Research Report

Heuristic Approaches for QoS-Aware Cloud Data Center Selection

OVER THE LAST DECADE, IT PROVISIONING VIA CLOUDS HAS BECOME A COMMON PRACTICE. MEANWHILE, COMPLEX SOFTWARE SERVICES WITH STRINGENT QUALITY OF SERVICE (QOS) REQUIREMENTS ARE DELIVERED BY CLOUD PROVIDERS OVER THE INTERNET. TO ACHIEVE A COST-EFFICIENT AND QOS-AWARE SERVICE PROVISIONING, THE SELECTION OF APPROPRIATE CLOUD RESOURCES IS A HIGHLY IMPORTANT TASK. IN THIS REPORT, WE OFFER CONCEPTS AND TOOLS TO SERVICE PROVIDERS FOR AN ACCELERATED RESOURCE SELECTION IN LARGE ENVIRONMENTS.

Ronny Hans

Ralf Steinmetz

Introduction

Information Technology (IT) can both be seen as a major enabler and a major expense for service provisioning in the financial industry. Because of the fierce and continuously growing competition, e.g., through FinTechs, cost savings remain mandatory. Besides the cost, flexibility, scalability, and a high service quality are further requirements for IT systems.

In terms of cost, flexibility, and scalability, cloud computing may be a promising substrate to provide IT services. Over the last decade, cloud computing has become a key paradigm for the provisioning of IT services. The seminal idea was to provide elastic infrastructure resources in order to enable users to adapt their demand to usage cycles and load surges. Today, the requirements go beyond merely supplying resources to applications with high-quality requirements, i.e., Quality of Service (QoS) constraints.

Cost-savings in cloud computing are accomplished by consolidation and centralization of resources (Creeger, 2009) with the consequence of high latencies. As a consequence, when using the public cloud infrastructure, providers are only partly able to provide software services with rigid latency constraints (Choy et al., 2012).

Thus, provisioning of only cost-driven cloud infrastructures appears inadequate for sophisticated and highly interactive applications. In order to optimize future or existing cloud infrastructures to software service providers, we address the following research questions:

1. How to efficiently plan the utilization of new and of existing resources in cloud infrastructures?
2. How to determine reliable approaches that improve and guarantee the quality of the solution to the resource assignment problem?

Optimization Approach for Data Center Selection

In our model, we consider a (private) cloud provider which aims to choose among a given number of geographically distributed cloud resources, i.e., data centers. Here, each data center provides different amounts of various resources, which results in different types of costs – fixed and variable costs. The provided resources are characterized by QoS guarantees.

The data centers provide their services to user clusters, which represent a group of users in a certain area. These user clusters are characterized by a specific demand and certain QoS requirements, e.g., latency requirements for specific services.

A basic example is given in Figure 1. Herein, a (private) cloud provider aims to serve four user clusters (U1 to U4) through its data centers (D1 and D2). The different sizes of the symbols refer to the particular resource demand of each user cluster and the resource supply of the data centers, respectively. Furthermore, the respective latencies are denoted at the connecting edges and differ depending on the network topology.

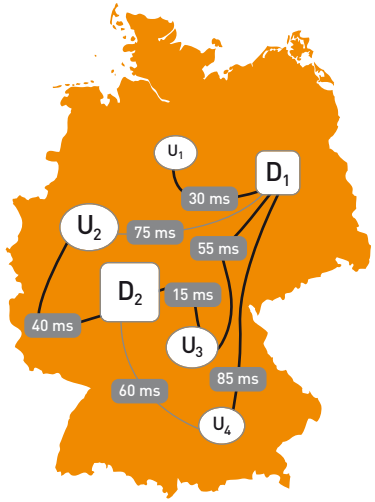

Figure 1: Simplified Example of a Cloud Data Center Selection Problem

The optimization problem consists in the fact of minimizing the costs for selected resources while meeting the QoS constraints of the clients.

The corresponding mathematical model can be solved by off-the-shelf solver frameworks (Hans et al., 2013). However, in the worst case, the computation time of such integer programs grows exponentially. For large environments, such an approach is hardly feasible, even if it delivers the optimal result for a given

