Yield management for IT resources on demand: Analysis and validation of a new paradigm for managing computing centres

Laura Wynter,* Parijat Dube and Yezekael Hayel

Received (in revised form): 28th October, 2004

*IBM T. J. Watson Research Center, PO Box 704, Yorktown Heights, NY 10598, USA
Tel: +1 914 784 7318; Fax: +1 914 784 6040; E-mail: lwynter@us.ibm.com

Laura Wynter *is a Research Scientist with IBM Watson Research Center. She works at the frontier between network equilibrium, optimisation and communications, and has developed techniques for the pricing and allocation of IT services over networks. She is leading the effort to integrate Yield Management within IBM's On Demand offerings. Dr Wynter worked previously on modelling transportation network equilibrium at the Université de Versailles, where she holds an associate professorship, at INRIA and INRETS, in Paris. Her Bachelors, Masters and PhD degrees are from MIT and Ecole Nationale des Ponts et Chaussees, Paris. She has numerous publications in journals and conference proceedings, and is an Associate Editor of Networks and Spatial Economics.*

Parijat Dube *received his MS in Electrical Communication Engg. from the Indian Institute of Science, Bangalore, in 2001 and his PhD in Computer Science from the University of Nice-Sophia Antipolis in 2002, where he was affiliated to INRIA, Sophia Antipolis, France. He joined IBM T. J. Watson Research Center, Hawthorne, NY, in 2002. His research interests include queuing theory, performance evaluation and control of communications networks, sensor networks, revenue management and pricing.*

Yezekael Hayel *received a MS in Statistics and Stochastic Modelling and a MS in Computer Science, both from University of Rennes I, France, in 2001 and 2002. He is currently in the third year of his PhD. His research interests include network modelling, performance evaluation, queuing network, pricing for communication networks and QoS.*

ABSTRACT

The field of information technology (IT) services, and, in particular, on-demand IT utilities is an emerging field of application for yield management. A detailed analysis of one instance of that model is provided, both in simplified cases where an analytical analysis is possible, and numerically on larger problem instances, and the significant increase in revenue that can accrue through use of yield management in an IT on-demand operating environment is studied.

KEYWORDS: information technology, on-demand business, yield management, optimisation, discrete choice function, model and analysis

Journal of Revenue and Pricing Management, Vol. 3, No. 4, 2005, pp. 000–000
© Henry Stewart Publications, 1476–6930

INTRODUCTION

Business environments today, in the *information technology* (IT) era, are both dynamic and unpredictable. To compete, firms must react quickly, in terms of both their pricing and the services or goods that they offer. Just-in-time paradigm and supply chain optimisation have been a part of this tendency, but e-commerce has made it more acute. Thus, there is a need for more efficient business models that can pave the way for transforming the way firms conduct business.

'IT on demand' is a business paradigm touted by the major players in the IT sector. In short, an on-demand business is one whose business processes are integrated end-to-end and who can respond quickly to changes in customer demand or market characteristics. IT on demand, in particular, concerns the ability of a business to access information technology — software, online services, computational or memory capacity — when it is needed, without any visible impact to the business or its clients.

IT on demand takes advantage of high network speeds and sophisticated middleware which allow seamless operation of IT resources, remotely. IT on demand is clearly a win–win proposition: it opens new markets for the IT provider as well as new capabilities for the customer. The on-demand IT provider experiences considerable scale economies through resource sharing; the customer saves on outlay expenses, converts purchases to operating costs, and reaps the savings of the scale economies passed on by the provider.

One example of an on-demand IT service that exists today is the case of dynamic off-loading of Web content. When a customer, such as an online retailer, experiences very heavy website traffic, that retailer may have its excess traffic automatically redirected to an off-loading service. The process is invisible to end-users of the retailer. Many other potential applications of on-demand IT are on the horizon: application service providers running software applications on their own cluster of servers and allowing customers, for a fee, to use those applications remotely is one such example.

Yield management is a potentially valuable paradigm for achieving profitable resource allocation in an on-demand IT centre. For example, yield management could be used by the Web content off-loading service provider to allocate its own capacity optimally and profitably; in this case, the provider would set capacity allocations (server use, storage and bandwidth) and multiple price points to offer to customers, depending on the available resource level of the service provider, as well as the market demand. Similarly, computing centres that rent processing capacity to customers can operate more profitably and more efficiently by incorporating yield management.

Yield management is the technique used by the airline reservation systems to set booking limits on seats at each price class. Similar to the airline setting, in an on-demand operating environment, customers and jobs or service requests arrive at random. Whereas some of the IT system resources are pre-reserved, the real-time arrival of new customers introduces the means to accomplish any number of desired service objectives by setting prices judiciously.

For example, when spare capacity exists, introducing dramatically low prices serves to induce new demand into the system. Yield management allows the provider to set dramatically low prices without sacrificing profits. On the contrary, it was proved by Wynter *et al.* (2004) that, under certain conditions, as the number of price points increases, the revenue increases. When usage costs are increasing linearly or sub-linearly in the number of users, as is generally the case, profits can be shown to increase monotonically as the number of

price points increases as well, in spite of the fact that some price points can be set below cost. The key to the remarkable increase in revenue and profits is that the number of slots available at each price point is limited and set optimally so as to maximise revenue, given the demand model and available resource level.

This paper analyses the model introduced by Wynter *et al*. (2004), both analytically and numerically. The analytical study is carried out on a simplified version of the model with only price points and fixed job sojourn times; as such, it provides a bound on what can be said about the full-scale model. The numerical study then illustrates the benefits that accrue and the present model can provide and confirm the tractability of the yield management paradigm to the IT on-demand context.

The setting of this work is the adaptation of yield management techniques to the management of an on-demand computing centre. A 'computing centre' can be defined as a group of computing resources that perform one or more related functions. Typically, a computing centre may be composed of a number of processor resources (eg servers), disk storage, application, software tools etc. A computing centre in an on-demand scenario needs to support customers with varying job sizes and varying computing resource requirements in a 'profitable' manner. Though existing resource management techniques used in computing centres aim to satisfy constraints associated with the computing needs of the customers, the pricing of service offerings is typically handled separately from and independent of the resource management system.

While airline yield management models are clearly of great relevance to the problem of yield management in on-demand IT services, there are notable differences which lead to significantly higher complexity in the present setting. First and foremost, the service under consideration in IT on demand does not have a fixed duration nor does it occupy a predetermined percentage of the resource capacity. That is, an airline seat is occupied precisely for the duration of the flight, and the number of seats to sell on any flight is known in advance. For example, such work as is found in Kleywegt (2001), Littlewood (1972), van Ryzin and Vulcano (2003) and Talluri and van Ryzin (2001) does not have these features.

