European Journal of Operational Research 207 (2010) 401–408

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Revenue management model for on-demand IT services

Tieming Liu [a,*], Chinnatat Methapatara [a], Laura Wynter [b]

[a] Oklahoma State University, Stillwater, OK 74078, United States
[b] IBM Watson Research Center, Yorktown Heights, NY 10598, United States

ARTICLE INFO

Article history:
Received 19 August 2009
Accepted 27 April 2010
Available online 9 May 2010

Keywords:
Revenue management
On-demand
IT service
Multinomial
Logit model

ABSTRACT

This paper presents a model for applying revenue management to on-demand IT services. The multinomial logit model is used to describe customer choice over multiple classes with different service-level agreements (SLAs). A nonlinear programming model is provided to determine the optimal price or service level for each class. Through a numerical analysis, we examine the impacts of system capacity and customer waiting incentives on the service provider's profit and pricing strategies.

Published by Elsevier B.V.

1. Introduction

Today's demand for products and services changes at a much finer time scale than in previous decades. Companies need to adapt themselves to these changes quickly in order to operate profitably. In parallel to the efforts taking place within companies to respond to a rapidly changing marketplace, the "IT on demand" paradigm has emerged as a partial solution to some of these companies' woes.

Using third-party on-demand IT services, firms can access IT resources – software, web-site hosting, computational or memory capacity – when it is needed and in the quantity that it is needed. The use of IT services on demand enables a firm to concentrate on its core business and outsource its IT functions to outside vendors. Vendors ensure their functionality and handle software upgrades, maintenance, capacity expansion, etc. The firm pays for IT as a service, where the price paid depends upon usage, rather than a capital and labor outlay; in particular, when usage needs change, no new acquisition/capacity planning decisions need to be made.

A common example of on an demand e-service is dynamic off-loading of web content. When a customer, e.g., an online retailer, experiences heavy web site traffic, the excess traffic is automatically redirected to an off-loading service. Web sites are stored remotely, and jobs are run remotely. The change is usually seamless from the point of view of end users.

Contracts for on-demand IT services typically specify both price structures and service-level agreements (SLAs). Common price structures vary from a fixed unit price per quantity of IT resource consumed (where quantity may be measured in terms of CPU-time, for instance) to multi-tier price structures, in which a different unit price is charged when the usage goes beyond a threshold.

SLAs stipulate the quality metrics, and for each of those metrics, the contracts specify the desired targets that the service provider should meet. If a target is not met, a penalty may be paid to the customer, and in some cases the total penalty may increase with the degree of nonsatisfaction. Another important issue in a contract for on-demand IT services is the billing period. The quality metrics, prices, and penalty may vary from peak periods to un-peak periods.

While the gains may be clear to customers, the IT service provider may face a great challenge to operate profitably. To attract business, the service provider must offer a price visibly lower (in some cases, 25% or more) than what the customer believes he or she is spending on in-house management of IT. The customer also expects a comparable service quality level. In order to satisfy the SLAs, the service provider's capacity must be sufficient to avoid excessive SLA breach penalties. Hence, from the service provider's point of view, cost reduction is difficult and revenue generation become critical for success. Therefore, in this paper we examine how revenue management strategies could help service providers to improve their bottom lines.

The rest of the paper is organized as follows. The relevant literature is reviewed and a summary of our main contribution is provided in Section 2. Next, we present the model's major components and its properties in Section 3. In Section 4, we conduct numerical analyses to examine the impact of system capacity and customer waiting incentives on the service provider's profit and pricing strategies. Finally, we conclude in Section 5.

* Corresponding author. Tel.: +1 405 744 6055; fax: +1 405 744 4654.
E-mail address: tieming.liu@okstate.edu (T. Liu).

0377-2217/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.ejor.2010.04.024

2. Literature review

Revenue management was first practiced in the airline industry for seat overbooking control (see Littlewood, 1972; Belobaba, 1987). The last two decades, however, saw a rise in the application of revenue management techniques to many other fields such as the hotel industry, car rental agencies, retail stores, and restaurants (see Bitran and Mondschein, 1995; Petruzzi and Dada, 1999), and models in general became more industry specific. There exists a vast literature on revenue management and we refer interested readers to McGill and van Ryzin (1999), Bitran and Caldentey (2002), Elmaghraby and Keskinocak (2003) and the monograph by Talluri and van Ryzin (2004) for detailed reviews.

