RC23984 (W0606-107) June 22, 2006
Mathematics

Revenue Management for e-Services: Joint Optimization of Price and Service Levels

Parijat Dube, Tieming Liu*, Laura Wynter
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

*Oklahoma State University
Stillwater, OK 74078

Revenue Management for e-Services: Joint Optimization of Price and Service Levels

Parijat Dube[§], Tieming Liu[†], Laura Wynter[§]

[§]*IBM Watson Research Center, Yorktown Heights, NY 10598*
[†]*Oklahoma State University, Stillwater, OK 74078*

Abstract

This paper presents a framework for revenue management in on-demand e-services, such as e-business hosting, e-services, software as service, capacity on demand, etc. A critical component of the framework presented here is that price segmentation and quality-of-service segmentation are performed jointly. As is often the case in the revenue management literature, the relation between demand modeling and price segmentation is captured explicitly. However, unlike much of the revenue management literature, the relation between number of customers served and the quality-of-service level offered is also explicitly modeled, in this case through queueing relations. The framework presented here therefore involves models of customer behavior, customer servicing and resource allocation. We argue that extending revenue management models beyond dynamic pricing to include other, related, decisions facing a business is critical operating profitably in an on-demand market of e-services.

1 Introduction

Because of the way the internet is used in doing business today, demand for products and services changes at a much finer time scale than in previous decades. Furthermore, because of the increasingly accurate ability of customers to compare prices and other characteristics of the products and services available on the market, companies are finding that they need to adapt themselves to these changes quickly in order to operate profitably. The principal measures that a firm can use to deal with the rapid pace of business changes in such a marketplace include dynamic pricing, resource allocation, and resource provisioning. Some of these measures are facilitated by new software platforms for engaging in e-commerce.

In parallel to the efforts taking place within companies to respond profitably 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 information technology (IT) services, firms can access IT - software, online services, web-site hosting, computational or memory capacity - when it is needed and in the quantity that it is needed. The use of on-demand IT enables a firm to concentrate on its core business and outsource its IT functions to an outside vendor, who makes assurances on its functionality and takes responsibility for ensuring software upgrades, maintenance, capacity expansion, etc.

In particular, "e-Services On-Demand" means that firms can outsource much of the IT work that is required by the firm: from servers and software hosting to maintenance and upgrades. Depending upon the service level agreed to between the firm and the service provider, users need not perceive the difference; jobs and software are run remotely; web sites are stored remotely, etc; the change should be seamless from the point of view of the users. The firm then 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. One very common example of an on demand e-service that exists today is dynamic off-loading of web content. When a firm, such as an online retailer, experiences heavy web site traffic, the retailer in many cases has its excess traffic automatically redirected to an off-loading service. The process is usually invisible to web customers of the retailer.

While the gains may be clear to the customer of the e-service, the on-demand IT service provider may have trouble operating profitably. To attract business, the service provider must offer a price visibly lower (in some cases, 25% or more) than what the customer believes it is spending on its in-house management of IT. The customer also expects a service quality level comparable to that which he would have in-house. Clearly, then, if the service provider equips its own infrastructure with the same amount of IT equipment and software licenses as the client, the savings that the service provider can generate will be low. However, in order to satisfy the service level agreement (SLA) with the customer, the service provider's capacity must be sufficient or else SLA breach penalties would need to be paid to the customer, and customer satisfaction may be sacrificed as well. Hence, from the service provider's point of view, cost reduction and revenue generation become critical to the success of on demand e-services business. Furthermore, it is clear that judicious price-setting on the part of the service provider is not sufficient for a successful on-demand IT service; setting differentiated service-quality levels, and installing an appropriate level of total capacity are critical elements of the business process, from both the point of view of the service provider, and that of the customer. In other words, the decisions of segmented pricing and resource allocation, or service-level determination, are heavily intertwined, and in the instance of revenue management for IT services, cannot be effectively treated separately.

In this work, we consider the joint problem of price and service-level setting along with capacity allocation for a service provider. Specifically, we present a modeling framework for revenue management of on-demand e-services where quality-of-service parameter optimization is performed jointly with pricing optimization. The framework also involves models of customer behavior, customer scheduling and resource allocation and portrays the benefits of revenue management to e-services; that is, we believe that price segmentation alone is insufficient.

The rest of the paper is organized as follows. Section 2 offers a review of relevant literature. We present the model and its principal components in Section 3. We discuss some structural properties of the optimization problem in Section 4. Section 5 presents computational analysis with the model, and considers a scenario in which a service provider offers Enterprise Resource Planning (ERP) software on demand through a hosted facility. Finally, we conclude in Section 6 with some suggestions for future work in the area.

2 Literature review

Historically, revenue management was first practiced in airlines for seat overbooking control (see Littlewood (1972)[22] and Belobaba (1987)[3]). The last decade saw a rise in the application of revenue management techniques to diverse fields such as the hotel industry, car rental agencies, retail stores, restaurants etc. (see Petruzzi (1999)[30]), and models became more industry specific. There exists a vast literature on revenue management and we refer interested readers to McGill and van Ryzin (1999) [24] , Bitran and Caldentry (2002) [5], Elmaghraby and Ksskinocak (2003) [14] and the recent monograph by Talluri and van Ryzin (2004) [32] for detailed reviews.

In the context of Internet services, pricing is also proposed as a tool to prevent network congestion while simultaneously optimizing the revenue of service providers. However, since the internet is basically a *best-effort* service model, works on Internet pricing consider price-only decisions by the service provider with possibly muliple price classes and the service level offered to the customers results as a by-product of pricing and customer choices. More often than not, the literature on internet pricing deals with how to achieve system optimality and reduce congestion for all users through judicious pricing, rather than profit maximization for any particular provider. Nonetheless, there is some relation between the body of literature on internet pricing and our revenue management model for e-services in that the underlying mechanics of the system works in much the same way.

On that topic, Odlyzko (1997)[27] proposed a simple approach for internet pricing which he called Paris Metro Pricing (PMP), to provide differentiated services on the internet and prevent congestion. Here again, different classes of services (with different fixed prices and with static capacity allocation) was proposed with the expectation that the pricing will segment the incoming demand among different classes. Paschalidis and Tsitsiklis (2000)[29] studied congestion-based and they show that the performance of an optimal dynamic pricing strategy is closely matched by a set of suitably chosen static prices in some asymptotic sense. Yuksel and Kalyanaraman (2002)[36] discuss implementation issues for congestion based pricing of the Internet. Dube, Borkar and Manjunath (2002)[8] propose a pricing scheme where incoming customers make decisions to join a particular service class based on the price and the congestion level in the class. They showed that under heavy load the individually optimum decisions made by users (to minimize their dis-utility) leads to social optimality.

Another related branch of literature is the resource allocation literature for queueing systems; in this body of work, researchers are concerned with the optimal allocation of servicer capacity, for example as input into a scheduler or workload management software. In many cases, pricing is part of the equation, especially in the literature called differentiated or priority-bsed resource allocation for queueing systems. The principal difference of our work with respect to this line of research is that our work is aimed at providing input for a reservation system for revenue management of e-services, rather than the real-time operation and scheduling of jobs in the system. In that respect, our framework focuses more on modeling customer behavior, price sensitivity, etc. than the following literature.

In e-services, where service-level agreements (SLA) to be offered by the service provider are negotiated during the contractual phase, a service class is identified the tuple of price and service level and not by price alone. The SLAs can be described through soft Quality of Service (QoS) constraints such as probabilistic bounds on service delays or through hard QoS constraints such as deterministic bounds. Mendelson (1985) [25] work studied the effect of congestion due to queueing effects and users' related costs on the pricing and capacity decisions at a computing center with a single processor. Mendelson and Whang (1990) extended Mendelson's (1985) model and obtained optimal incentive compatible pricing for the single processor node with Poisson arrival of customers. Konana, Gupta and Whinston (2000) [21] proposed dynamic priority pricing where each priority class is associated with a price and a delay, and priority pricing used for admission control and for job scheduling. Van Mieghem (2000)[33] studied the optimal prices and service quality grades of a service provider facing heterogeneous utility-maximizing customers in queueing models and discusses both centralized systems (maximizing total system utility) and decentralized systems (maximizing the server provider's profit), under either full information or asymmetric information. Hampshire, Massey, Mitra and Wang (2002)[18] used queueing models to solve QoS provisioning and pricing problems. Fulp and Reeves (2004) [16] study a problem with joint decisions on pricing and bandwidth provisioning, and they also discuss the class promotions in off-peak hours. Dewan and Mendelson (1990) [7] studied optimal pricing and capacity decisions by assuming a general delay cost structure which incorporates the tradeoff between capacity and delay cost. Afeche and Mendelson (2004) [1] proposed a generalized delay cost structure by accounting for the interdependence between delay costs of customers and their service valuations. Finally, Dube, Touati and Wynter (2006) [9] studied the pricing strategies under competitive environments by exploiting the interdependence between price, system capacity, demand and customer preferences.