Conversely, in on-demand IT utilities, the duration of a job depends on the type of server upon which it runs, and the number of servers, if it is parallelisable; further, the number of servers it requires depends on the type of servers that are used. In other words, both the capacity needed and the time taken by a job are not simple, exogenous parameters in the compute on-demand yield management problem.

Some features of this time variability can be observed in other sectors, such as *hotellerie*, in which hotel stays span varying numbers of nights, restaurant yield management, in which visits span varying numbers of hours, and even golf course yield management (see, for example, Kimes *et al*., 1999; Kimes, 2001). Nonetheless, the capacity and percentage of capacity occupied in these latter examples are still fixed and exogenous, as opposed to the setting with which the present study is faced. However, like the airlines, there may be the possibility to have 'fences' (eg Saturday night stay requirements) in the yield management of computing resources to prevent significant revenue dilution. This can be implemented by selling memory or CPU in chunks of predefined units, so that a job which requires a fractional unit of memory/CPU has to buy the full unit.

Work on the pricing of information services, such as the pricing strategies of Internet service providers (ISPs) has

traditionally considered some of these issues of job duration and capacity occupation through queuing formulae. The literature on that and related areas is quite vast, and a thorough survey of it is not the focus of this work, but a few relevant references are Elazouzi *et al.* (2003), Liu *et al.* (2003) and Mason *et al.* (1995). The difference between the decisions optimised in those and related work is the degree of segmentation of the demand and choice functions. In the Internet pricing world, a single price per type of service is proposed. It is sometimes the case that multiple qualities of service are discussed but, in that case, each quality of service (QoS) level has a single fixed price associated with it. The number of such price levels is generally limited to three, for example, gold-, silver- and bronze-level service. The yield management strategy takes customer segmentation to a much finer level and does so through the incorporation of demand models.

Some work in computing resource allocation is relevant to this work as well. In particular, Liu *et al.* (2001) make use of service-oriented characteristics and perform the resource allocation by maximising revenue less a unit cost (penalty) for non-satisfaction of the service level agreements offered to that service class. However, this approach takes as *given*, some fixed revenue or cost that accrues to the system if the request is satisfied or not.

The structure of the paper is as follows. The next section presents a simplified version of the model proposed in Wynter *et al.* (2004) and studies it analytically. The third section contains a numerical study of the model, and the final section concludes.

MODEL AND ANALYSIS

It is assumed that the on-demand service infrastructure is composed of a pool of homogeneous nodes (processing units) to allocate to different fee classes. The optimisation problem that needs to be solved is then the following: in a particular time epoch (this paper considers only one), one would like to reserve the available resource for the different fee classes. The resource should be allocated so as to maximise expected provider revenue, which is related to the distributions of different customer arrival types, their preferences (in terms of service/price trade-offs) as well as their service requirements, and to the number of nodes assigned to each fee class, on each server type.

Fee classes are defined as follows: for an identical resource, several different prices may coexist; each fee class then has a maximal number of users, and once that number is reached within the time period for that fee class, new requests are offered only at the next higher level fee for that resource.

Resources are also defined in a broad way. While a server and storage are clearly aspects of the resource, so are the Service Level Agreement (SLA) parameters, such as availability, advance notice, penalties in the case of non-satisfaction of SLAs by the provider, etc. The broad scope of the resource in this manner allows the price differentiation to become still finer grained; that is, for an identical server/storage combination, different SLA offerings create new sets of fee classes.

With respect to notation, T_c is the (here, deterministic) sojourn time of a job of class c in the system. While job sojourn time generally depends upon the workload or size of the job W, and, especially, the number of slots allocated to that job, that dependence leads to non-convexities; indeed the number of slots n_k to allocate to each price class k is the decision variable. Therefore, this paper assumes the job sojourn time to be externally provided.

The choice probability of a user accepting a slot of segment-type k can be expressed in general as a function of W and n, where n is the vector of n_ks. Again

for analytical simplicity, the dependence on the particular workload has been suppressed, and the use of a choice probability of the form $P_k(n)$ has been used. The decision variables are denoted by n_k, representing the number of resource slots to reserve for price segment k. Recall that only one time epoch is being considered here. The parameters r_k are the price points of the resource. By enumerating a wide range of such prices, the optimisation model works by identifying those price segments which are most profitable to offer, given the characteristics of the available demand and resource levels.

Simplified two-variable model

As stated earlier, a simplified model with single customer class and two different prices per node is considered, ie $r_1 \neq r_2$; further, let $T_c = T$. Under these simplifications, the model can be analysed and bounds obtained on parameters for the existence of an explicit closed form solution. The simplified yield management for IT resource model can be expressed as

$$\max_{n_1,n_2\geq 0} F(n_1,n_2) = \sum_{k=1}^{2} T r_k n_k P_k(n) \qquad (1)$$
$$n_1 + n_2 \leq N$$

Recall that the decision variable is $n = \{n_k\}$, the number of slots to allocate to each price class k. Alternatively, one can assume the resource limits as *soft constraints* and include the possibility to surpass those limits at a cost associated with having to make use of remote resources or to pay a penalty to the customers. While the results in this paper cannot be extended in general for any number of parameters, they, along with the larger-scale numerical results, provide a valuable insight into the nature of the problem under study.

To model the behaviour of customers or job requests, a *stochastic* discrete choice function is introduced. That is, the logit model is used, which randomises the utility of choice i by a Weibull random variable and normalises that quantity by the sum of all randomised choice utilities. For more details on the logit discrete choice function, Ben-Akiva and Lerman (1985) remains a good introductory source.

That is, for all k

$$P_k(n) = \frac{e^{-\theta U_k(n)}}{\sum_{j=1}^{k} e^{-\theta U_j(n)}}$$

where θ is the control parameter that determines the degree of randomness of the user choice model. Thus, $\theta = 0$ means that the choice is purely random and does not depend upon the utilities (but rather is constant at $1/K$), and $\theta = \infty$ means that the choice is purely deterministic in that, when k is the minimum disutility choice, the probability of choosing it is equal to 1, and the probability of choosing any other option $k \neq j$ is 0. Here, as the number of price segments is set to $K = 2$, the pair of logit preference functions is obtained

$$P_1(n) = \frac{1}{1 + e^{\theta(U_1 - U_2)}}$$

and

$$P_2(n) = \frac{1}{1 + e^{\theta(U_2 - U_1)}}$$

where the (dis)utility function for $k = 1,2$ is

$$U_k(n_k) = \zeta_1 T r_k n_k + \zeta_2 T$$

The parameters ζ_1 and ζ_2 are constants that define the price–time trade-offs and render the utility U_k unitless. There are different ways to define these parameters, but a single, deterministic, parameter vector was chosen for all customers. The utility function is thus linear in the decision variable n_k.

