European Journal of Operational Research 244 (2015) 637–647

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

# Revenue management for Cloud computing providers: Decision models for service admission control under non-probabilistic uncertainty

Tim Püschel [a], Guido Schryen [b,*], Diana Hristova [b], Dirk Neumann [a]

[a] Chair for Information Systems Research, Albert-Ludwigs-Universität Freiburg, Platz der Alten Synagoge, 79108 Freiburg, Germany
[b] Management Information Systems, Universität Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany

ARTICLE INFO

*Article history:*
Received 4 March 2014
Accepted 15 January 2015
Available online 30 January 2015

*Keywords:*
Admission control
Informational uncertainty
Revenue management
Cloud computing

ABSTRACT

Cloud computing promises the flexible delivery of computing services in a pay-as-you-go manner. It allows customers to easily scale their infrastructure and save on the overall cost of operation. However Cloud service offerings can only thrive if customers are satisfied with service performance. Allowing instantaneous access and flexible scaling while maintaining the service levels and offering competitive prices poses a significant challenge to Cloud computing providers. Furthermore services will remain available in the long run only if this business generates a stable revenue stream. To address these challenges we introduce novel policy-based service admission control models that aim at maximizing the revenue of Cloud providers while taking informational uncertainty regarding resource requirements into account. Our evaluation shows that policy-based approaches statistically significantly outperform first come first serve approaches, which are still state of the art. Furthermore the results give insights in how and to what extent uncertainty has a negative impact on revenue.



## 1. Introduction

Cloud computing denotes a computing model that enables ubiquitous and on-demand network access to a shared pool of configurable resources, which can be rapidly provisioned and released with minimal management effort (Mell & Grance, 2009). Resources typically refer to IT infrastructures, platforms or software, which are provided as services on a per-usage basis. While Cloud computing enjoys wide popularity for users, Cloud providers face fierce competition that has led to an erosion of the margins from $1 per CPU/hour to just a few cents (Amazon, 2015).

As the revenue side is reduced by the competition, Cloud providers need to minimize their operation costs to remain competitive. Reducing the operation costs is, however, quite difficult as the workload is highly uncertain, as the exact distribution of job arrivals is unknown. To the Cloud providers' dismay, most of modern applications are online services that require immediate processing (e.g. Outlook.com, Dropbox, GDrive). As a consequence of the immediacy traditional batch processing is not applicable; nor is rescaling of the Cloud possible due to the set up time of adding additional resources to the Cloud. Thus, Cloud providers need to manage the trade-off between maintaining excess resources as a buffer and operating the Cloud with minimal resources. While the former strategy assures to meet all quality of service assertions even if uncharacteristically many jobs arrive, the latter strategy clearly optimizes on the operating costs potentially leaving customers unsatisfied with unmet quality of service assertions.

An effective Admission control that determines which job requests are processed, may alleviate this trade-off: If the workload exceeds a critical threshold, the Cloud is susceptible to fail the quality of service assertion. In extreme cases this can even result in a system overload compromising the stability of the entire system. Admission control can act as an instrument for Cloud providers to control the exact number of jobs that are confronting the Cloud in the short run. The admission control decision is, however, hampered by the uncertain jobs arrivals and, in addition, by resource uncertainty. Resource uncertainty accounts for the fact that it is literally impossible to predict the exact resource requirements necessary to meet the quality of service assertions, due to the involved complexity in the underlying IT infrastructure (cf. Kounev, Nou, & Torres, 2007).

The field of revenue management has developed many solutions to related problems in other as well as related industries. Those solutions, however, are not applicable as they do not account for the peculiarities of online applications run in Clouds, such as resource uncertainty. Cloud systems are too complex, to obtain good predictions on required resources for a certain job. Thus, resource uncertainty

* Corresponding author. Tel.: +499419435634; fax: +499415435635.
*E-mail addresses:* tjpueschel@gmail.com (T. Püschel), guido.schryen@wiwi.uni-regensburg.de (G. Schryen), diana.hristova@wiwi.uni-regensburg.de (D. Hristova), dirk.neumann@is.uni-freiburg (D. Neumann).

http://dx.doi.org/10.1016/j.ejor.2015.01.027
0377-2217/

results in a tremendous resource overestimation (Caglar & Gokhale, 2014). We extend the revenue management literature by introducing an admission control scheme that is tailored toward the needs of services requiring instantaneous access. We cope with the complicated issue of resource uncertainty by applying fuzzy set theory to revenue management. Our work incorporates feedback and significantly extends models presented in Püschel and Neumann (2009) and Püschel, Schryen, Hristova, and Neumann (2012).

Our work contributes to the literature by suggesting and testing various service admission control policies using extensive simulations. We show that the admission problem can be solved in polynomial time, which is prerequisite for instantaneous decisions. Furthermore, we show that our policies succeed in satisfying the technical requirements stemming from Cloud computing while contemporaneously securing additional revenue for the Cloud provider. In our analysis, we demonstrate how uncertainty can affect the tradeoff between the revenue base and the service request acceptance rate.