In the context of pricing of information goods in the on-demand paradigm, there has been some recent work. Gurnani and Karlapalem (2001) [17] studied the economic viability of pay-per-use licensing of packaged software over the internet by using a monopoly pricing model. Paleologo (2004) [28] accounted for the impact of uncertainty in the decision process in the pricing of on-demand e-services and proposed a novel methodology, "Price-at-Risk". Sundararajan (2004) [31] established 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) [35]). Bakos and Brynjolfsson (1999) [2] showed that bundling a large number of unrelated information goods can be signifcantly more profitable for a multi-product monopolist. Vishwanathan and Anandalingam (2005) [34] provide a review of literature on pricing of information goods with particular focus on customisation, bundling and versioning strategies adopted by service providers. A recent paper by Huang and Sundararajan (2005) [19] studied three pricing models for on-demand computing under different settings of service provider's infrastructure choice and cost structure, and customers' valuation; however, this work did not model queueing behavior or its impact on the SLA offered by the service provider.

Revenue management in the context of IT resources has been advocated by Dube, Liu and Wynter (2003)[12] and by Dube, Hayel and Wynter (2005)[11], the idea being that revenue management techniques within a reservation system can be used by IT service providers to use their own capacity optimally and profitably. In this case, the provider would set capacity allocations (server use, storage, and bandwidth) and offer multiple price points to customers, depending on the available resource level of the service provider, as well as the market demand. In [11], the problem was formulated as an optimization model with fixed sojourn times of customers/jobs at the facility. The model was analyzed in a simplified setting with only two classes, while a numerical study using a large data set made a convincing case for improved pricing and resource allocation in IT on demand through revenue management techniques. However the assumption of fixed sojourn times in [11] is too restrictive for e-services where, in general, the sojourn times are stochastic and at best can be modeled through some probability distributions. Further in e-services, requests from different customers are processed in parallel and hence yield management framework for e-services should involve more realistic service models (see Liu, Squillante and Wolf (2001)[23]). Recently Dube and Hayel (2006) [10] proposed a real-time revenue management framework by explicitly accounting for the dependence of expected sojourn times on the provider's capacity and using a fixed point formulation for getting feasible set of SLAs. However, because of the need to solve a fixed-point system just to obtain the set of feasible SLAs, that model is not practicable as input into a large-scale reservation system.

3 The Model

3.1 Service Level Agreement over a Day

We assume a given time horizon is separated into N equal-length time periods, $t = 1 \dots N$. Demand for computing and software services is divided into J *demand classes*, $j = 1 \dots J$. Transactions, or jobs, are submitted by each demand class at the start of each period; a demand class is defined so that transactions in the same class share the same features: the same workload per transaction, identical sensitivities to price, the same preference for higher/lower quality of service (QoS), and the same degree of propensity (or not) for waiting for service.

In this framework, users may wait until a later period to have their transactions processed; the motivation from the user's point of view would be to obtain a lower price or higher quality of service (QoS). Consequently, we denote by t the period when a transaction is submitted, and by s the period when the that request is actually processed.

Demand for service is characterized by two input parameters, the arrival rate of transactions, or jobs, τ_j^t, submitted by demand class j at period t, and the is the average workload per transaction in class j, w_j.

In addition to the demand classes, we assume that the service provider has the possibility of offering up to K *price-service classes*, $k = 1 \dots K$. All transactions in the same price-service class during the same time period experience the same level of service, and are offered at the same price per unit of workload.

We define the promised level of service by a multiplicative upper bound on the delay experienced by the transaction. For example, if, for some demand class, k, the shortest sojourn time of one unit of workload is T_k^0, then the promised sojourn time of a transaction with workload w in price-service class k at period s is bounded from above, with probability α_k^s, by $z_k^s w T_0$. The variable, $z_k^s \in [1, \infty)$, is thus the promised multiplicative upper bound on the delay rate of class k at period s.

The objective of the model is to determine, for the service provider, the optimal prices, $\{r_k^s\}$, and service levels, $\{z_k^s\}$, for each price-service class and each time period, that should be offered to its customers over the time horizon, T, so as to maximize the provider's revenue less the expected penalty costs given the optimal service levels offered in each price-service class.

We shall make the following blanket assumption.

Assumption 1 *The duration of any transaction in the system is much smaller than the length of the time period.*

Under Assumption 1, we shall isolate the transactions submitted in each period. Indeed, while some small fraction of the transactions, arriving late in any period, may cross into the next period, we neglect the effect of those on the next period's capacity in our model.

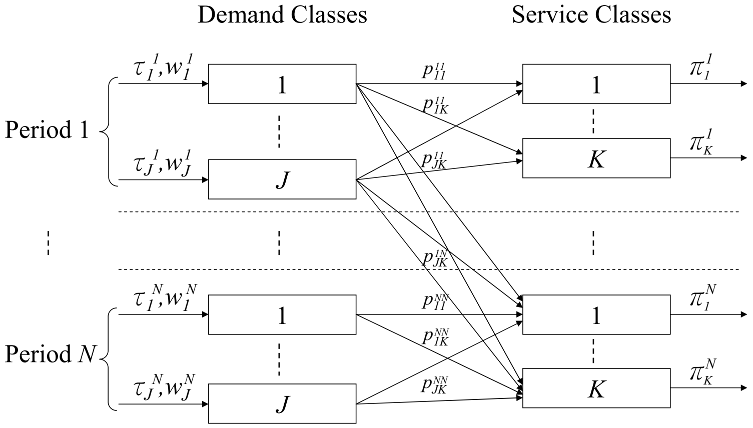

Figure 1: Customers' Choices under Service Level Agreements

Figure 1 illustrates the demand model graphically. On the left side of the figure are the demand classes, $j = 1 \ldots J$ for time periods 1 to N. The parameter τ_j^t is the customer arrival rate of demand class j at the start of period t, and w_j is the average workload per transaction in class j. On the right side of the figure are the price-service classes. The service usage rate, π_k^s, represents the total workload of service class k at period s. This rate can be obtained by

$$\pi_k^s = \sum_{t=1}^{s} \sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}. \tag{1}$$

where p_{jk}^{ts} is the probability of a user in demand class j submitting a transaction at the start of period t choosing price-service class k at period s. This probability is given by the multinominal logit function, summarized below.

3.2 Multinominal Logit Function

The logit function can elegantly describe the probability that a user chooses one amongst a finite, discrete set of options. The logit function is widely used in travel demand forecasting and marketing, and it is more robust than linear regression (see Ben-Akiva and Lerman [4]).

We define a simple, linear dis-utility function composed of price and transaction delay bound variable (upper bound on the promised delay):

$$U_{jk}^{ts} = \begin{cases} -v_j^t + r_k^s + \eta_j^t z_k^s + \zeta_j^t (s-t) & if \quad s \geq t, \\ \infty & if \quad s < t, \end{cases} \tag{2}$$

where v_j^t is the a user's value for one unit of workload of a job in demand class j arriving at period t, r_k^s is the price per unit work load and z_k^s the transaction delay bound for price-service class k at period s, η_j^t is the scaling factor converting a user's preference of high QoS into a dollar amount, and ζ_j^t is the parameter representing a class j user's willingness (arriving at period t) to wait for service in a later period in order to have lower price or higher QoS. Demands cannot be served in a period earlier than its arrival time; hence such dis-utility values are set to infinity.

In addition, in order to model the probability of non-purchase, we define the dis-utility for potential customers who choose to go to other service providers.

$$U_{j0}^t = -v_j + r_0^t + \eta_j z_0^t. \tag{3}$$

where r_0^t is price per unit workload, and z_0^t is the delay multiplier per unit workload available on the market at period t. These values represent average market values of price and service level guarantees across competitors of the service provider.

Then, the probability that a user from demand class j submits a transaction at period t to price-service class k to be processed at period s is given by the multinomial logit function:

$$p_{jk}^{ts}(\mathbf{r}, \mathbf{z}) = \begin{cases} \dfrac{e^{-\theta_j U_{jk}^{ts}(r_k^s, z^s, k)}}{e^{-\theta_j U_{j0}^t} + \sum_{n=t}^{N} \sum_{l=1}^{K} e^{-\theta_j U_{jl}^{tn}(\mathbf{r}, \mathbf{z})}} & if \quad s \geq t, \\ 0 & if \quad s < t. \end{cases} \tag{4}$$

As transactions cannot be processed in a period earlier than they are submitted, the probability is zero when $s < t$. When $s > t$, U_{jk}^{ts} is the dis-utility of a job from demand class j arriving at period t served in price-service class k at period s, and U_{j0}^t is the dis-utility of a potential customer from class j processing its job through another service provider. Due to mis-information and other unmeasurable parameters involved in decision making, there is some randomness involved in customer choices; $\theta_j \in [0, \infty)$ is the parameter characterizing the degree of randomness involved in class j customers' choices.

For example, in Figure 2, by varying this logit scaling parameter, θ, we isolate three distinct cases for a model with two options, $K = 2$.

- When $\theta = 0$, customer choices are totally random, and we can substitute the probability function with $p_k = 1/K$ for each choice $k = 1, \ldots, K$.
- When $\theta = \infty$, customer choices are purely deterministic, as customers choose the option with the minimal dis-utility.
- When $\theta = (0, \infty)$, there is some error in the perception of the dis-utility of each choice, or some degree of randomness involved in customer choices. In this case, the logit function is differentiable, but non-linear over R.