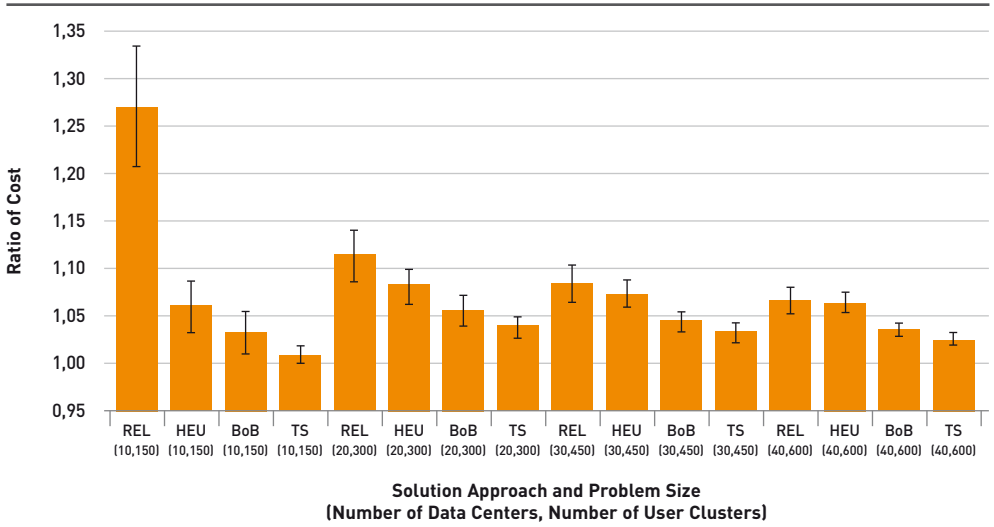

Figure 2: Ratio of Costs (Based on Macro-Average; with 95% Confidence Intervals) Between the Exact Approach and the Heuristic Approaches by Heuristic Approach and Test Case (Sample Size n = 100 per Test Case)

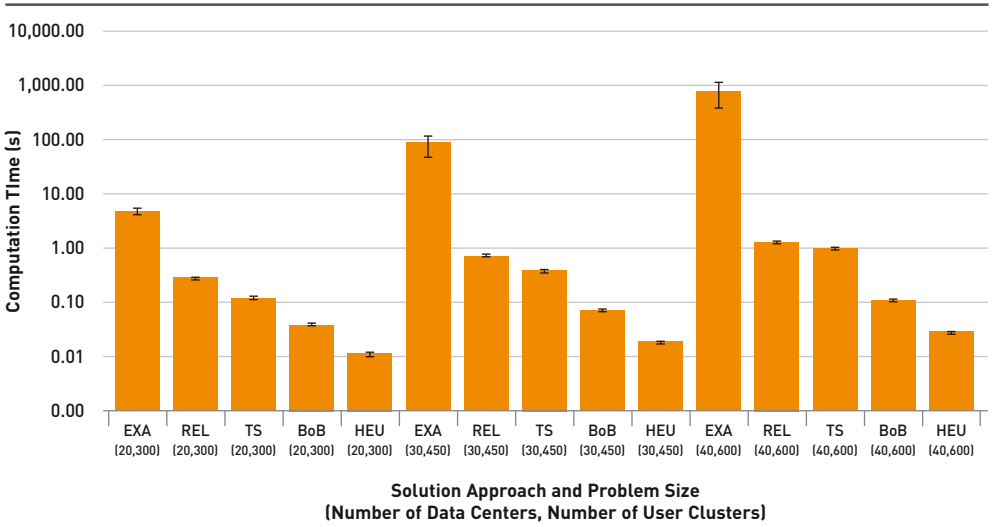

Figure 3: Observed Mean Computation Times (with 95% Confidence Intervals) by Heuristic Approach and Test Case (Sample Size n = 100 per Test Case; Please Note the Logarithmic Scaling of the Ordinate)

problem. To overcome this issue, we developed different heuristic approaches and evaluate their suitability.

Heuristic Approaches

In general, heuristics trade solution quality against performance. Thus, an increase in performance usually happens on the expense of the solution quality. Further, for increased efficiency, heuristics need to be developed or adapted subject to the given problem and the application scenario.

As an initial approach to solve the problem described earlier, we introduced a relaxed version of the model that can be solved using a linear program (LP). This heuristic approach quickly delivers less accurate solutions by relaxing some given constraints (Hans, 2013). The advantage of the approach is its simplicity and, again, the possibility to use off-the-shelf solver frameworks. But this simplicity is at the same time a major drawback. It is a very general solution and ignores the specific structure of the problem. Using specifically developed or adapted approaches, substantial improvements in both solution quality and performance can be achieved.

A very good performance, i.e., a very low computation time, is delivered by simple heuristics, such as greedy algorithms. In our research, we use priority-driven heuristics to find valid solutions for the optimization problem described earlier. Therefore, we evaluate various priority and cost allocation roles for an efficient resource assignment. Combining these rules in our priority-based framework, we are able to generate numerous heuristics, each with different solution quality and performance (Hans et al., 2015).

Besides depending on the selected rules, the solution quality and performance also depend on the given problem instance. Since, in real word scenarios, the characteristic of a problem instance, e.g., the demand, is uncertain, the selection of appropriate rules is hard to handle and the solution quality cannot be guaranteed.