There is a parallel stream of literature in the networking community, some of which has been treating not only the question of resource allocation in networked computer systems, but also the price to be charged for network use. This field emerged as a result of advances in hardware and software that enable the use of priority classes in networked computer systems. Models previously used for scheduling and admission control have been extended to include pricing decisions and revenue maximization, as a result of the ability to handle multiple priority classes. A non-exhaustive list of such papers includes Afeche and Mendelson (2004), Dewan and Mendelson (1990), Fulp and Reeves (2004), Hampshire et al. (2003), Konana et al. (2000), Liu et al. (2001), Mendelson (1985), Mendelson and Whang (1990), Nair and Bapna (2001), Paschalidis and Tsitsiklis (2000) and Van Mieghem (2000). While many of those models are not directly applicable to on-demand IT service, that line of work is of interest to us in that queueing and congestion due to network traffic are modeled, and hence some of the mechanics of the applications (Liu et al., 2001) are similar to ours. The objectives of the above papers are to determine long-term, asymptotically optimal policies. This contrasts to our goal of running an IT system over a finite horizon, with prices and service quality guarantees.

Research has also been conducted on the pricing structures in on-demand IT services. Paleologo (2004) accounted for the impact of uncertainty in the decision process in the pricing of on-demand e-services and proposed a method to set prices by maximizing net present value with probabilistic bounds on the gross profit margin. Sundararajan (2004) showed that a combination of usage-based pricing and unlimited-usage fixed-fee pricing is optimal for the pricing of information goods in the presence of transaction costs, contrasting the well-known results from nonlinear pricing which suggests the optimality of purely usage-based pricing (see Wilson, 1993). Huang and Sundararajan (2005) studied three pricing models for on-demand computing under different settings of service provider's infrastructure choice and cost structure, and customers' valuation.

Unfortunately, none of these papers model the queueing behavior in on-demand IT services or its impact on the quality of service, and hence on the total price paid and demand induced. In addition, the pricing structures discussed in these papers, while of interest for designing long-term contracts, are not directly implementable in on-demand IT services with varying quality metrics and prices over the time. With the consideration of time-dependent qualities and prices, the above pricing structures in the literature would make the model analytically intractable.

Finally, another revenue management model for IT resources has been advocated by a group of researchers at IBM (Dube and Hayel, 2006; Dube et al., 2005; Dube et al., 2007). In Dube and Hayel (2006), a single-period, multi-service-class yield management model was developed based on linear demand function and weighted-utility customer choice model. In Dube et al. (2005), the problem was formulated as an optimization model with fixed sojourn times of customers/jobs at the facility, and the model was studied analytically in a simplified setting with two price-service-level classes. In Dube et al. (2007), pricing and outsourcing decisions were examined at the competitive equilibrium between two firms.

Although the queueing behavior in IT services has been considered in those three papers, the quality metrics and prices are still time-independent. In a single-period model, customers' decision is to either accept or reject the price and service offering, as opposed to the model in this paper that allows users to defer their usage to a later time, for a more attractive price or for a better quality of service (QoS).

In this paper, we consider the problem for a service provider to determine optimal prices (or service levels) along with capacity allocations to multiple service classes for multiple periods. The revenue management model developed in this paper is based on the multinomial logit demand function, which describes customer choices facing any number of discrete choices.

The paper contributes to the literature on on-demand IT services by considering both the queueing behavior and the deferred service option. The queueing behavior in IT services and its impact on QoS are considered in both the capacity constraints and the firm's expected profit. The deferred service option redirects some peak-period demands to be processed in off-peak periods. The deferred service option has the additional benefit of inducing users to think of IT as a shared and limited resource, and to make efficient use of it not only individually but at a company-wide level. For example, tasks that can be accomplished at off-peak times can therefore be performed at lower cost, leaving resources available at the more costly peak times to higher-priority tasks. Indeed, the use of IT on demand from an external service provider leads naturally to a shift in the way IT resources are used (Dube et al., 2007).

3. Model

3.1. Service classes

We assume that the planning horizon (e.g., a 24-h day starting at 8 AM) is separated into N time periods. For simplicity, we assume these periods are of equal length. However, the model can be modified with periods of different length, by using the expected amount of demand and service capacity in a period to replace the arrival and service rates in the optimization model. Alternatively, periods of different lengths can always be chopped into small intervals of the same length. We denote by t the period when a job is submitted, and by s the period when it is actually processed. Users may wait until a later period to have their transactions processed. The motivation from the user's point of view would hence be to obtain a lower price or a higher level of service.

We consider transaction-based IT service. Transactions, or jobs, are submitted by customers in a Poisson process with arrival rate λ_t in period t. Without loss of generality, we assume each transaction requires the same workload, since long jobs could be split into small independent transactions. We also assume that the processing time for each transaction is much smaller than the length of the time period, and no transactions will be processed across periods. Given the focus in this work on transaction-based IT services, this assumption is not overly restrictive.

We assume that the service provider offers K service classes. All transactions processed in the same service class during the same time period experience the same level of service, but may be charged different prices depending on the submission periods. Denote r_k^{ts} as the price of a unit transaction submitted in period t and processed in period s in service class k. To encourage customers to