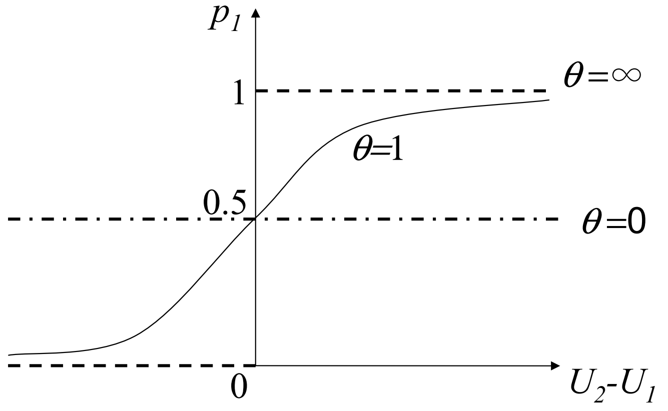

Figure 2: Logit Demand Function

3.3 Computational Capacity Allocation

Let ϕ_k^s be the portion of computational capacity, c_s, used by class k at period s, and let μ be the service rate of one unit of computational capacity. Let T_0 be the baseline sojourn time for a single workload unit, (note that the number of workload units per transaction varies across demand classes through w_j), and let $z_k^s T_0$ be the promised delay bound for a single workload unit in price-service class k. Due to the dynamic nature of the service, there is some probability that an actual service time for one unit of workload in price-service class k at period s, T_k^s, exceeds the promised sojourn time. According to a result of queueing theory [20], this probability has an upper bound,

$$P[T_k^s > z_k^s T_0] \leq \begin{cases} e^{-(\phi_k^s c_s \mu - \sum_{t=1}^{s} \sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z})) z_k^s T_0} & if \quad \pi_k^s \leq c_s \mu, \\ 1 & if \quad \pi_k^s > c_s \mu, \end{cases} \tag{5}$$

where as before, $\pi_k^s = \sum_{t=1}^{s} \sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z})$. We henceforth denote this upper bound by α_k^s. When $\alpha_k^s < 1$, we require that $\pi_k^s \leq c_s \mu$, and have:

$$\alpha_k^s = e^{-(\phi_k^s c_s \mu - \sum_{t=1}^{s} \sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z})) z_k^s T_0}. \tag{6}$$

which is equivalent to

$$\phi_k^s(\mathbf{r},\mathbf{z}) = \frac{\sum_{t=1}^{s} \sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z}) - \frac{log\ \alpha_k}{z_k^s T_0}}{c_s \mu}. \tag{7}$$

Hence, the capacity required by class k at period s can be obtained from the optimal values of the decision variables, r^* and z^* as well as the constant upper bound on probabilistic service-level degradation α_k^s.

3.4 Problem Formulation

To make the problem tractable, we use the upper bound, α_k^s, to approximate the probability of delay in the objective function[23]. According to [37] and [23], this approximation always holds for exponential distributions of the service times of the jobs, and for general distributions, is asymptotically correct when $z_k^s \to \infty$, if the service time distribution has a heavy tail. We assume that the service provider must pay a penalty of l_k if the actual service time of a unit of workload in price-service class k at period s exceeds the promised sojourn time, $T_k^s > z_k^s T_0$, and the delay ratio is bounded from above by some tolerance level, $z_k^s \leq \bar{z}_k^s$ and from below by 1, since it is a multiplicative delay bound on the nominal delay. The formulation of the IT revenue management optimization problem is given as follows,

$$\max_{\mathbf{r},\mathbf{z}} \quad \sum_{t=1}^{N}\sum_{s=t}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z})(r_k^s - l_k^s \alpha_k^s) \tag{8}$$

$$s.t. \quad \sum_{k=1}^{K} \phi_k^s(\mathbf{r},\mathbf{z}) \leq 1, \qquad for \quad s = 1,\ldots,N, \tag{9}$$

$$1 \leq z_k^s \leq \bar{z}_k^s, \qquad for \quad k = 1,\ldots,K, \quad s = 1,\ldots,N, \tag{10}$$

$$\phi_k^s(\mathbf{r},\mathbf{z}), r_k^s \geq 0, \quad for \quad k = 1,\ldots,K, \quad s = 1,\ldots,N. \tag{11}$$

where r_k^s and z_k^s are the decision variables, l_k^s, α_k^s, $\bar{z}_k^s$, and w_j are parameters, $p_{jk}^{ts}(\mathbf{r},\mathbf{z})$ and $\phi_k^s(\mathbf{r},\mathbf{z})$ are obtained from (4) and (7), (9) is the total capacity constraint, (10) is the upper bound constraint of delay rate, and finally, (11) are the constraints ensuring non-negativity.

4 Properties of the Model

In this section, we present an analysis of the model and its properties. First, we consider the existence of a feasible solution to the model. Note that since we are treating bounds on the service quality as part of the revenue management reservation, and that we model explicitly the link between customers served and service quality offered, a feasible solution may not always be guaranteed. On the same topic, we provide a bound on the total level of service provider capacity to ensure feasibility. Then, we discuss some properties of the revenue function and of the optimal solution.

4.1 Feasibility and bounds

Proposition 1 [Feasibility of the system] *Let the delay ratio bound, $\bar{z}_k^s \to \infty$ for all $k = 1 \ldots K$, $s = 1 \ldots N$. Then, the optimization problem (8)-(11) admits a solution for any value of the arrival process, $\tau_j^t, w_j \geq 0$.*

Proof: We must show that, for any fixed value of $\alpha \in [0,1]$, it holds that

$$\frac{\sum_{t=1}^{s}\sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z}) - \frac{log\ \alpha_k}{z_k^s T_0}}{c_s \mu} \geq 0, \ k = 1 \ldots K, \ s = 1 \ldots N, \tag{12}$$

and

$$\sum_{k,s} \phi_k^s(\mathbf{r},\mathbf{z}) = \sum_{k,s} \frac{\sum_{t=1}^{s}\sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z}) - \frac{log\ \alpha_k}{z_k^s T_0}}{c_s \mu} \leq 1 \tag{13}$$

for any $\tau_j^t, w_j \geq 0$. Recall that the logit probability, $p_{jk}^{ts}(\mathbf{r}, \mathbf{z})$, for any class includes the dis-utility function of the competitor, whose price and service level are fixed and exogenous to the model. Hence, the competitor's capacity is infinite. As $\bar{z}_k^s \to \infty$,

$$p_{jk}^{ts}(\mathbf{r}, \mathbf{z}) = \frac{e^{-\theta_j U_{jk}^{ts}(r_k^s, z^s, k)}}{e^{-\theta_j U_{j0}^t} + \sum_{n=t}^{N} \sum_{l=1}^{K} e^{-\theta_j U_{jl}^{tn}(\mathbf{r}, \mathbf{z})}} \to \infty. \tag{14}$$

Indeed, when $s \geq t$, $U_{jk}^{ts} = -v_j^t + r_k^s + \eta_j^t z_k^s + \zeta_j^t (s - t)$, and $z_k^s \to \infty$; also, $U_{jk}^{ts} = \infty$ for $s < t$. Hence, as $z_k^s \to \infty$,

$$e^{-\theta_j U_{jk}^{ts}(r_k^s, z^s, k)} \to 0. \tag{15}$$

Furthermore $U_{j0}^t = -v_j + r_0^t + \eta_j z_0^t$ is a constant. Thus,

$$p_{jk}^{ts}(\mathbf{r}, \mathbf{z}) \to \frac{0}{const + 0} = 0 \tag{16}$$

as $z_k^s \to \infty$. In addition,

$$\frac{\log \alpha_k}{z_k^s T_0} \to 0 \tag{17}$$

as $z_k^s \to \infty$, since both α_k and T_0 are constants. Hence, there exists a $\hat{z}_k^s$ for all k, s, such that

$$\sum_{k,s} \phi_k^s(\mathbf{r}, \hat{\mathbf{z}}) = \sum_{k,s} \frac{\sum_{t=1}^{s} \sum_{j=1}^{J} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r}, \hat{\mathbf{z}}) - \frac{log\ \alpha_k}{\hat{z}_k^s T_0}}{c_s \mu} \leq 1 \tag{18}$$

Nonnegativity of each ϕ_k^s is straightforward, since $-\log(\alpha_k) \geq 0$ for all $\alpha \in [0, 1]$, and hence all terms in (18) are nonnegative. ⊐⊔

Remark 1 *Observe that for $\bar{z}_k \to \infty$, $\forall k$, for any $\alpha \in [0, 1]$ we can eliminate the constraints on ϕ_k^s (this is because the constraints on ϕ_k^s are only design constraints and ϕ_k^s does not enter explicitly in the objective function) in our optimization problem. Thus the problem from Sec. 3.4 can be reformulated as:*

$$\max_{\mathbf{r}, \mathbf{z}} \quad \sum_{s=1}^{N} \sum_{t=1}^{s} \sum_{j=1}^{J} \sum_{k=1}^{K} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r}, \mathbf{z})(r_k^s - l_k^s \alpha_k^s) \tag{19}$$

$$s.t. \quad r_k^s, z_k^s \geq 0, \qquad for \quad k = 1, \ldots, K, \quad s = 1, \ldots, N. \tag{20}$$

The next result shows the effect of increasing service provider's capacity on the model.

