia services is cloud gaming, where video games are centrally executed in a cloud data center and delivered to the user as audio/video stream [1]. Unfortunately, multimedia services also pose high Quality of Services (QoS) requirements, e. g., with respect to latency, which renders the selection of data centers challenging and potentially cost-prone [2]. Based on this notion, we examine the QoS-aware, cost-efficient selection of cloud data centers in our ongoing work. In this context, selection may either refer to choosing among *potential* data centers for construction at *design time*, or choosing among *existing* data centers for service delivery at *run time*. Due to the structural similarity of these two challenges, they can be addressed in an identical manner.

In the following Section II, we explain the problem in further detail and introduce formal notations. Section III presents an exact optimization approach, which is briefly evaluated in Section IV. Conclusions and an outlook on future work are given in Section V.

II. Problem Statement

We assume that a cloud provider considers a set of (potential or existing) data centers, $D \subset \mathbb{N}$, in different physical locations. The prospective data centers should serve a set of user clusters, $U \subset \mathbb{N}$, where every cluster represents a set of individual service clients in a certain geographical area, such as a state, county, or city. The provider further defines a set of relevant QoS attributes $Q \subset \mathbb{N}$.

Each user cluster $u \in U$ exhibits a certain service demand, $S_u \in \mathbb{N}$, which can be expressed in a standardized resource unit, e. g., servers. Furthermore, the cluster has certain QoS requirements, $QR_{u,q} \in \mathbb{R}^+$, which are expressed with respect to each QoS attribute $q \in Q$. Without loss of generality, we assume that QoS requirements are expressed as upper bound, e. g., maximum latency.

Inversely, each data center $d \in D$ may provide between $K_d^{min} \in \mathbb{N}$ and $K_d^{max} \in \mathbb{N}$ resource units. It makes a QoS guarantee $QR_{d,u,q} \in \mathbb{R}^+$ with respect to the user cluster u and QoS attribute q. If a certain data center is selected, fixed costs $CF_d \in \mathbb{R}^+$ apply, either for its construction and future operation or its lease. In addition, each provisioned resource unit results in additional variable cost $CV_d \in \mathbb{R}^+$.

The problem of the provider consists in cost-minimally selecting among the potential data centers, as well as setting the respective resource capacity and allocation to different user clusters for each data center. This process is subject to the constraints that the service demands of all user clusters are met and the QoS requirements are matched by corresponding guarantees. We refer to this problem as *Cloud Data Center Selection Problem* (CDCSP).

III. Optimization Approach CDCSP-EXA.KOM

In order to provide an exact, i. e., optimal solution to the CDCSP, we map it into a mathematical optimization model. The result is provided in Model 1.

In the model, Eq. 7 defines the decision variables: x_d are binary variables, which indicate whether data center d will be constructed respectively used or not. $y_{d,u}$ are integer variables that denote how many resource units data center d provides to user cluster u in order to satisfy its service demand. Depending on the decision variables, the total cost C is determined in the objective function in Eq. 1.

Eq. 2 represents the constraint that the service demands of all user clusters must be satisfied by corresponding data center capacities. Eqs. 3 and 4 functionally link the decision variables x and y and also assure that the capacity of each data center is chosen from the specified interval, i. e., K_d^{min} to K_d^{max}. Eq. 5 constrains the assignment between data centers and user clusters, depending on the variables $p_{d,u}$ from Eq. 6, which indicate whether the QoS requirements of a user cluster u are met by data center d or not.

As can easily be seen, Model 1 constitutes an *Integer Linear Program* (ILP), which can be solved using off-the-shelf solver frameworks. The corresponding optimization approach is referred to as *CDCSP-EXA.KOM*.