To this end, we provide a best-of-breed approach that enables assembling heuristics with different characteristics and thus also different solution qualities. Our approach aims at a steady solution quality of the optimization problem compared to single heuristics. The main idea is to efficiently use different heuristics for the same cloud resource assignment problem. Proposed heuristics are either executed concurrently or sequentially, and the best solution in terms of minimal total cost is returned.

We determine the set of (priority-based) heuristics to be used based on quality comparisons and the statistical paired t-test. We use this tool to infer whether a selected heuristic delivers better quality of statistical significance, or not. Those heuristics that deliver the highest solution quality without statistical differences

among each other are grouped into a candidate heuristic group. With respect to performance, we select the candidate heuristic group and identify the heuristics that have the lowest computation time. For different test cases, we thus identify the best and fastest heuristics. The identified heuristics form our best-of-breed approach.

Besides the simple heuristics and the best-of-breed approach, we adopt the metaheuristic tabu search for our optimization problem. This heuristic approach is used to guide a local search procedure to avoid local optima and thus to improve the solution quality. Since tabu search is an improvement procedure, it requires an initial solution which can be calculated with one of the approaches described earlier. For a current solution, tabu search analyses the solution neighborhood and tries to find a better one. If a better solution is found, it is stored in a long term memory. If only inferior solutions are available, the one with the best solution quality is used. To avoid directly switching to an already considered solution, a short term memory, i.e., tabu list, stores already visited solutions. Such solutions are forbidden for a given number of iterations. To assess our heuristic approaches we compare them to the exact solution approach.

For further details regarding the latter two approaches, we refer the interested reader to our recent publication (Hans et al., 2016), which contains a detailed description of the heuristic approaches.

Evaluation Results

We evaluate our approach based on two variables, i.e., the cost ratio and the computation time. The first variable assesses the solution quality while the latter assesses the corresponding performance. The independent variables include the number of data centers and user clusters. We consider latency as the desired QoS parameter.

We evaluated the following approaches: Exact/optimal approach (EXA), the LP relaxed approach (REL), a priority-based heuristic (HEU), the best-of-breed heuristic (BoB), and the tabu search heuristic (TS). For the sake of readability, we listed the heuristics in descending order regarding the evaluation result.

Figure 2 shows the solution quality provided by our approaches. First, we observe that, in general, the solution quality compared to the exact approach improves with an increasing number of data centers and user clusters.

We also observe that the difference between the best-of-breed approach and the tabu search approach decreases with growing problem size. In addition, Figure 3 shows the performance measured through the computation time of the different approaches. Here, we show significant savings when dropping the exact approach and using heuristics instead. It is also noteworthy that the difference in computation time between tabu search and the best-of-breed approach grows constantly under an increasing number of data centers and user clusters.

Hence, through sacrificing a small fraction of the solution quality, the best-of-breed approach provides a much higher performance at a still very high quality of the cloud resource allocation.

Conclusion

Cloud computing provides the infrastructure for modern IT services with high quality of service requirements. A cloud provider seeking to minimize initial and running costs requires optimal resource selection to enable QoS-aware IT service provisioning.

In this report, we briefly described the cloud data center selection problem and discussed some corresponding advanced heuristics approaches. Since the particular approaches differ in solution quality as well as in performance, we consider and compare multiple approaches to solve this optimization problem. We present a best-of-breed approach that combines the benefits of different heuristics to provide a high solution quality and low computation costs. Further, we compare the approaches with our tabu search heuristic.

References

Choy, S.; Wong, B.; Simon, G.; Rosenberg, C.:
The Brewing Storm in Cloud Gaming: A Measurement Study on Cloud to End-User Latency.
In: 11th Annual Workshop on Network and Systems Support for Games, Venice, Italy, 2012.

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

Hans, R.:
Selecting Cloud Data Centers for QoS-Aware Multimedia Applications.
In: PhD Symposium at the 2nd European Conference on Service-Oriented and Cloud Computing, Malaga, Spain, 2013.

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.

Hans, R.; Steffen, D.; Lampe, U.; Richerzhagen, B.; Steinmetz, R.:
Setting Priorities: A Heuristic Approach for Cloud Data Center Selection.
In: Proceedings of the 5th International Conference on Cloud Computing and Services Science, Lisbon, Portugal, 2015.

Hans, R.; Steffen, D.; Lampe, U.; Stingl, D.; Steinmetz, R.:
Short Run: Heuristic Approaches for Cloud Resource Selection.
In: Proceedings of the 9th International Conference on Cloud Computing, San Francisco, CA, United States, 2016.