Proposition 2 *Let the service provider's capacity be fixed over the time horizon, i.e., $c_s = c$ for all s. Then, there is a threshold value of c, called $\widehat{c}$, such that when $c \geq \widehat{c}$, there exists a solution to the optimization program (8)–(11) for any value of the arrival process, $\tau_j^t, w_j \geq 0$. Furthermore $\widehat{c} \leq \frac{1}{\mu} \max_s \left\{ \sum_{j,k,t=1..s} \tau_j^t w_j - \frac{\log \alpha_k}{T_0} \right\}$.*

Proof: When $c_s = c$ for all s, the constraint on service capacity says that for every $s = 1, ...N$,

$$\sum_k \phi_k^s(r,z) = \sum_k \frac{\sum_{t=1}^s \sum_{j=1}^J \tau_j^t w_j p_{jk}^{ts}(r,z) - \frac{log\ \alpha_k}{z_k^s T_0}}{c\mu} \leq 1,$$

which can be rewritten as

$$\frac{1}{\mu} \sum_k \sum_{t=1}^s \sum_{j=1}^J \tau_j^t w_j p_{jk}^{ts}(r,z) - \frac{log\ \alpha_k}{z_k^s T_0} \leq c.$$

From above, we see that as c gets large, the contraint is always satisfied. Since for all t, s, j, k, $p_{jk}^{ts}(r,z) \leq 1$, $z_k^s \geq 1$, and $\log \alpha_k \leq 0$, as $\alpha_k \in \{0,1\}$, the left-hand side is bounded by $\frac{1}{\mu} \sum\limits_{j,k,t=1...s} \tau_j^t, w_j - \frac{\log \alpha_k}{T_0}$. Hence, when the capacity is larger than

$$\frac{1}{\mu} \max_s \left\{ \sum_{j,k,t=1...s} \tau_j^t w_j - \frac{\log \alpha_k}{T_0} \right\},$$

the constraint is always satisfied. ◻

Note that the above result is a special case of the stability condition on a queueing system. Indeed, the stability of a queue requires that $\lambda \leq \mu c$, where λ is the arrival rate of jobs into the queue, and μc the service rate and capacity, respectively. In this case, $\sum\limits_{j,k,t=1..s} \tau_j^t w_j$ is equivalent to an arrival rate, as it is the product of the workload per user type and the number of such users. The difference between the classic stability condition and the condition above is that the arrival rate must be lower than the processing capability, μc, by the quantity $\frac{\log \alpha_k}{T_0}$. That quantity is the upper bound on the additional capacity that is needed to achieve the service quality promised, which is independent of the arrival rate of customers into the system.

4.2 Analysis of the optimal objective and price and service-level variables

The next result confirms similar results on other revenue management models and says essentially that indeed, for this model as well, the revenue management strategies enable a provider to obtain more of the consumer surplus than without revenue management and increased market segmentation leads to increased potential revenues.

Proposition 3 *Let $r_k^s - l_k^s \alpha_k^s \geq 0$ for all s, k. Then, the revenue function(8) increases with the number of service classes.*

Proof: The revenue function (8) is given by

$$\sum_{s,t,j,k} \tau_j^t w_j p_{jk}^{ts}(\mathbf{r}, \mathbf{z})(r_k^s - l_k^s \alpha_k^s),$$

where the difference $r_k^s - l_k^s \alpha_k^s \geq 0$ for all s,k. Furthermore, the parameters $\tau_j^t w_j \geq 0$ for all t, j. Consider next the probability function $p_{jk}^{ts}(\mathbf{r}, \mathbf{z})$. Given the non-negativity of the other parameters, we must now show that the sum of the probabilities increases as K increases. Note that the probability of customers going to the competitors is given by $p_{j0}^{ts}(\mathbf{r}, \mathbf{z})$, and furthermore that

$$\sum_{s,k} p_{jk}^{ts}(\mathbf{r}, \mathbf{z}) = 1 - p_{j0}^{t}(\mathbf{r}, \mathbf{z}).$$

In addition, $p_{j0}^{t}(\mathbf{r}, \mathbf{z})$ can be expressed as

$$p_{j0}^{t}(\mathbf{r}, \mathbf{z}, \mathbf{K}) = \frac{e^{-\theta U_0(r_0, z_0)}}{e^{-\theta U_0(r_0, z_0)} + \sum\limits_{s,k=1..K} p_{jk}^{ts}(\mathbf{r}, \mathbf{z})}.$$

Hence, the competitors' probability $p_{j0}^{t}(\mathbf{r}, \mathbf{z}, \mathbf{K})$ is decreasing in K for fixed $\mathbf{r}, \mathbf{z}$, since each term in the denominator satisfies $p_{jk}^{ts}(\mathbf{r}, \mathbf{z}) \geq 0$, and the number of terms is increasing as K increases. Since $p_{j0}^{t}(\mathbf{r}, \mathbf{z}, \mathbf{K})$ is decreasing in K, we have that $\sum\limits_{s,k} p_{jk}^{ts}(\mathbf{r}, \mathbf{z})$ is increasing in K. This concludes the proof. ⊔⊓

The above result says that the objective function of our revenue management formulation with explicit price and service level models increases as the number of available price-service classes increase, in the absence of capacity constraints. In our framework, though, the effect of increasing the mass of customers served on service quality is explicitly modeled. In particular, when service provider capacity is below a threshold, there is an effect on service quality as the number of customers served increases. Hence, depending upon the market structure, it is in many cases preferable to offer a lower number of price-service classes when capacity is low, with the optimal classes serving the higher-paying segments of the population.

Next, we study how customer's choice changes with the optimal price and service level of each class. For the purpose of simplicity, we focus on a single-period system, i.e. no demand will be served in a later period. However, the results we obtain in the following are not limited to single-period system.

Proposition 4 *In a single-period system, the Logit probability has the following properties (taking p_{jk} as example):*

$$\frac{\partial p_{jk}}{\partial r_k} = -\theta_j p_{jk}(1 - p_{jk}); \tag{21}$$

$$\frac{\partial p_{jk}}{\partial r_l} = \theta_j p_{jk} p_{jl}; \tag{22}$$

$$\frac{\partial p_{jk}}{\partial z_k} = -\theta_j \eta_j p_{jk}(1 - p_{jk}); \tag{23}$$

$$\frac{\partial p_{jk}}{\partial z_l} = \theta_j \eta_j p_{jk} p_{jl}. \tag{24}$$

Proof: For a single period system, let us denote $e_{jk} = e^{-v_j + r_k + \eta_j z_k}$. Then the Logit probability that a class j customer will choose service class k can be calculated as $p_{jk} = e_{jk} / \sum_{i=0}^{K} e_{ji}$. Take derivatives over r_k, r_l, z_k, and z_l respectively, we have,

$$\begin{aligned}
\frac{\partial p_{jk}}{\partial r_k} &= \frac{-\theta_j e_{jk}}{\sum_{i=0}^{K} e_{ji}} - \frac{-\theta_j e_{jk} e_{jk}}{(\sum_{i=0}^{K} e_{ji})^2} = -\theta_j \frac{e_{jk}[(\sum_{i=0}^{K} e_{ji}) - e_{jk}]}{(\sum_{i=0}^{K} e_{ji})^2} = -\theta_j p_{jk}(1 - p_{jk}); \\
\frac{\partial p_{jk}}{\partial r_l} &= \frac{\theta_j e_{jk} e_{jl}}{(\sum_{i=0}^{K} e_{ji})^2} = -\theta_j p_{jk} p_{jl}; \\
\frac{\partial p_{jk}}{\partial z_k} &= \frac{-\theta_j \eta_j e_{jk}}{\sum_{i=0}^{K} e_{ji}} - \frac{-\theta_j \eta_j e_{jk} e_{jk}}{(\sum_{i=0}^{K} e_{ji})^2} = -\theta_j \eta_j \frac{e_{jk}[(\sum_{i=0}^{K} e_{ji}) - e_{jk}]}{(\sum_{i=0}^{K} e_{ji})^2} = -\theta_j \eta_j p_{jk}(1 - p_{jk}); \\
\frac{\partial p_{jk}}{\partial r_l} &= \frac{\theta_j \eta_j e_{jk} e_{jl}}{(\sum_{i=0}^{K} e_{ji})^2} = -\theta_j \eta_j p_{jk} p_{jl}.
\end{aligned}$$

□

Equation (21) indicates that as the price for service class k increases, the probability that a customer choose class k decreases, and the impact of r_k is prominent when $p_{jk} = 0.5$. In other words, when $p_{jk} \to 1^-$ or when $p_{jk} \to 0^+$, customer's choice is not so sensitive to price change as when $p_{jk} \approx 0.5$. Equation (22) indicates that as the price for service class l increases, the probability that a customer choose class k increases. The sensitivity is proportional to both p_{jk} and p_{jl}. Since $p_{jl} < 1 - p_{jk}$, the impact of r_l on p_{jk} is smaller than that of r_k.

Comparing equations (21) and (23), we find that the impact of the sojourn time in class k, z_k, is proportional to the impact of r_k with rate η_j, which is class j customers' sensitivity to delay. Similar properties are found by comparing (22) and (24).

Nonconcavity of the objective function is clearly a concern as concerns solving the revenue management model. Empirically, however, we observe that when capacity is sufficiently large, standard nonlinear programming algorithms are quite well-behaved on the model. Figure 3 shows the *optimal* objective function surface from the computation study of Section 5 obtained from different starting point vectors indexed from 1 to 10. Note that the surface is not the objective function but rather the optimal objective function value according to starting point. Observe that for much of the region explored, the optimal revenue is the same, i.e., the surface is flat, confirming a relative insensitivity of the optimal solution found to starting point.

