Ronny Hans: Selecting Cloud Data Centers for QoS-Aware Multimedia Applications. In: Wolf Zimmermann: Proceedings of the PhD Symposium at the 2nd European Conference on Service-Oriented and Cloud Computing, p. 11-16, University Halle-Wittenberg, September 2013.

Selecting Cloud Data Centers for QoS-Aware Multimedia Applications

Ronny Hans (Advisor: Prof Dr.-Ing. Ralf Steinmetz)

Multimedia Communications Lab (KOM), Technische Universität Darmstadt,
Rundeturmstr. 10, 64283 Darmstadt, Germany

Abstract. Cloud computing infrastructures are increasingly used to deliver sophisticated multimedia services. Since these services commonly pose stringent Quality of Service (QoS) requirements, the appropriate selection of data centers arises as a new research challenge. The corresponding *Cloud Data Center Selection Problem* is addressed in my work. In this paper, I provide a state of the art overview, initial research results, and an extensive outlook on future extensions.

Keywords: cloud computing; multimedia; quality of service; cost; data center; selection

1 Introduction

For many years, cloud computing has been used to deliver Information Technology (IT) services in countless application scenarios. Today, the delivery of sophisticated multimedia services increasingly gains in importance. A popular example of such multimedia services is cloud gaming. Thereby, video games are executed in the data centers of cloud providers and the content is delivered as audio/video stream via the Internet [1]. In such context, the fulfillment of stringent Quality of Services (QoS) requirements plays an outstanding role. For example, latency – which determines the quality of experience in video gaming – highly depends on the selection of appropriate data centers that are located geographically close to the users [2].

Nowadays, most cloud services are provisioned by a few, centralized data centers around the globe. The locations of these data centers are selected with the aim to minimize costs [2]. Due to these facts, the current cloud infrastructure is hardly able to provide multimedia services with stringent QoS requirements [2]. Hence, to ensure and improve the provisioning of multimedia software services, two basic questions arise: (1) How to design future cloud infrastructures, i. e., where to place new data centers? (2) How to distribute resources of existing data centers to offer QoS-sensitive software services to a maximum number of users?

Thus, the selection process 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*. Both problems are closely related and essentially map onto a similar research problem, which we have previously introduced as

Cloud Data Center Selection Problem (CDCSP) [3]. The aim of my work consists in the development of corresponding optimization approaches, which permit to address the CDSCP and hence allow for a cost-efficient, QoS-aware selection of data centers both a design time and run time.

The remainder of this paper is structured as follows: In Section 2, the current state of research is presented. Section 3 outlines initial optimization approaches and gives preliminary evaluation results. Future research directions are discussed in Section 4. Section 5 concludes the paper with a brief summary.

2 Current State of Research

In my work I address the cost-efficient selection of data centers for multimedia applications during *design time* and *run time*. Related issues have been addressed by other researchers in the past.

To start with, Goiri et al. [4] present an approach for efficient data center placement. Thereby, a location is determined by several factors, e. g., network backbones, cost of electric energy, and proximity to potential customers. The authors used a combination of optimal approaches and heuristics to find a solution. While Goiri et al. focus on the placement of new data centers at design time, my work additionally aims to provide solutions for appropriate data center resource distribution at run time.

Larumbe and Sans [5] present an optimization approach which addresses three distinct, yet interlinked problems: First, the authors address the geographical location of data centers. Second, they address the location of software components that are hosted in network nodes. Finally, they investigate the issue of routing. Because the authors see a close connection between these problems, they integrated it in one mathematical framework using an optimal approach. Due to the chosen approach, which uses Integer Programming, the algorithm tends to be more suitable for design time, whereas my work covers design time and run time.

Choy et al. [2] study the network delay of the existing Amazon EC2 cloud infrastructure. The authors show that the existing small numbers of large scale data centers are only able to meet latency requirements of multimedia applications for fewer than 70% of the US population. The authors propose to augment existing data centers by specialized servers, so called edge servers, located nearby end users. They claim that their proposal would allow 90% coverage in the U.S. However, they do not propose an optimization approach to decide on the placement or selection of these data centers and servers.

Wang et al. [6] identify several cloud gaming challenges: Low round-trip latency, high bandwidth for video streaming, and high computation needs for servers, saying that these challenges could result in high costs. The authors propose an approach to schedule computing and network resources simultaneously. Thereby, they assume a dynamically changing resource demand. However, the proposed algorithm makes decisions for new requests only and does not try to find a solution for all uses at the same time. The focus on run time is also a major difference to my work.