Explicitly including the logit discrete choice model into the objective function for this two-price-segment model gives

$$\max_{n_1,n_2\geq 0} T \left\{\frac{r_1 n_1}{1+e^{\theta\zeta_1 T(r_1 n_1 - r_2 n_2)}} + \frac{r_2 n_2}{1+e^{\theta\zeta_1 T(r_2 n_2 - r_1 n_1)}}\right\} \quad n_1 + n_2 \leq N \tag{2}$$

As the revenue is maximum when all resources are occupied, this implies that the inequality constraint will be active at the solution. In this case, to simplify the added complexity posed by the logit model, the equality constraint will be used in eliminating one of the two decision variables from the formulae, since they sum to a constant. That is, the expressions are rewritten in terms only of n_1 as $n_2 = N - n_1$. The following objective function is thus obtained, where the first and second terms of (2) are labelled $f(n_1)$ and $g(n_1)$, respectively

$$F(n_1) = T[f(n_1) + g(n_1)]$$

Analytical solution of the model

The advantage of expressing the yield management model for a single time period and with only two possible price segments is that it can be solved analytically and the form of the objective function can be examined. Setting

$$\gamma(n_1) := e^{\theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N} \tag{3}$$

after some manipulation, the derivative of the objective function with respect to the single variable n_1 is (denote the derivative of any function by putting a dot over the function)

$$\dot{F}(n_1) = T\left[\frac{r_1}{(1+\gamma)^2}(1+\gamma - n_1\dot{\gamma}) - \gamma^2\frac{r_2}{(1+\dot{\gamma})^2} (1 = \gamma^{-1} - (N-n_1)\dot{\gamma}/\gamma^2)\right]$$

By setting $=0$, to obtain optimal n_1, one needs to solve the following equation

$$\gamma^2 r_2 + \gamma(r_2 - r_1) + \dot{\gamma}[r_1 n_1 - r_2(N-n_1)] = r_1 \tag{4}$$

In addition, one has that

$$\dot{\gamma}(n_1) = \theta\zeta_1 T(r_1 + r_2)\gamma(n_1)$$

and thus (4) becomes

$$\gamma^2 r_2 + \gamma H(n_1) = r_1 \tag{5}$$

with

$$H(n_1) := r_2 - r_1 + \theta\zeta_1 T(r_1 + r_2)[r_1 n_1 - r_2(N-n_1)] \tag{6}$$

Solving (5), one obtains

$$\gamma\sqrt{r_2} = \sqrt{\frac{H^2(n_1)}{4r_2} + r_1} - \frac{H(n_1)}{2\sqrt{r_2}}$$

Define

$$M(n_1) := \frac{1}{\sqrt{r_2}}\sqrt{\frac{H^2(n_1)}{4r_2} + r_1} - \frac{H(n_1)}{2r_2} \tag{7}$$

Thus, the optimal is characterised as a solution to the following fixed point equation

$$\gamma(n_1) = M(n_1) \tag{8}$$

Lemma 1

Let

$$r_1 \geq e^{-\theta\zeta_1 T r_2 N}[r_2 - r_1 - \theta\zeta_1 T(r_1 + r_2) N r_2 + r_2 e^{-\theta\zeta_1 T r_2 N}]$$

and

$$r_1 \leq e^{\theta\zeta_1 T r_2 N}[r_2 - r_1 + \theta\zeta_1 T(r_1 + r_2) N r_1 + r_2 e^{\theta\zeta_1 T r_2 N}]$$

Then, $M(0) = \gamma(0)$ and $M(N) = \gamma(N)$.

Proof

Using (3), (6) and (7), the result follows using straightforward algebra. □

Proposition 1
Let the number of slots for each price class be defined over a closed, bounded subset of $\Re^+$; that is, $n_1 n_i \in [0, N]$, for each $i=1,2$. Then, one has a necessary and sufficient condition to the existence of a valid solution of the fixed point equation (8).

Proof
The left-hand side of (8) is strictly increasing as its derivative is

$$\dot{\gamma}(n_1) = \theta\zeta_1 T(r_1 + r_2)\mathrm{e}^{\theta\zeta_1 T n_1(r_1+r_2)-\theta\zeta_1 T r_1} > 0$$

and the right-hand side, , is strictly decreasing as its derivative is

$$\dot{M}(n_1) = \frac{\theta\zeta_1 T(r_1 + r_2)^2}{2r_2} \left(\frac{H(n_1)}{\sqrt{H^2(n_1) + 4r_1 r_2}} - 1 \right) < 0$$

The continuity of the functions γ and M and the result of Lemma 1 give the desired result. □

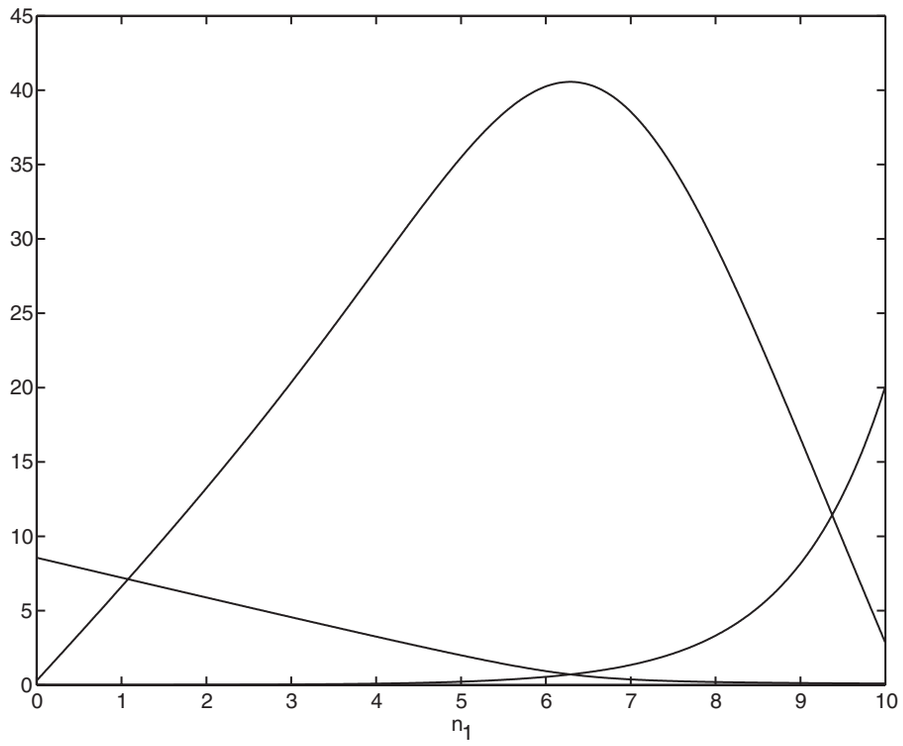
Figure 1 presents an example of this fixed point problem with constant job times $T=2$, logit scaling parameter $\theta=0.05$, value of time constant $\zeta_1=1$, prices at the two price levels $r_1=3$, $r_2=6$, and total available capacity $N=10$. In this particular case, the optimal solution is and slots at each of the two price levels.