Remark 2 [Simple Processor Sharing System] *Note that our model reduces to a processor sharing system when in each period s all the classes get equal share of the total capacity c_s,*

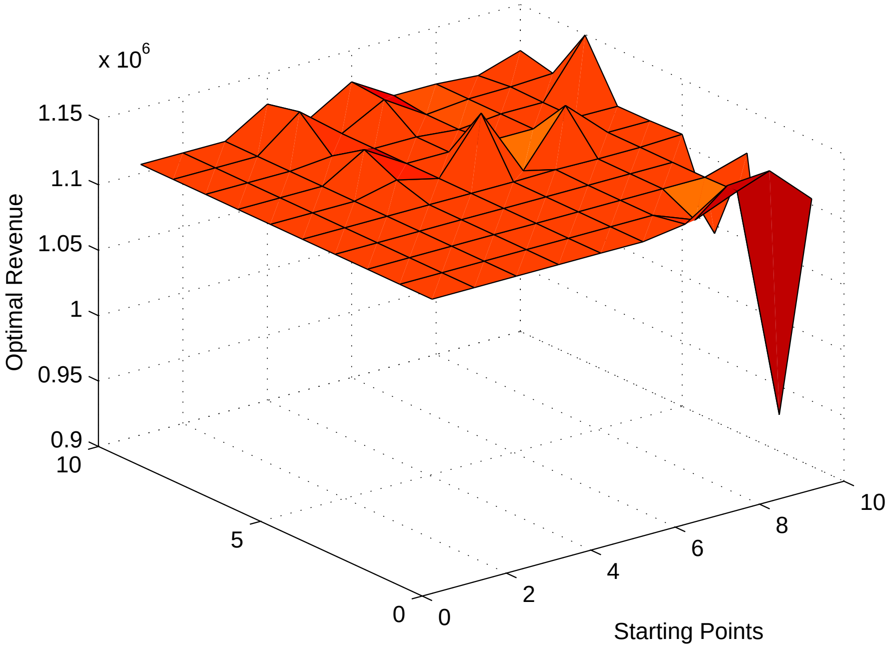

Figure 3: The optimal revenue value as a function of the starting point vector, where the starting points are indexed from 1 to 10

that is, $\phi_k^s = \frac{1}{K}$*. For such a system we have from (7):*

$$\sum_{t=1}^{s}\sum_{j=1}^{J}\tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z}) = \frac{c_s\mu}{K} + \frac{\log\alpha_k^s}{z_k^s T_0} \tag{25}$$

From (7) and (19) we can hence write the objective function as:

$$\sum_{s=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{s}\sum_{j=1}^{J}\tau_j^t w_j p_{jk}^{ts}(\mathbf{r},\mathbf{z})(r_k^s - l_k^s\alpha_k^s) = \sum_{s=1}^{N}\sum_{k=1}^{K}(r_k^s - l_k^s\alpha_k^s)\left(\frac{c_s\mu}{K} + \frac{\log\alpha_k^s}{z_k^s T_0}\right)$$

Defining a new variable $q_k^s = \frac{1}{z_k^s}$*, we can write the optimization problem of (19)-(20) as:*

$$\max_{\mathbf{r},\mathbf{z}} \quad \sum_{s=1}^{N}\sum_{k=1}^{K}(r_k^s - l_k^s\alpha_k^s)\left(\frac{c_s\mu}{K} + \frac{q_k^s\log\alpha_k^s}{T_0}\right) \tag{26}$$

$$s.t. q_k^s, r_k^s \geq 0 \qquad\qquad for \quad k = 1,\ldots,K, \quad s = 1,\ldots,N. \tag{27}$$

Observe that by restricting the choice of ϕ_k^s *to* $\frac{1}{K}$ *for all s and k we have introduced an additional dependency between the decision variables r and z through (25). This restricts the feasible set of variables r and z.*

5 Computational Analysis: Example of an On-Demand Software as a Service (SaaS) e-Services Utility

In this section, we report results of a computational study aimed at obtaining insights into the revenue management framework proposed for a typical on demand IT service provider. Our goal is two-fold: (i) examine the profit improvement for the service provider by using our joint pricing and service-level strategy, which depends on factors such as number of classes that can be opened and number of distinct classes among those that are opened, time-of-day, user demand, and provider capacity, and (ii) identify how the optimal price and service levels changes with the provider's total capacity and demand. We discuss as well the impact of the revenue management policy on the fraction of total demand that is captured by the service provider.

5.1 Data and problem setup

In this section, we study an application example of revenue management at a Software as a Service (SaaS) center supporting on-demand access to an Enterprise Resource Planning (ERP) application software, such as PeopleSoft or SAP. ERP software includes typically many modules, from finance to manufacturing to human resources, in a single platform, thereby facilitating the sharing of information and the cross-optimization of tasks and resources. Modules of an ERP software package include for example Customer Relationship Management (CRM), Enterprise Performance Management (EPM), Financial Management, Human Capital Management (HCM), Supply Chain Management (SCM), and Project Management (PM). Clearly, then, usage patterns, willingness to pay for peak-time usage and ability to run tasks offline in batch mode (and hence at possibly cheaper time periods) vary across these modules. It is precisely this variability that allows a revenue management approach to reap benefits in terms of profit increase and demand smoothing.

ERP software has traditionally been installed on the customer's computer infrastructure. In this case, it is generally deployed and maintained by the customers themselves. The on-demand software-as-a-service paradigm, or e-business hosting (eBH), involves a third-party service provider installing and maintaining a sufficient number of licenses for the software and users access the software remotely from the customer's site. As long as the service provider's capacity is sufficient, no conflicts will arise and response times for the user would be very short.

In this study, the time horizon is a single day. Let us consider a hypothetical e-services provider supporting six different ERP application suites including HRM, Finance, EPM, CRM, SCM and PM. The provider faces the problem of allocating resources (database and application servers) to these different applications so as to maximize its revenue from hosting on-demand PeopleSoft applications. The provider also needs to allocate servers for Administrative (Admin) purposes. Hence, in total there are 7 ERP modules, each of which each has a different usage pattern.

Typically for the purpose of resource provisioning, users for different applications are classified into three categories: heavy users, moderate users and light users, defined as follows:

- Heavy User (HU) - a user whose principal job requirement is to enter, request and update data. They submit transactions frequently. This type of user is sometimes referred to as a "power user" operating in a "heads down" data entry or update mode.
- Moderate User (MU) - a user that enters data, requests data and updates the database regularly but their requirements are not as intense as those of a heavy user. You may consider a moderate user workload is half as much as a heavy user workload.
- Light User (LU) - a user that submits application requests or updates infrequently.

Table 1 gives the workload of each ERP application and the percentage of customers of different customer types for each application. Apparently, for any function $G \in Func$, we have

$$\psi^t_{Gh} + \psi^t_{Gm} + \psi^t_{Gl} = 1 \qquad for \quad t = 1, \ldots, N.$$

Service	Work Load	# Transactions	% HUs	% MUs	% LUs
HCM	$w_H = 2$	$x^t_{Hh} = 10, x_{Hm} = 5, x_{Hl} = 2$	$\psi^t_{Hh} = 0.1$	$\psi^t_{Hm} = 0.3$	$\psi^t_{Hl} = 0.6$
Finance	$w_F = 3$	$x^t_{Fh} = 50, x_{Fm} = 25, x_{Fl} = 10$	$\psi^t_{Fh} = 0.7$	$\psi^t_{Fm} = 0.25$	$\psi^t_{Fl} = 0.05$
EPM	$w_E = 3$	$x^t_{Eh} = 10, x_{Em} = 8, x_{El} = 4$	$\psi^t_{Eh} = 0.25$	$\psi^t_{Em} = 0.6$	$\psi^t_{El} = 0.15$
CRM	$w_C = 4$	$x^t_{Ch} = 80, x^t_{Cm} = 45, x^t_{Cl} = 20$	$\psi^t_{Ch} = 0.65$	$\psi^t_{Cm} = 0.25$	$\psi^t_{Cl} = 0.1$
SCM	$w_S = 2$	$x^t_{Sh} = 45, x^t_{Sm} = 22, x^t_{Sl} = 6$	$\psi^t_{Sh} = 0.2$	$\psi^t_{Sm} = 0.65$	$\psi^t_{Sl} = 0.15$
Admin	$w_A = 1$	$x^t_{Ah} = 15, x^t_{Am} = 7, x^t_{Al} = 3$	$\psi^t_{Ah} = 0.1$	$\psi^t_{Am} = 0.5$	$\psi^t_{Al} = 0.4$
PM	$w_P = 4$	$x^t_{Ph} = 100, x^t_{Pm} = 50, x^t_{Pl} = 20$	$\psi^t_{Ph} = 0.85$	$\psi^t_{Pm} = 0.13$	$\psi^t_{Pl} = 0.02$

Table 1: Fraction of different types of users and their average workloads at e-services provider X.