3 Initial Approach and Results

In the following, I describe my preliminary research results. These results include an exact (optimal) and a heuristic (non-optimal) solution approach, as well as initial evaluation results for those. As a basis, I will first introduce a set of formal notations.

3.1 Formal Notations

I assume that the cloud provider considers a set of (potential or existing) geographically distributed data centers, $D = \{1, 2, ..., D^{\#}\} \subset \mathbb{N}$, for selection. These data centers should serve a set of user clusters $U = \{1, 2, ..., U^{\#}\} \subset \mathbb{N}$. Such user clusters are representations of a number of clients, which are located in certain geographical areas. Further, a provider defines a set of relevant QoS attributes $Q = \{1, 2, ..., Q^{\#}\} \subset \mathbb{N}$. Each user cluster $u \in U$ has a specific demands of services, $S_u \in \mathbb{N}$, expressed in, e. g., server units. Additionally, for the provided services, each cluster expects certain QoS requirements, $QR_{u,q} \in \mathbb{R}$, for each specific QoS attribute $q \in Q$. Without loss of generality, the QoS requirements are expressed as as upper bound, e. g., maximum latency; lower bounds can be expressed through negation.

Each data center $d \in D$ may provide server units within a minimal and maximal boundary, $K_d^{min} \in \mathbb{N}$ and $K_d^{max} \in \mathbb{N}$. Regarding the user cluster u and QoS attribute q, a data center makes a QoS guarantee $QR_{d,u,q} \in \mathbb{R}^+$, depending, e. g., on the network topology and distance. Each selected data centers results in is assumed to result in fixed cost $CF_d \in \mathbb{R}^+$ and variable $CV_d \in \mathbb{R}^+$ per provisioned server unit.

The challenge for the provider is the cost-minimal selection of data centers, such that each user cluster is served its service demand at minimal cost under the given QoS constraints.

3.2 Exact Optimization Approach CDSCP-EXA.KOM

An exact solution for the CDCSP, based on Integer Programming (IP), is provided Model 1. This approach – which has been proposed in past research [3] – is referred to as CDSCP-EXA.KOM in the following.

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.

Model 1 Cloud Data Center Selection Problem

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

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

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

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

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

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

$$x_d \in \{0,1\} \quad \forall d \in D$$

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

. . . . . .

$$x_d \in \mathbb{R}, 0 \leq x_d \leq 1 \quad \forall d \in D$$

$$y_{d,u} \in \mathbb{R}, y_{d,u} \geq 0 \quad \forall d \in D, \forall u \in U \tag{8}$$

3.3 Heuristic Optimization Approach CDSCP-REL.KOM

As explained before, Model 1 constitutes an IP. This model can be solved using off-the-shelf solver frameworks, most notably using the *branch-and-bound* algorithm. Unfortunately, this algorithm is based on the principle of (intelligently) enumerating the solution space and thus features worst-case exponential time complexity. Accordingly, the past research has shown that solving larger problem instances may result in computation times in the order of magnitude of hours, which renders the approach unsuitable for application at run time and highlights the need for a heuristic approach.

Hence, as an initial measure, I propose the application of Linear Program (LP) relaxation [7] to the IP formulation. The corresponding approach is referred to as *CDCSP-REL.KOM*. Through the relaxation, the decision variables are defined as real, rather than binary and integer numbers, i. e., Eq. 7 is replaced by Eq. 8. The resulting problem can be solved using commonly much more efficient methods, e. g., the *Simplex* algorithm or *interior point* approaches [8]. However, the relaxed formulation may result in suboptimal solutions, thus trading reduced computation time for higher cost of the solution.

3.4 Implementation and Preliminary Evaluation

In order to assess the performance of the proposed optimization approaches CDCSP-EXA.KOM and CDCSP-REL.KOM, I have prototypically implemented

them in Java. As solver, I employ the commercial IBM ILOG CPLEX framework. The focus of the evaluation is on the required computation time and the solution quality, i. e., total cost, and the tradeoff between these two factors. For the evaluation, I created six different test cases with a predefined number of data centers ($|D|$) and user clusters ($|U|$), respectively. Latency was considered as sole QoS attribute. Each test case involved 100 problems that were randomly generated, based on actual data from the 2010 United States census[1]. The evaluation was conducted using a dedicated laptop computer, equipped with an Intel Core i5-450M processor and 2 GB of memory, operating under Windows 7.