Analytical solutions in two variables

It is possible to express the optimal solution analytically as a function of the problem parameters, T, θ, ζ_1, r_1, r_2 and N. To do so, however, one makes use of a Taylor expansion. One first makes the following assumption.

Proposition 2
Let the logit scaling parameter θ satisfy

Figure 1: Fixed point solution of (8) with two classes and logit discrete choice model and the corresponding revenue



$$\theta \leq \frac{\varepsilon}{\zeta_1 TN \max(r_1 r_2)}$$

Then, the Taylor expansion is valid with a given precision ε.

Proof

Refer to Appendix. □

The Taylor expansion of the exponential term in the logit function gives

$$\begin{aligned} e^{\theta\zeta_1 Tn_1(r_1+r_2)-\theta\zeta_1 Tr_2N} &= 1 \\ &+ \theta\zeta_1 Tn_1(r_1+r_2) \\ &- \theta\zeta_1 Tr_2N + o(n_1^2) \end{aligned} \tag{9}$$

Lemma 2

Making use of Proposition 2 and Lemma 1, one has the following condition on the price points r_1 and r_2

$$\begin{aligned} &e^{\varepsilon}(r_2 - r_1 - \varepsilon(r_1+r_2) + r_2 e^{-\varepsilon}) \\ &\leq r_1 \leq e^{\varepsilon}(r_2 - r_1 + \varepsilon(r_1+r_2) + r_2 e^{\varepsilon}) \end{aligned}$$

Remark

From Lemma 2, observe that, as ε tends towards 0, the two price levels must converge to a single price, ie $|r_1 - r_2| \leq D(\varepsilon)$, with $\lim_{\varepsilon\to 0} D(\varepsilon) = 0$ and $\dot{D}(\varepsilon) < 0$. It is not necessary, however, for $\varepsilon \longrightarrow 0$, as it is sufficient that ε be small for the expansion to be valid.

Now, using the Taylor approximation from (9) in (8), one has (with $o(n_1^2) \approx 0$)

$$\begin{aligned} &1 + \theta\zeta_1 Tn_1(r_1+r_2) - \theta\zeta_1 Tr_2N \\ &\frac{1}{\sqrt{r_2}}\sqrt{\frac{H^2(n_1)}{4r_2} + r_1} - \frac{H(n_1)}{2r_2} \end{aligned} \tag{10}$$

Solving (10) (refer to Appendix) and after some manipulation, one obtains an explicit expression for n_1^* (See equation below).

One may now compare this analytic solution found using the Taylor approximation with the solution obtained numerically from the optimisation code. Consider an example with $\theta = 0.05$, $\zeta_1 = 1$, $\zeta_2 = 2$, $r_1 = 2$, $r_2 = 3$, $N = 1$ and $T = 2$. By running the optimisation code, one obtains the solution $n_1^* = 0.1973$, with a corresponding optimal revenue of 2.6007, and using the Taylor approximation, one obtains $n_1^* = 0.2131$, with a maximum revenue of 2.6004. Thus, the optimal number of slots for class 1 obtained through the approximation has an error of $\Delta n_1^* = 8$ per cent, and the revenue difference, taking into account class 2 as well, is essentially zero.

Figures 2 and 3 show a different example with three price levels, ie $K = 3$, and a heterogeneous sojourn time in that each job class c offers a different sojourn time, T_c. The parameters for this numerical example are $\theta = 0.05$, $\zeta_1 = 1$, $\zeta_2 = 2$, $r_1 = 2$, $r_2 = 4$, $N = 10$. The solution is $n_1^* = 5.2281$, $n_2^* = 2.9909$ and $n_3^* = 1.8110$.

While it is observed from Figures 2 and 3 that the solution map is neither concave nor quasi-concave (ie its level sets are not convex), it is a 'nice' non-convex function in that a standard gradient ascent algorithm will generally converge to the global maximum, as can be observed well from the two figures and from the numerical experiments.

Induced demand curve

In Wynter *et al.* (2004), the IT yield management model was described in terms of

$$\begin{aligned} n_1^* = &\frac{(2\theta\zeta_1 Tr_2N + \theta\zeta_1 Tr_1N - 2)}{\theta\zeta_1 T(r_1+r_2)(2r_2+r-r_1)} \\ &\pm \sqrt{\frac{\left[(2-2\theta\zeta_1 Tr_2N - \theta\zeta_1 TrN)^2 - (2r_2+r_1)\left(2 - 2\frac{r_1}{r_2} + \theta\zeta_1 Tr_2N(-4+\theta\zeta_1 Tr_2N)\right)\right]}{\theta\zeta_1 T(r_1+r_2)(2r_2+r_1)}} \end{aligned} \tag{11}$$