Interval	Characteristic	Activity	# Active Customers, ν							
			HCM	Finance	EPM	CRM	SCM	Admin	PM	
0:00-8:00	night-time	very low	2	4	2	4	4	2	2	
8:00-9:00	start of day	moderate	10	20	10	20	20	10	10	
9:00-12:00	peak-time	high	30	60	30	60	60	30	30	
12:00-14:00	lunch-time	moderate	10	20	10	20	20	10	10	
14:00-16:00	afternoon	high	30	60	30	60	60	60	30	
16:00-19:00	end of office	moderate	10	20	10	20	20	10	10	
19:00-24:00	end of day	low	6	12	6	12	12	6	6	

Table 2: Average number of active customers, ν_j, in different periods over a typical business day

We divide a day into different intervals depending upon the intensity of usage of applications, with periods of low, moderate and heavy activity. We capture this by taking different arrival rates of customers in different periods over a typical business day given in Table 2. Observe that interval from 0:00-8:00 (night time) has very low activity, 8:00-9:00 (start of business for some) has moderate activity, 9:00-12:00 (start of business for most) has high activity, 12:00-14:00 (lunch time) has low activity, 14:00-16:00 has high activity, 16:00-19:00 (end of day) has moderate activity, 19:00-24:00 is the end-of-day period of low activity.

Users in different classes have different sensitivities to prices and service qualities. Given the three user types and the 7 ERP modules, there are a total of 21 demand classes. The customer arrival rates of each are given by:

$$\tau^t_1 = x^t_H \psi^t_{Hh} \nu^t_H, \qquad \tau^t_2 = x^t_H \psi^t_{Hm} \nu^t_H, \qquad \tau^t_3 = x^t_H \psi^t_{Hl} \nu^t_H,$$
$$\vdots$$
$$\tau^t_{19} = x^t_P \psi^t_{Ph} \nu^t_P, \qquad \tau^t_{20} = x^t_P \psi^t_{Pm} \nu^t_P, \qquad \tau^t_{21} = x^t_P \psi^t_{Pl} \nu^t_P.$$

The job workloads are given by:

$$w_1 = w_2 = w_3 = w_H, \quad \ldots, \quad w_{19} = w_{20} = w_{21} = w_P.$$

5.2 Analysis of computational results

It is well-known that, in general, increasing the number of service classes increases the revenue that can be gained, but with a decreasing marginal gain; i.e. that the optimal revenue is concave increasing in the number of price-service classes. The experimental results of this section confirm that theoretical result. In particular, from having a single price-service class to a maximum of four, the revenue increased up to 25% when capacity is high, and 23% when capacity is low. The greatest jump in revenue is when the service provider goes from 1 to 2 price-service classes open, which ranged from 12-14% in the three capacity scenarios, high to low. The jump in revenue was 3–8% when a third class was added, and a further increase of 4-6% in revenue could be obtained by going from three to four price-service classes open.

In the figures 5–7, one can see the effect of permitting an increasing number of price-service classes to be made available. In these experiments, the average competitor's price is set to $50 and maximum delay bound set to 5, i.e., 5 times the minimum CPU time needed to execute the job. These values can be interpreted as a medium to low quality of service and a relatively low price. Note that the left-hand y-axis in units of price, while the right-hand y-axis is in units of the maximum delay bound. In this set of figures, the provider's total capacity is quite high.

In Figure 5, up to two service classes can be opened, and indeed at all but the nighttime low period (the 7th time period), two distinct classes are open in the optimal solution. The optimal price for class 1 is considerably higher than the average competitor's price and the corresponding service quality is significantly better. Clearly, then the optimization model selects class 1 to offer to the high-end customer, i.e. the customer with important and time-sensitive functions. On the other hand, the price of class 2 is some of the time also above that of the competitor, and the service quality mirrors the price trend, either above or below that offered by the competitor. It is of interest to note how the optimal price-service offering varies with the demand and time of day. In particular, during the 5th and 6th time periods, the promised service level of the high-priced class degrades; this can be explained by the fact that the 5th time period is one of heavy demand, but also that it is towards the end of the day. Since jobs can be processed after their arrival time into the system, there is an accumulation of jobs towards the end of the day, which includes those submitted at earlier times. Hence, even though time period 16:00-19:00 has only moderate arrivals, the jobs sent earlier and processed in that time slot make it another heavily-subscribed time period.

In Figure 6, the scenario is the same, with the exception that $K = 3$, i.e., the model allows up to three price-service classes to be opened. From the figure, one observes that

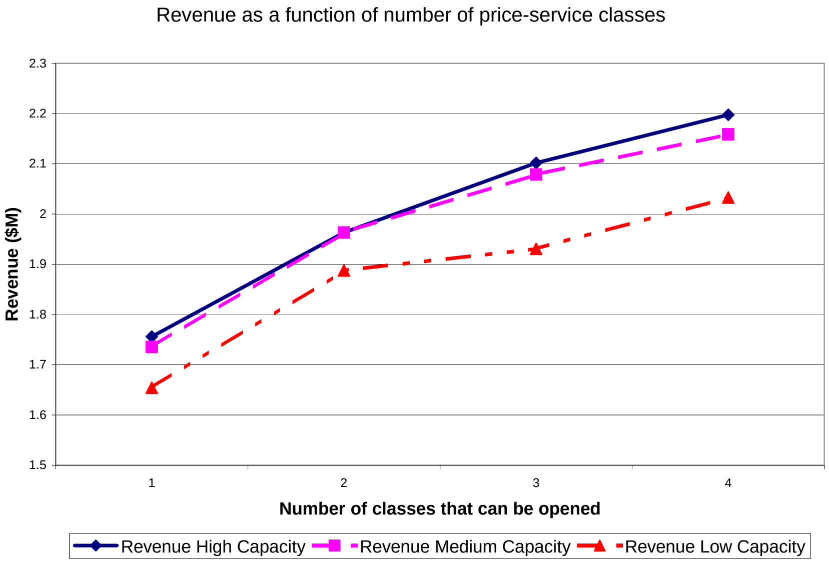

Figure 4: The revenue (less expected penalty costs) is concave increasing in number of classes that can be opened, with little difference across the three total capacity levels: high, medium, and low

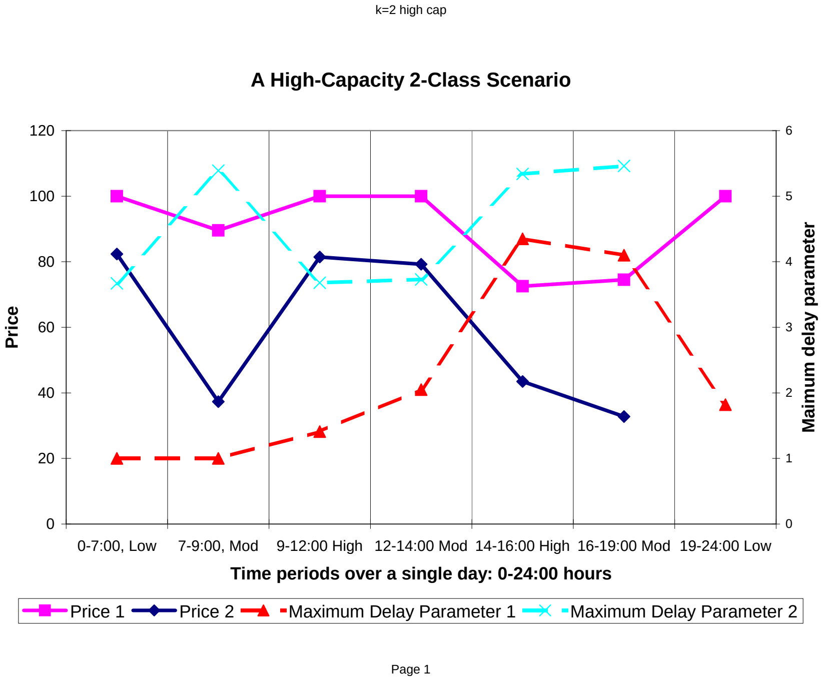

Figure 5: Scenario in which a maximum of two price-service classes can be opened, i.e., $K = 2$. In this case capacity is high.

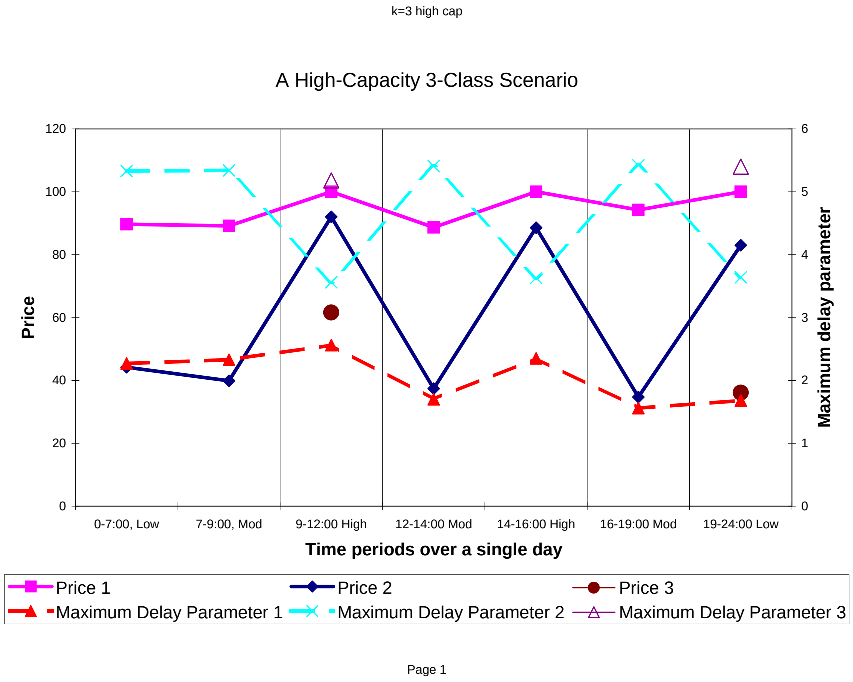

Figure 6: Scenario in which a maximum of three price-service classes can be opened, i.e., $K = 3$. In this case capacity is high.

three distinct classes are opened in only two time periods: the first high demand early morning period, and the last period. The first class is again the high-end price-service class, with less variation across time periods than what was seen in the previous example with two price-service classes. The second class now has some variation of the course of the day, with once again average price and service level values over the day approaching that of the competitor. The third price-service class takes the role of a class very close to the competitor's in the time slots for which the second class offers significantly better service, at a higher price.