The remainder of this paper is structured as follows: In the second section, we discuss the determinants of the "Cloud Admission Control Problem". Subsequently, we review related work in Section 3 based on these determinants. In Section 4, we propose decision models that account for various real-time admission control policies of Cloud providers that focus uncertainty regarding resource demands. The fifth section comprises an evaluation, where we test and compare our models with respect to their attractiveness in terms of revenue and service request acceptance rate. In Section 6, we provide managerial implications. The paper concludes with a summary and an outlook on new research avenues.

## 2. Determinants of the Cloud Admission Control Problem

Before we formulate the Cloud Admission Control Problem (henceforth CACP), it is useful to state the characteristics of Cloud applications representing the requirements on the CACP. In total we account for seven different characteristics:

**Production inflexibility:** Providers usually maintain a Cloud infrastructure, which consists of a fixed amount of resources (e.g. servers). While resources can easily be added to the Cloud infrastructure, it takes some time until the Cloud is reconfigured. This reconfiguration delay dictates the Cloud infrastructure to be fixed in the short run. In case of online job processing the Cloud infrastructure cannot be adapted to the job admission decision.

**Perishability:** Resources managed as Cloud offer computation and storage capacities. If the Cloud is not (fully) used it is not possible to store the excess capacity for later consumption.

**Real-time decision-making:** Due to the trend toward online processing, Cloud providers face the challenge to control job admission in real time. Effective mechanisms need to work in online and real-time scenarios as well. With respect to the CACP this translates into the requirement that the admission control mechanism needs to be of very low computational cost to be computationally tractable.

**Limited/no information on future demand/jobs:** Due to the "pay-as-you-go environment", Cloud service providers have to serve customers with lack of information (i.e. non-clairvoyant). In contrast to machine scheduling problem where relevant data are available (e.g. distributions of job arrivals and job characteristics), Cloud providers have only vague information on (i) the job arrival rate, (ii) the exact resource need of jobs, and on (iii) the customer's willingness to pay.

**Non-probabilistic uncertainty of required resources:** Cloud customers typically provide estimates on their jobs' resource requirements. As customers usually do not utilize the estimated resources for the entire job lifecycle, Cloud customers can exploit the flexibility of Cloud infrastructures by devoting excess resources to other customers. The uncertainty in accurate resource prediction stems from two sources: (i) the customer's ability to make accurate predictions and (ii) the type of the job. The Cloud provider may gather information on how well the customers calculate their predictions. Customers tend to overestimate their job resource requirements to have a buffer in case the job needs more resources. In practice, customers just use a fraction of their requested and consequently allocated resources. This claim can be verified by observing the actual usage of Google's data centers. Apparently, the used resources are way lower than the allocated resources (Reiss, Tumanov, Ganger, Katz, & Kozuch, 2012). The Cloud provider can exploit this buffering behavior, as the unused resources can be used for other jobs, increasing the overall utilization of the Cloud. The information on the customers is, however, limited, if Cloud providers offer their services to the public, as there will be a frequently varying customer base. As such, it is impossible to have probabilistic information on the accuracy of the resource predictions. The second uncertainty in resource prediction stems from the type of the job that is directed to the Cloud. For example, for routine jobs (e.g. weekly accounting or controlling tasks of customers), the required resources may be exactly predicted drawing on prior observations. For jobs that are rarely conducted (e.g. data mining jobs) the required resources largely depend on the analyzed data. Predicting the job resources for those jobs is naturally very difficult.

**Best-effort vs. priority-based processing:** Cloud providers typically serve two different customer groups, which can be distinguished with respect to their quality of service (QoS) requirements (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). While some customers accept service delivery on a best-effort base, i.e. without guarantee that the job is executed, other customers may require a guarantee that the job is executed according to the contracted service levels. The former customer group pays less for their service execution, while the latter group will only pay in the case the service levels are met. Clearly, Cloud providers favor customers with contracted service levels (henceforth gold clients) over best-effort customers, who are processed only when resources are left.

**Resistance to strategic behavior:** Since customers also strive to maximize their utility, it is appealing for them to act strategically in order to gain advantages. Basically customers have three ways in which they can adapt their actions. Firstly they might be able to shift their demand in time. They might also be able to split jobs into several smaller ones or merge several jobs into one big one. Their last option is to vary their price bid. Certain strategic behavior can lead to a significant reduction in revenue of the Cloud provider. Furthermore, it can reduce customer satisfaction. Customers might be unwilling to shoulder the additional effort necessary for strategic behavior.

## 3. Related work

Quite recently, the CACP has gained in attention by the service literature. These approaches are, however, driven by the technology and not founded by management literature. Obviously, the determinants of the Cloud Admission Control Problem suggest the use of revenue management mechanisms. However, revenue management has been developed primarily for the airline industry—as such, not all requirements of Cloud services are met. Thus, we review both literature streams and evaluate them with regard to the CACP.

### 3.1. CACP in IT-service environments

The first stream stems from computer science and focuses on technical aspects. Ferguson, Nikolaou, Sairamesh, and Yemini (1996) discuss the general applicability of economic theories to resource management. Being a primer, their results are general, but not specifically associated with actual implementations. An interesting approach to realize high service levels and end-to-end QoS is the Globus Architecture for Reservation and Allocation (Foster et al., 1999). This approach uses advance reservations to guarantee QoS. A related approach to