Figure 2: Solution with three classes and logit discrete choice model

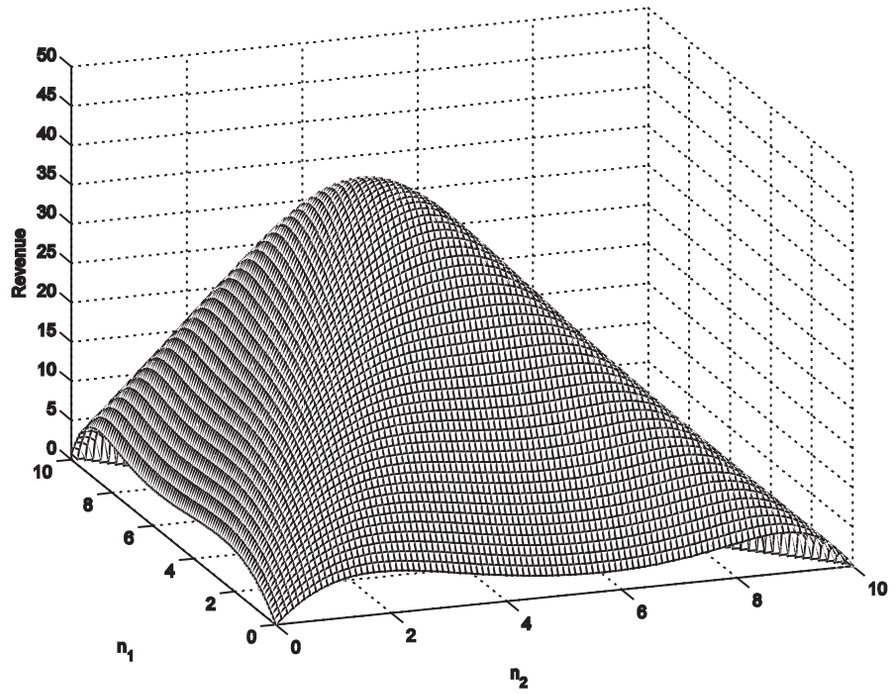
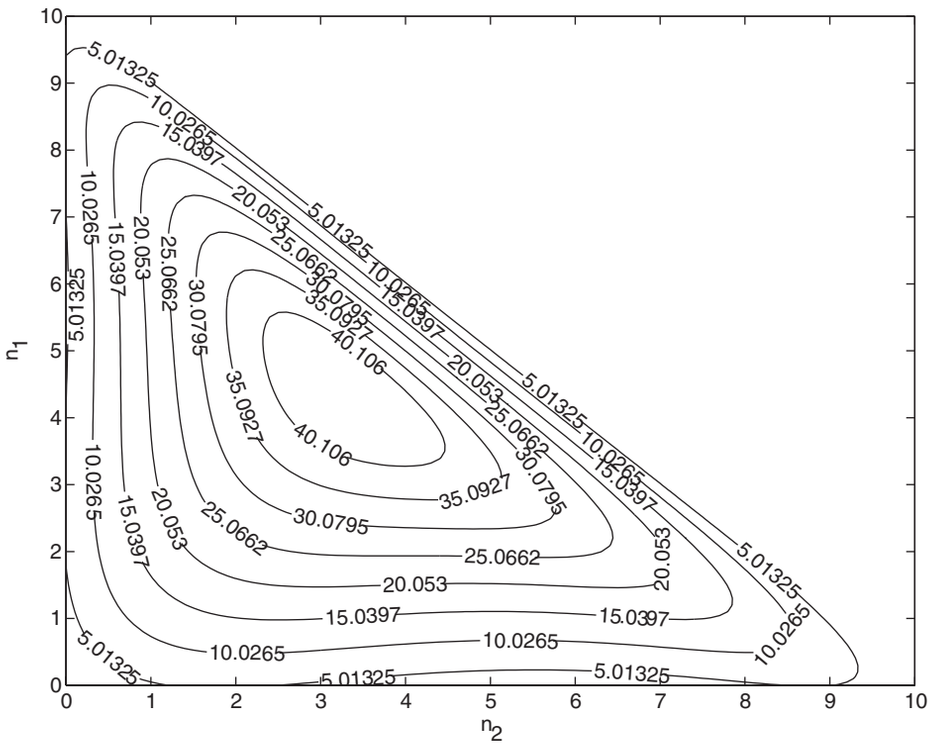


Figure 3: Level sets of the solution with three classes and the logit discrete choice model



price–demand curves. While it is straightforward to use demand curves in computing optimal yield management prices and quantities, it is difficult to calibrate such curves for multiple parameters such as price, delay, reliability, etc. For the sake of understanding the logit-based, multi-parameter model, and comparing it to the understanding of the optimal yield management prices and quantities with a demand curve, an *induced* demand function from the logit model within the optimisation code with K classes is computed here. For example, if $K=2$, the equations used in the optimisation code are given by (1).

The induced demand function is generated by determining the expected quantity that subscribes to the IT service based on the multivariate logit model at a given price, all other data being fixed. Then, this is repeated at a number of different prices to permit tracing the curve. One expects at the very least that the induced demand function is a decreasing function of price. The induced price–demand function is illustrated first in two dimensions. In Figure 4, one such demand curve (solid line) is generated in this manner. The dotted line above illustrates the revenue curve with price, that is $R(p)=pd(p)$, where $d(p)$ is the induced demand curve shown below it.

Analytical solution using the induced demand curve

Figure 5 plots the optimal yield management solution, as obtained analytically from the induced demand function (see Wynter *et al.* (2004) for more details on solving analytically for the yield management results using demand curves). The

Figure 4: Demand function for the kth class with the logit discrete choice model and $\theta=0.5$

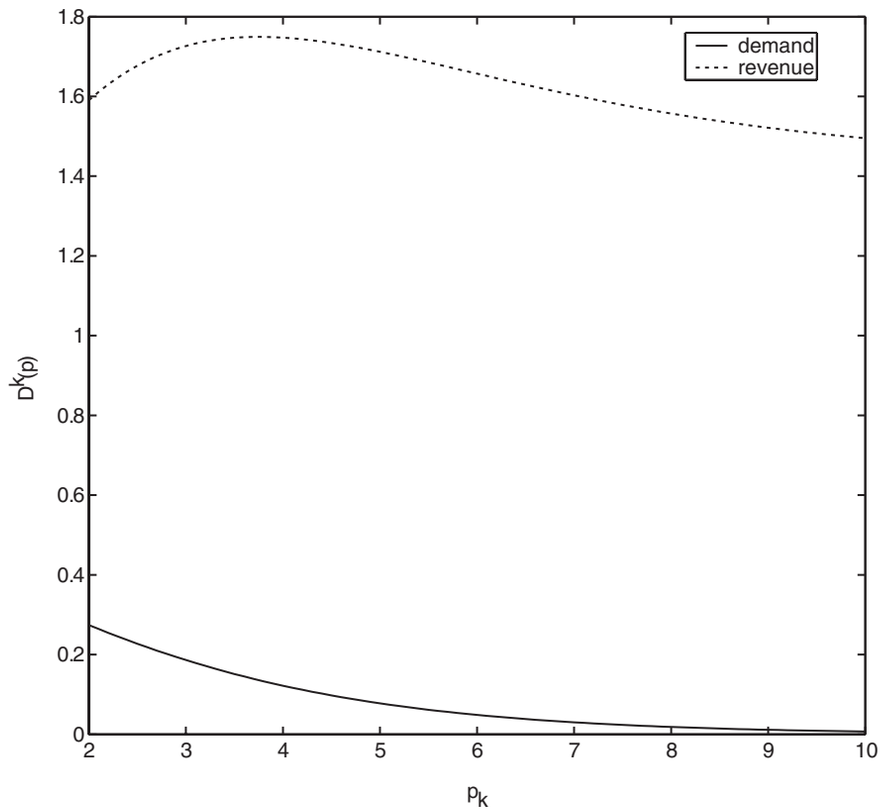
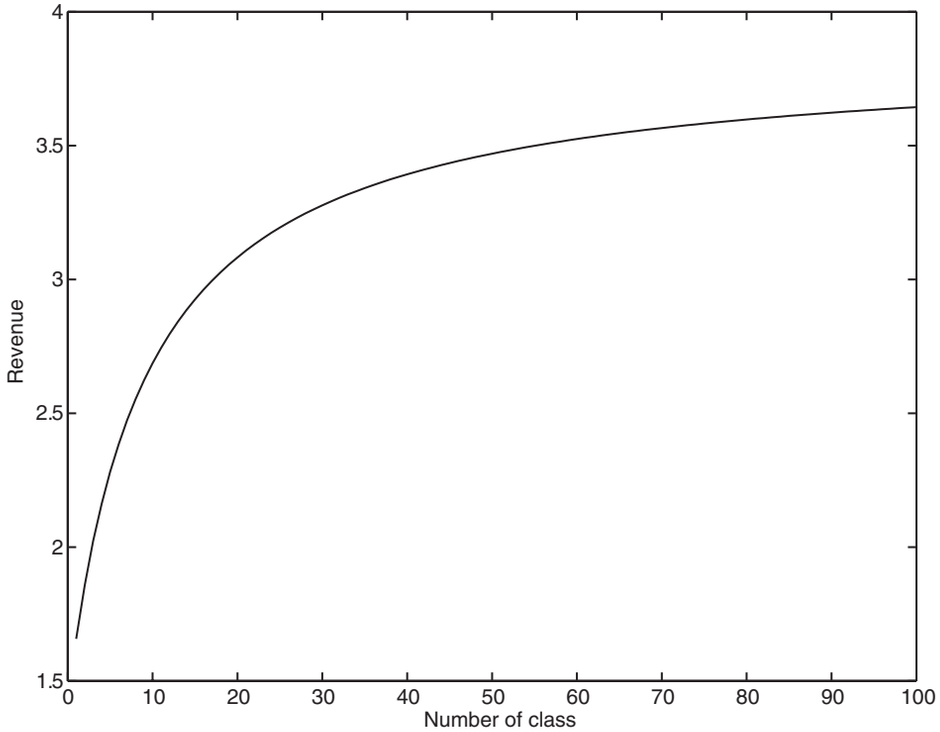