Figure 7 shows now the same scenario with a maximum of five price-service classes that the service provider may open. We see that the model chooses not more than three distinct classes. In fact, because of the form of the model, all classes are open but the additional classes reduce to a maximum of three. The third class plays the same role as in the previous

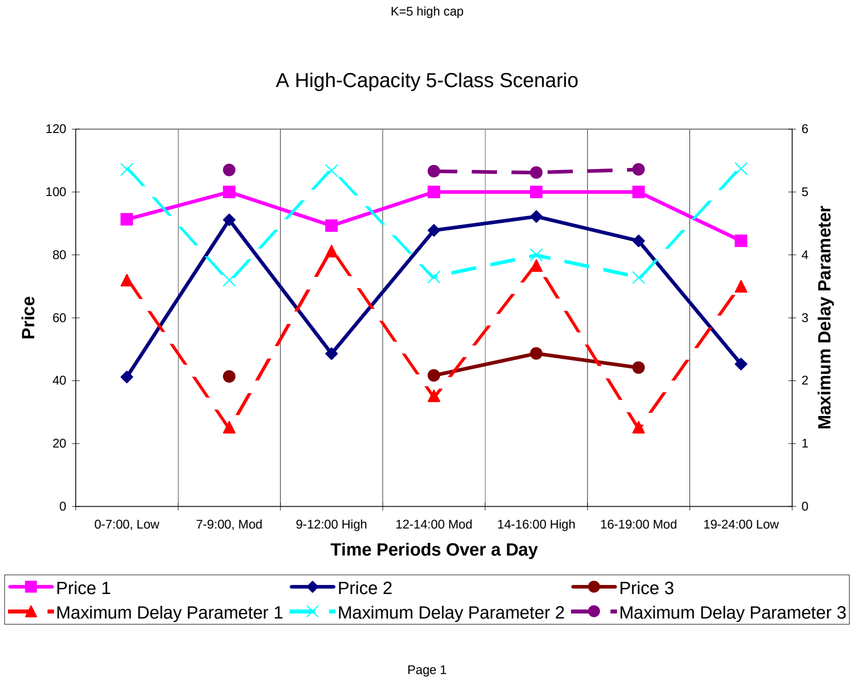

Figure 7: Scenario in which a maximum of five price-service classes can be opened, i.e., $K = 5$. In this case capacity is high.

example; it offers an option close to that of the competitor when the second class is set closer to the high-end class.

The total capacity of the service provider has an impact on the price-service classes offered in the optimal solution. Figure 8 illustrates a scenario with a maximum of four open classes, $K = 4$. As before, only three distinct classes are proposed by the optimal solution; however, their structure is somewhat different than was the case with a high total capacity.

Recall that when the total capacity of the service provider is low relative to the market demand, the service provider benefits by satisfying high-paying customers rather than by increasing market share through offering more classes; at the same time, in order to offer low delay bounds to those customers, fewer customers will be admitted into the system. In particular, observe from Figure 8 that at the second peak period, 14-16:00, only a single

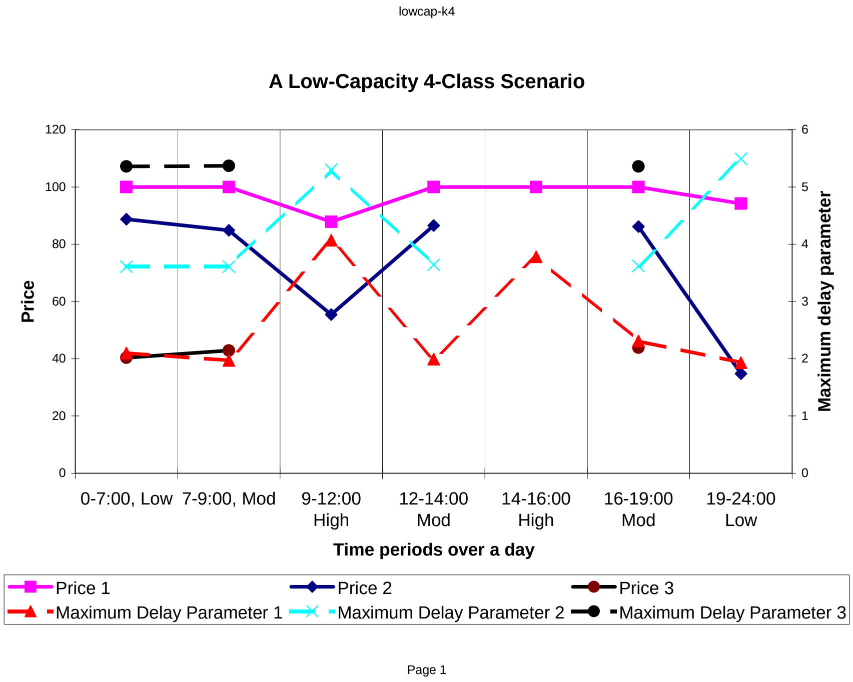

Figure 8: Scenario in which a maximum of four price-service classes can be opened, i.e., $K = 4$. In this case capacity is low.

class is opened; note though that the price is set very high–to \$100, which is, in fact is the upper bound on price. Service quality is still reasonably good, at 3.78. In other words, because of scarce capacity at a high-demand time slot, the solution suggested by the model is to go with a single high-price offering and maximize revenue on the price, rather than by offering a palette of service levels. This result is typical of the results obtained in the low capacity scenarios.

It is also important to examine the effect of a revenue management framework on the demand captured by the service provider, as a fraction of the total market demand. In Figure 9, two curves are presented: in one the service provider's capacity is quite high, and in the other the capacity is insufficient to handle the customer load that would like to be served.

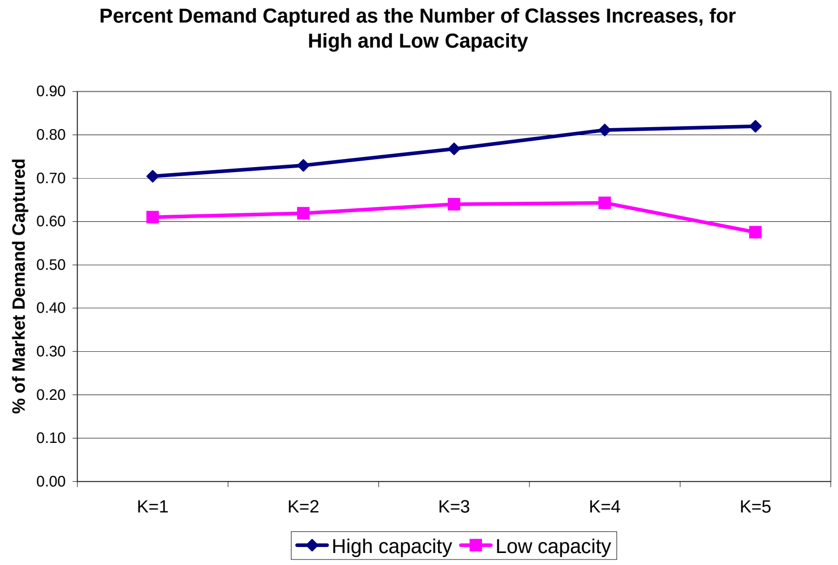

Figure 9: Fraction of market share captured by the service provider, as a function of the number of classes that can be opened and the time of day. Capacity is sufficient to accommodate as much demand as can be captured by judicious price-service level offerings.

When the service provider's total capacity is insufficient to handle all the demand that would otherwise patronize the provider, the optimal solution favors, in some cases, reducing the number of price-service classes and offering primarily those with a better quality-of-service and a higher price. Since capacity is linked to service quality through the explicit queueing relations, offering a better quality-of-service amounts to admitting less customers when total capacity is low. This is what we observe in Figure 9: whereas the market share increases in general with increasing number of service classes, as expected in the absence of capacity constraints from, e.g. Proposition 3, when capacity constraints are active, this need no longer be the case. Indeed, the market share obtained when capacity is low actually decreases from 4 to 5 price-service classes, although as we saw in Figure 4, revenue continues to increase.