Model 1 Cloud Data Center Selection Problem

$$\text{Min. } C(x,y) = \sum_{d \in D} x_d \times CF_d + \sum_{d \in D, u \in U} y_{d,u} \times CV_d \tag{1}$$

$$\sum_{d \in D} y_{d,u} \geq S_u \quad \forall u \in U \tag{2}$$

$$\sum_{u \in U} y_{d,u} \leq x_d \times K_d^{max} \quad \forall d \in D \tag{3}$$

$$\sum_{u \in U} y_{d,u} \geq x_d \times K_d^{min} \quad \forall d \in D \tag{4}$$

$$y_{d,u} \leq p_{d,u} \times K_d^{max} \quad \forall d \in D, \forall u \in U \tag{5}$$

$$p_{d,u} = \begin{cases} 1 & \text{if } QG_{d,u,q} \leq QR_{u,q} \quad \forall q \in Q \\ 0 & \text{else} \end{cases} \tag{6}$$

$$x_d \in \{0,1\} \quad \forall d \in D$$
$$y_{d,u} \in \mathbb{N} \quad \forall d \in D, \forall u \in U \tag{7}$$

IV. Evaluation

In order to assess the performance of our proposed optimization approach CDCSP-EXA.KOM, we have prototypically implemented it in Java. As solver, we employ the commercial IBM ILOG CPLEX framework[1]. The focus of our evaluation is on the required computation time as a determinant of the practical applicability of the algorithm.

For that purpose, we created 12 different test cases with a predefined number of data centers ($|D|$) and user clusters ($|U|$), respectively. Each test case involved 50 problems that were randomly generated, based on actual data from the 2010 United States census[2]. We solved each problem on a desktop computer, equipped with an Intel Core 2 Quad Q9450 processor and 4 GB of memory, operating under Microsoft Windows 7, and measured the required computation time.

Figure 1 provides the results of our evaluation. As can be seen, the mean computation times quickly increase with the number of data centers and user clusters. The effect appears slightly more pronounced for the first variable, i. e., the number of data centers. Despite the relatively small problem sizes in our evaluation, we observed computation times of up to 180 minutes for individual instances. For the last three test cases (with $|D| = 40$), even the *mean* computation times reach the order of magnitude of minutes.

Thus, in order to permit for a very fine-grained representation of users – which implies using tens or hundreds of thousands of user clusters in the model – and a large set of potential data centers, the development or adaptation of a heuristic adaptation approach may be beneficial. This specifically applies if the approach is to be used at run time for the allocation of user clusters to data centers, rather than the placement of the latter at design time (cf. Section I).

[1]http://www.ibm.com/software/integration/optimization/cplex-optimizer/

[2]http://www.census.gov/geo/maps-data/data/gazetteer.html

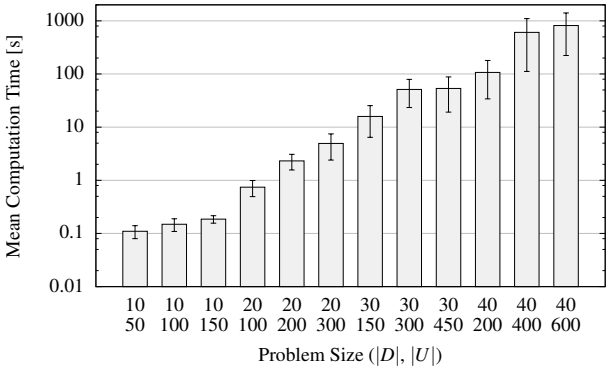

Figure 1: Mean computation times of CDCSP-EXA.KOM, depending on the problem size (with 95% confidence intervals; sample size $n = 50$). Please note the logarithmic scale of the ordinate.

V. Conclusions and Outlook

The QoS-aware, cost-efficient selection of data centers is an important challenge that arises with the increasing delivery of multimedia services through the cloud. In this work, we have introduced the *Cloud Data Center Selection Problem* (CDCSP) and proposed an exact optimization approach, based on Integer Linear Programming. A preliminary evaluation indicated high computational requirements for solving larger problem instances that are of practical relevance.

Hence, the primary focus of our future work will be the development of heuristic optimization approaches. Furthermore, we strive to incorporate additional parameters – such as fluctuating service demands, different service classes, or data center redundancy – into the model.

Acknowledgments

This work has partly been sponsored by the E-Finance Lab e. V., Frankfurt a.M., Germany (www.efinancelab.de).

References

[1] U. Lampe, Q. Wu, R. Hans, A. Miede, and R. Steinmetz, "To Frag Or To Be Fragged - An Empirical Assessment of Latency in Cloud Gaming," in *Proc. of CLOSER 2013*, 2013.

[2] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "The Brewing Storm in Cloud Gaming: A Measurement Study on Cloud to End-User Latency," in *Proc. of NetGames 2012*, 2012.