Figure 5: Total revenue as a function of the number of price segments used, where the location and quantities offered at each price segment are optimised



'analytic' optimal solution illustrates the increase in revenue as the number of price segments increases, using the induced demand curve directly. That is, each point on the optimal revenue curve is obtained by plotting the expected revenue when the value (in terms of the actual prices to offer and quantities to offer at each price) of each segment is determined optimally.

The optimal revenue curve, obtained through the use of the induced demand curve, is useful in pointing out where the trade-off in increasing complexity due to a high number of price segments is balanced by a revenue that increases very little. For example, Figure 5 shows that the revenue is close to its asymptotic value at around 70 price segments, and within 10 per cent of that value at around 30 segments.

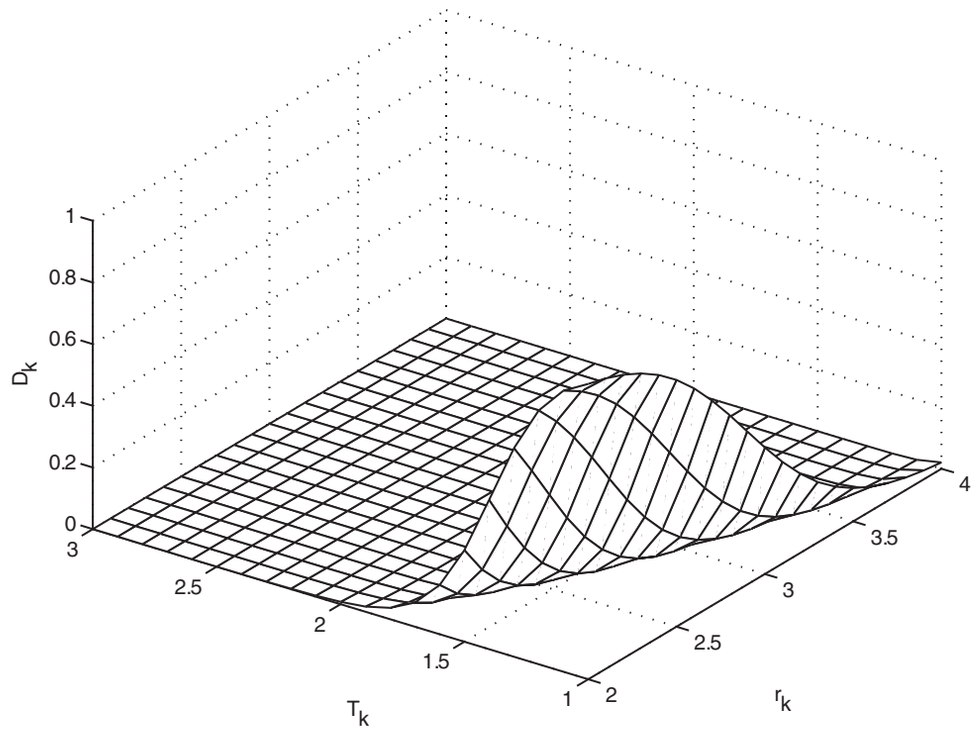
It is possible to induce such a demand curve in three dimensions as well, by taking into account both price and service quality on separate axes.

Figure 6 illustrates such a three-dimensional demand curve which depends on the unit price and on the sojourn time separately. Note that the form of the induced demand curve in three dimensions is the natural extension of what is seen in two dimensions, in that the behaviour of each of the parameters on the demand is similar. Making use of such curves in a higher dimension would be difficult in practice, however, owing to the large amount of data that would be necessary for their calibration.

YIELD MANAGEMENT FOR WEB TRANSACTION DATA

The optimisation model is applied to Web transaction data over an eight-day horizon.

Figure 6: Three-dimensional demand function, in price and sojourn time, for some class k using the logit discrete choice model with $\theta = 0.5$



The data do not include job durations; therefore all jobs are considered to have unit duration (here, the time unit is one hour).

The subscription works as follows: some users, not willing to pay high prices for service, subscribe only if they can obtain the service at an acceptable price level to them. If no such acceptable price is available (not offered, or the maximal quantity is attained) those customers 'go elsewhere'. Other users with higher willingness-to-pay can still subscribe, until their threshold is reached, and so on. Therefore, depending upon the prices offered and the available quantities of each, a different share of the market can be captured, and revenue will thus vary as well.

The objective of the yield management system is to determine which offerings to propose to customers, and the optimal quantity of each offering to propose, so as to maximise the potential revenue. Here, the output of such a system is illustrated in terms of the optimal number of slots to propose at each of the price levels, and then the resulting revenue stream is compared with the base case, in which a single price per QoS is charged.

The transaction data represent the demand at each point of time. The yield management system model allows for the possibility that a user does not accept any of the offerings proposed. This series of examples considers a single QoS level and multiple prices for that QoS, with the quantities of slots available at each price limited by a number to be determined by the yield management system. Possible price levels are determined in advance,

Table 1: Input data on possible prices for each simulation, in which one to six price segments are offered to customers in limited quantities

K, *max. number of price points*	*Actual price points, normalised to* $p \in [0,1]$
1, single low price	0.2
1, single medium price	0.6
1, single high price	1
2	0.4 0.8
3	0.3 0.6 0.9
4	0.2 0.4 0.6 0.8
5	0.2 0.4 0.6 0.8 1
6	0.2 0.35 0.5 0.65 0.8 0.95

with not necessarily all price levels open in the optimal solution; the possible set of price points is given to the right in Table 1. In the left-hand column of the table, a variable number of price segments in each optimisation run is considered, from one single price (be it high, medium or low) to six price points.