What one can take away from this suite of computational experiments is, on the one hand, that optimal price variation across the time horizon is significant, i.e., dynamic pricing does appear to be advantageous to the service provider. There are clearly demand patterns and available capacity levels for which more than one price class significantly improves revenue, and the prices themselves vary considerably. It is also important to observe, on the other hand, that service level guarantees are a critical element to optimal market segmentation. Price variation alone would impact demand for the e-services, but in the absence of taking into account its effect on service quality, the market would re-adjust and the benefits of the price segmentation would not be achieved. One can observe from this suite of experiments that limited capacity impacts considerably the optimal structure of price-service level segmentation; in particular, when capacity was low it was advantageous to offer fewer, higher-quality classes. This result would not have been visible were the explicit relation between demand, capacity and service level not included in the model.

6 Conclusions and Future Research Directions

In this paper, we presented a revenue management framework for e-services provision, also known as IT on demand, e-business hosting, software as a service, etc. An important property of the framework is that it includes joint pricing and quality-of-service (QoS) decisions along with multiple demand classes, and the QoS is dependent explicitly on the number and types of customers served. Indeed, while it is usual in revenue management models to explicitly relate customer behavior to price, we argue that going forward it is equally important to treat more than just the price decisions; in treating service-level decisions for resource allocation, for example, one must model the dependences explicitly as well.

Properties of the model discussed in the paper include its non-convexity, but at the same time the empirical observation shows that the objective function surface appears to be changing shape rather slowly and in a limited range. In other words, the model appeared to be relatively insensitive to starting point. A more thorough theoretical assessment of this appearance of epsilon-optimality would be of great use for this class of models.

With respect to the use of the revenue management framework presented here for e-services, or IT on demand, a thorough set of experimental tests confirmed what theory and intuition had put forth. Namely, that an increasing number of price-service classes available increases revenue significantly (up to 25% from one to four service classes open), but the jump in revenue decreases as the number of such classes increases; i.e., marginal benefits are concave increasing in the number of price-service classes. Furthermore, it was shown that the number of distinct classes tended to three in most instances, while up to five distinct price-service classes could be opened. While explicit monetary costs of opening classes were not included here, its effect of accommodating a large market share on a finite service capacity was modeled, and the effect was to limit the number of open classes, and focus on pricing and serving the high-end customers.

In this paper, we assumed that the service capacity is given and fixed, and we did not consider the option of capacity expansion. However, capacity is the most costly investment in the IT service industry, and it should be carefully considered. An IT service provider may also lease contingent capacity itself, when facing dramatic increase of demand; hence, in that case the service provider itself would be an occasional customer of a different service provider. There would be a few ways of taking into account these interactions, one being a bi-level framework with the two service providers each at its own level, and another would be a peer-to-peer structure in which service providers share capacity as needed. In summary, two potentially valuable extensions of this work are to (i) incorporate explicit capacity provision decisions into the model, and (ii) to consider that the service provider itself may be a customer of additional capacity, as needed.

References

[1] Afeche, P. and Mendelson, H., 2004. Pricing and Priority Auctions in Queueing Systems with a Generalized Delay Cost Structure. *Management Science* 50, 869-882.

[2] Bakos, Y. and Brynjolfssonm E., 1999. Bundling Information Goods: Pricing, Profits and Eficiency. *Management Science* 49 (11), 1546-1562.

[3] Belobaba, P. P., 1987. Airline yield management: An overview of seat inventory control. *Transporation Science* 21, 63-73.

[4] Ben-Akiva, M. and Lerman, S., 1985. Discrete Choice Analysis. The MIT Press, Cambridge, Massachusetts.

[5] Bitran, G. and Caldentey, R., 2002. An overview of pricing models for revenue management. *Manufacturing Service Operations Management* 5, 203-229.

[6] Chen, P. and Wu. S., 2004. New IT Architecture: Implications and Market Structure. Workshop of Information Systems and Economics.

[7] Dewan, S. and Mendelson, H., 1990. User Delay Costs and Internal Pricing for a Service Facility. *Management Science* 36, 1502-1517.

[8] Dube, P., Borkar, V. and Manjunath, D., 2002. Differential Join Prices for Parallel Queues: Social Optimality, Dynamic Pricing Algorithms and Application to Internet Pricing. *21st Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Infocom)*, 276-283.

[9] Dube, P., Touati, C. and Wynter, L., 2006. Quality-of-Service and Price Wars in E-Markets. Submitted to *European Journal of Operations Research.* An earlier version appeared in Networking 2004, Lecture Notes in Computer Science, 442-453.

[10] Dube, P. and Hayel, Y., 2006. A Real-Time Yield Management Framework for e-Services. Proceedings of *IEEE Conference on E-Commerce Technology (CEC'06) and Enterprise Computing, E-Commerce and E-Services (EEE'06).*

[11] Dube, P., Hayel, Y. and Wynter, L., 2005. Analysis of a Yield Management Model for On Demand Computing Centers. *Journal of Revenue Management and Pricing* 4 (1), 24-38.

[12] Dube, P., Liu, Z. and Wynter, L., 2003. Yield Management for IT Provisiong. INFORMS Annual Meeting.

[13] Dube, P., Liu, Z., Wynter, L. and Xia, C., 2007. Competitive Equilibrium in e-commerce: Pricing and Outsourcing. To appear in *Computers and Operations Research (special issue on Operations Research and Outsourcing).*

[14] Elmaghraby, W. and Keskinocak, P., 2003. Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices, and Future Directions. *Management Science* 49, 1287-1310.

[15] Fishburn, P. and Odlyzko, A., 1999. Competitive Pricing of Information Goods: Subscription Pricing Versus Pay-per-use. *Economic Theory* 13, 447-470.

[16] Fulp, E. and Reeves, D., 2004. Bandwidth Provisioning and Pricing for Networks with Multiple Classes of Service. *Computer Networks Journal : Special Issue on Internet Economics - Pricing and Policies* 46 (1), 41-52.

[17] Gurnani, H. and Karlapalem, K., 2001. Optimal Pricing Strategies for Internet-Based Software Dissemination. *Journal of the Operational Research Society*, 52 (1), 64-70.

[18] Hampshire, R., Massey, W., Mitra, D. and Wang, Q., 2003. Provisioning for Bandwidth Sharing and Exchange. Telecommunications Network Design and Management, Ed. G.Anandalingam and S.Raghavan, Kluwer Academic Publishers, 207-225.

[19] Huang, K. and Sundararajan, A., 2005. Pricing Models for On-Demand Computing. Working paper. New York University.

[20] Kleinrock, L., 1975. *Queueing Systems Volume I: Theory.* John Wiley and Sons.

[21] Konana, P., Gupta, A. and Whinston, A., 2000. Integrating User Prferences and Real-Time Workload in Electronic Commerce. *Information Systems Research* 11, 177-196.

[22] Littlewood, K., 1972. Forecasting and Control of Passenger Bookings. Proceedings of *12th AGIFORS (Airline Group of the International Federation of Operational Research Societies) Symposium*, 95-128.

[23] Liu, Z., Squillante, M. and Wolf, J., 2001. On Maximizing Service-Level-Agreement Profits. Proceedings of *ACM Conference on Electronic Commerce*, 213-223.

[24] McGill, J. and van Ryzin, G., 1999. Revenue Management: Research Overview and Prospects. *Transportation Science* 33, 233-256.

[25] Mendelson, H., 1985. Pricing Computer Services: Queueing Effects. *Communications of the ACM* 28, 312-321.

[26] Mendelson, H. and Whang, S., 1990. Optimal Incentive-Compatible Prioriy Pricing for the M/M/1 Queue. *Operations Research* 38, 870-883.

[27] Odlyzko, A., 1999. Paris Metro Pricing for the Internet. Proceedings of *ACM Conference on Electronic Commerce*, 140-147.

[28] Paleologo, G., 2004. A Methodology for Pricing Utility Computing Services. *IBM Systems Journal* 43(1), 20-31.

[29] Paschalidis, I. and Tsitsiklis, J., 2000. Congestion-Dependent Pricing of Network Services. *IEEE/ACM Transactions on Networking* 8(2), 171-184.

[30] Petruzzi, N. and Dada, M., 1999. Pricing and the Newsvendor Model: a Review with Extensions. *Operations Research* 47, 183-194.

[31] Sundararajan, A., 2004. Nonlinear Pricing of Information Goods. *Management Science* 50, 1660-1673.

[32] Talluri, K. and vanRyzin, G., 2004. *The Theory and Practice of Revenue Management.* Kluwer Academic Publishers.

[33] Van Mieghem, J., 2000. Price and Service Discrimination in Queuing Systems: Incentive Compatibility of $Gc\mu$ Scheduling. *Management Science* 46, pp. 1249-1267.

[34] Viswanathan, S. and Anandalingam, G., Pricing Strategies for Information Goods. *Sadhana* 30 (2-3), 257-274.

[35] Wilson, R., 1993. *Nonlinear Pricing.* Oxford University Press, New York.

[36] Yuksel, M. and Kalyanaraman, S., 2002. Distributed Dynamic Capacity Contracting: A Congestion Pricing Framework for Diff-Serv. Lecture Notes in Computer Science, 198-210.

[37] Zwart, A and Boxma, O., 2000. Sojourn Time Assympotics in the M/G/1 Processor Sharing Queue. *Queueing Systems* 35, 141-166.