Table 1 provides the results of my evaluation. As can be seen, the heuristic approach CDCSP-REL.KOM has substantial benefits over the exact approach CDCSP-EXA.KOM with respect to the absolute computation time, specifically for larger problem instances. This is also confirmed by the macro-averaged ratio, which indicates reductions of up to 99.3%. This reduction is traded against a moderate increase in cost, which is up to about 11% for the smaller problem classes, but shrinks with increasing problem size. Hence, CDCSP-REL.KOM appears as a promising non-exact solution approach to the CDCSP, specifically for application at run time under stringent time constraints.

Table 1: Evaluation results, with 95% confidence intervals in parentheses. Ratios were computed using both macro- and micro-average and use CDCSP-EXA.KOM as baseline.

Test case	Abs. Comp. Time [ms]		Rel. Comp. Time			Rel. Cost		
$\|D\|, \|U\|$	EXA	REL	Macro		Micro	Macro		Micro
10, 50	165.0	13.5	11.6%	(2.7%)	8.2%	110.8%	(9.8%)	108.3%
10, 100	182.3	25.1	18.8%	(2.0%)	13.8%	108.7%	(5.2%)	107.3%
20, 100	1017.7	54.2	19.4%	(3.5%)	5.3%	103.9%	(1.5%)	103.7%
20, 200	2723.7	140.5	22.0%	(4.7%)	5.2%	103.2%	(0.7%)	103.1%
30, 150	14530.0	150.4	14.6%	(3.5%)	1.0%	102.0%	(0.5%)	101.8%
30, 300	63809.7	465.1	15.9%	(4.3%)	0.7%	102.3%	(0.3%)	102.3%

4 Future Research Directions

The approaches and corresponding results assume a scenario that involves deterministic data and a set of predefined QoS requirements. Furthermore, resources are represented in a coarse-granular form using server units, rather than individual resource types such as CPU power or bandwidth. This scenario and the corresponding optimization approaches could be extended in the future through the consideration of the following aspects: *Individual resource types:* The current approaches based on the assumption that the resource allocation take place in terms of server units. While this model may constitute a good approximation for

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

many application scenarios, I additionally plan to consider individual resource types such CPU, GPU, or memory in the future. *Stochastic parameters:* In future approaches, I plan to drop the assumption that service demands and QoS properties are precisely known in advance. For that purpose, I will adapt the model to permit for stochastic parameters as input. *Functional/non-functional correlation:* The current approaches treat service demands and QoS properties as independent. In the future, I plan to further investigate the correlation and potential tradeoff between these factors, e. g., potential reductions in latency through changes in data center resource usage.

5 Conclusions

The cost-efficient selection of cloud data centers for the provision of multimedia services is an important challenge. For the resulting *Cloud Data Center Selection Problem* (CDCSP), I proposed an exact optimization approach, CDCSP-EXA.KOM, based on Integer Programming and a heuristic optimization approach, CDCSP-REL.KOM, that uses Linear Programming relaxation. A preliminary evaluation indicated high computational requirements for solving larger problem instances using the exact approach. Using the heuristic, computation time can reduced by up to 99.3% with moderate cost increases of about 11%, hence providing a first viable to scenarios where a selection at run time is required.

My future work will focus on the extensions that were outlined in Section 4, the development of additional heuristic approaches, and a more extensive evaluation.

Acknowledgments

This work was partially supported by the Commission of the European Union within the ADVENTURE FP7-ICT project (Grant agreement no. 285220) and 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," in *CLOSER*, 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 *NetGames*, 2012.
3. R. Hans, U. Lampe, and R. Steinmetz, "QoS-Aware, Cost-Efficient Selection of Cloud Data Centers," in *CLOUD*, 2013.
4. I. Goiri, K. Le, J. Guitart, J. Torres, and R. Bianchini, "Intelligent Placement of Datacenters for Internet Services," in *ICDCS*, 2011.
5. F. Larumbe and B. Sansò, "Optimal Location of Data Centers and Software Components in Cloud Computing Network Design," in *CCGrid*, 2012.
6. S. Wang, Y. Liu, and S. Dey, "Wireless Network Aware Cloud Scheduler for Scalable Cloud Mobile Gaming," in *ICC*, 2012.
7. W. Domschke and A. Drexl, *Einführung in Operations Research*. Springer, 2004.
8. F. Hillier and G. Lieberman, *Introduction to Operations Research*. McGraw-Hill, 2005.