Figure 7 illustrates the optimal revenue over time when two to six price segments are made available to customers, in limited quantities. Note that the topmost (dashed) curve is the total demand, not the revenue, and illustrates the peaks and valleys in the demand pattern. The revenue accrued under each simulation (two to six price segments on offer) is indicated in the lower series of curves. The larger numbers of price segments (five to six) clearly gives higher revenue during peak periods, whereas during periods of lower demand, two to three price segments on offer is optimal. *A higher number of segments on offer maximises revenue when demand is high; for low demand (valleys), a more modest number of price segments is optimal.*

Figure 7: Revenue stream for different numbers of price segments on offer. Time periods on the x-axis are evenly spaced (in units of hours) over an eight-day planning horizon. Note that the units of demand (transactions) are different than that of revenue (dollars)

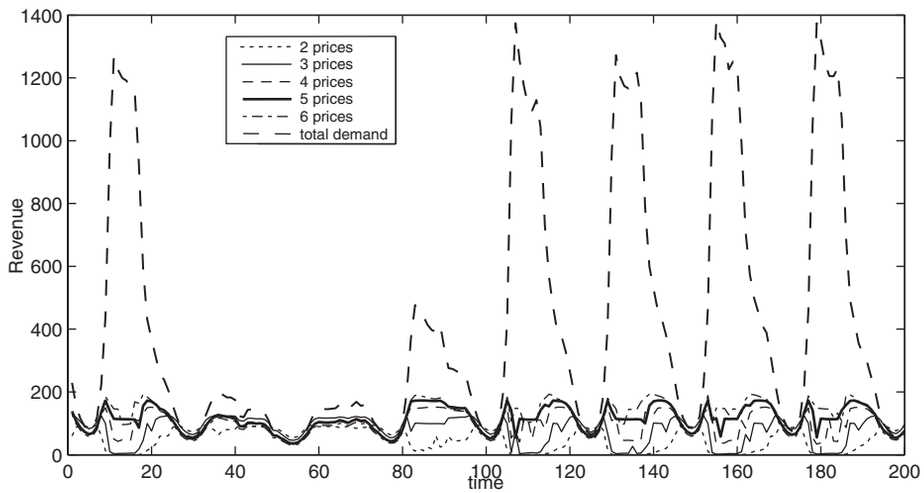


Figure 8: Optimal revenue for five different time periods (periods off-peak 40, medium 80, medium 120, peak 160 and off-peak 200) over the five different yield management strategies (offering two to six price segments)

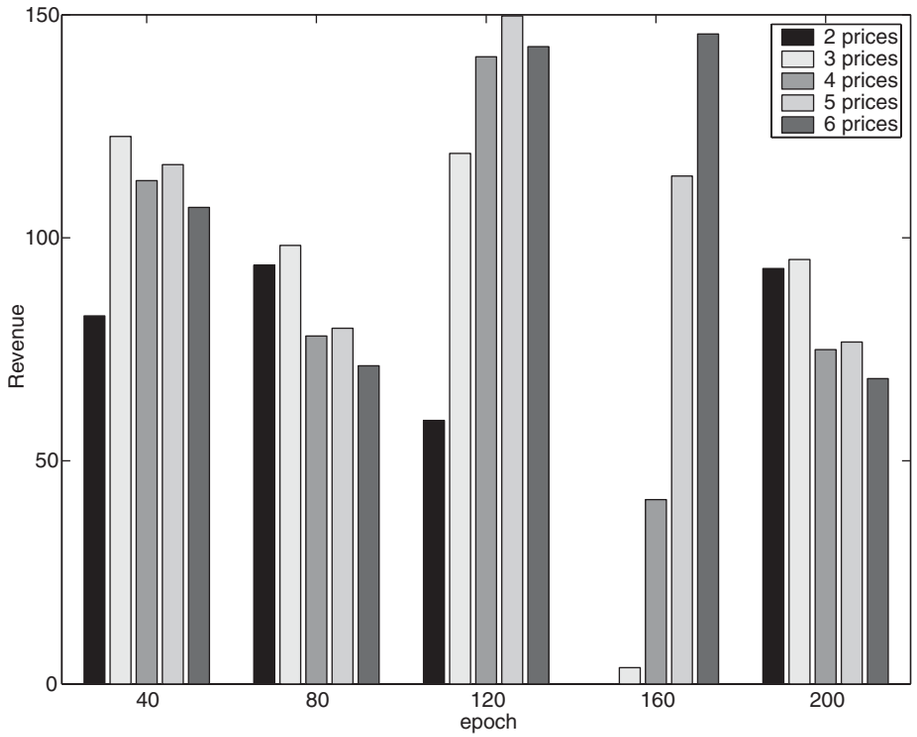


Figure 9: Comparison of the yield management strategy of offering five price segments with a single-price offering, where the single price is either low, medium or high

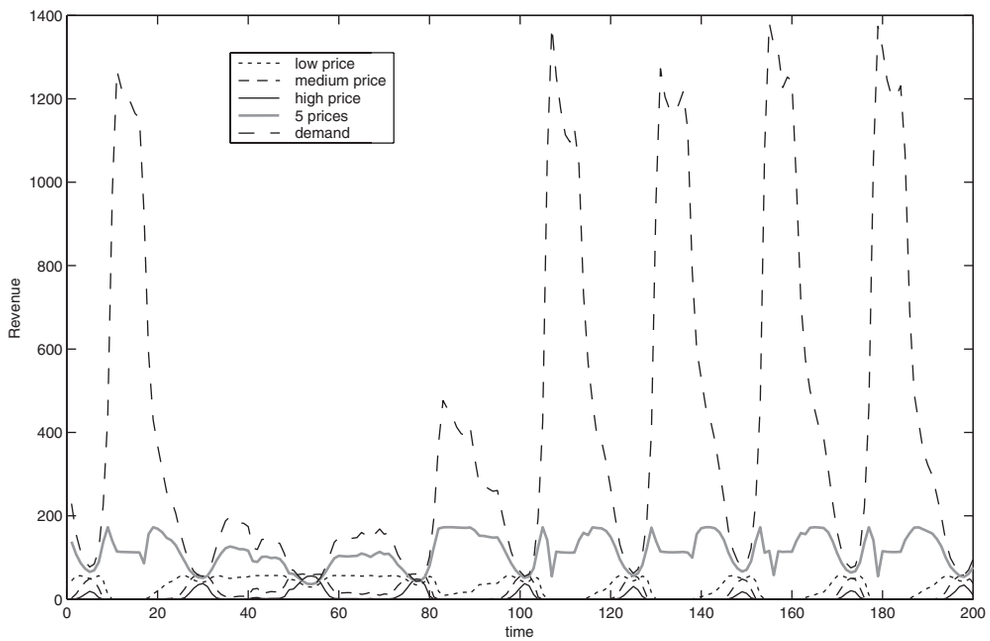


Figure 8 summarises the data in Figure 7 for certain time periods, for increased clarity. In particular, five time periods were chosen, with alternating peak flows and off-peak flows, to illustrate how the optimal number of price segments to offer varies.

Figure 9 compares the revenue when only one price segment is offered (for three cases: a low, medium or high price). Figure 9 shows the optimal revenue with a strategy of offering five price points (irrespective of the demand level). Observe that the five-price-segment offering is always superior to offering a single price, irrespective of whether a low, medium or high single price is offered. Furthermore, from the above figures, it is known that the yield management system would not suggest always proposing five price segments irrespective of the load level, but would allow further revenue increase by modulating the number of segments to offer with the demand level (fewer segments when demand is low, more when it is high).

CONCLUSION

This paper has analysed a yield management model for on-demand IT services such as e-commerce services or data processing centres. It has provided a means of determining an optimal reservation of resources in order to maximise expected revenue, as well as a detailed analysis of the resulting optimisation problem when the number of class of prices is small. Finally, it provides numerical results on time series data of Web transactions that illustrate the substantial impact of the approach on service-provider revenue.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their useful comments which, in particular, helped in an improved presentation of the paper.

REFERENCES

Ben-Akiva, M. and Lerman, S. (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA.

Elazouzi, R., Altman, E. and Wynter, L. (2003) 'Telecommunications network equilibrium with price and quality-of-service characteristics', paper presented at the 18th International Teletraffic Conference (ITC).

Kimes, S. E., Barrash, D. I. and Alexander, J. E. (1999) 'Developing a restaurant revenue management strategy', *Cornell Hotel and Restaurant Administration Quarterly*, **40**, 5, 18–29.

Kimes, S. E. (2001) 'Revenue management on the links: applying yield management to the golf course', *Cornell Hotel and Restaurant Administration Quarterly*, **41**, 1, 120–127.

Kleywegt, A. J. (2001) *An Optimal Control Problem of Dynamic Pricing*, Georgia Tech. Research Report.

Littlewood, K. (1972) 'Forecasting and control of passengers', *12th AGIFORS Symposium Proceedings*, 95–117.

Liu, Z., Squillante, M. and Wolf, J. L. (2001) 'On maximizing service-level-agreement profits', *ACM Conference on Electronic Commerce*, 213–223.

Liu, Z., Wynter, L. and Xia, C. (2003) 'Pricing and QoS of information services in a competitive market' (extended abstract), *ACM Conference on Electronic Commerce*, 188–189.

Mason, J. M. and Varian, H. (1995) 'Pricing the Internet', in Kahn, B. and Keller, J. (eds) *Public Access to the Internet*, Prentice Hall, Englewood Cliffs, NJ.

Ryzin, G. van and Vulcano, G. (2003) 'Simulation-based optimization of virtual nesting controls for network revenue management', Working Paper DRO-2003-01, Columbia Business School.

Talluri, K. and Ryzin, G. van (2001) 'Revenue management under general discrete choice model of consumer behavior', Working Paper DRO-2001-02, Columbia Business School.

Wynter, L., Dube, P. and Liu, Z. (2004) *Yield Management for On Demand Computing Services*, IBM Research Report.

APPENDIX

Proof of Proposition 2

Given that $0 \leqslant n_1 \leqslant N$, the Taylor expansion holds with precision $\varepsilon > 0$ if

$$-\varepsilon \leq \theta\zeta_1 T n_1(r_1 + r_2) - \theta\zeta_1 T r_2 N \leq \varepsilon$$

Considering that $0 \leqslant n_1 \leqslant N$, one has

$$\begin{aligned}&\theta\zeta_1 T n_1(r_1 + r_2) - \theta\zeta_1 T r_2 N \\ &\leq \theta\zeta_1 T r_1 N \leq \theta\zeta_1 T N \max(r_1, r_2)\end{aligned}$$

which should be less than ε by assumption. Secondly, one has

$$\begin{aligned}&\theta\zeta_1 T n_1(r_1 + r_2) - \theta\zeta_1 T r_2 N \geq -\theta\zeta_1 T r_2 N \\ &\geq -\theta\zeta_1 T N \max(r_1, r_2)\end{aligned}$$

which should be greater than $-\varepsilon$ by assumption. One obtains

$$\varepsilon \leq \theta\zeta_1 T n_1(r_1 + r_2) - \theta\zeta_1 T r_2 N \leq \varepsilon$$

the desired result, when

$$\theta \leq \frac{\varepsilon}{\theta_1 T N \max(r_1, r_2)}$$

Closed form solution for

From (10), one obtains

$$\begin{aligned}1 &= \theta\zeta_1 T n_1(r_1 + r_2) - \theta\zeta_1 T r_2 N + \frac{H(n_1)}{2r_2} \\ &= \frac{1}{\sqrt{r_2}}\sqrt{\frac{H^2(n_1)}{4r_2} + r_1} \\ &\Rightarrow \left(1 + \theta\zeta_1 T n_1(r_1 + r_2)\right. \\ &\left. - \theta\zeta_1 T r_2 N + \frac{H(n_1)}{2r_2}\right)^2 \\ &= \frac{H^2(n_1)}{4r_2^2} + r_1 2\end{aligned}$$

Hence, with $A = \theta\zeta_1 T(r1 + r_2)$ and $B = 1 - \theta\zeta_1 T r_2 N$, one has

$$(An_1 + B)^2 + (An_1 + B)\frac{H(n_1)}{r_2} = \frac{r_1}{r_2}$$

One thus obtains the following quadratic equation

$$\begin{aligned}&n_1^2\left(A^2 + \frac{A\theta\zeta_1 T(r_1 + r_2)^2}{r_2}\right) \\ &+ n_1\left\{2AB + \frac{A}{r_2}\right. \\ &[r - 2 - r_1 - \theta\zeta_1 T r_2 N(r_1 + r_2)] \\ &\left. + \frac{B}{r_2}\theta\zeta_1 T(r_1 + r - 2)^2\right\}\end{aligned} \tag{12}$$

The first term on the LHS is equal to

$$\theta^2\zeta_1^2 T^2(r_1 + r_2)^2(2r_2 + r_1)$$

the second term is

$$2\theta\zeta_1 T(r_1 + r_2)(2 - 2\theta\zeta_1 T r_2 N - \theta\zeta_1 T r_1 N)$$

and the third, constant term is

$$2 - 2\frac{r_1}{r_2} + \theta\zeta_1 T r_2 N(-4 + \theta\zeta_1 T r_2 N)$$

Solving (12), one obtains (11).