ADDITIONAL ACADEMIC PAPER

Pricing of shared computer services

Opher Baron,* Dirk Beyer and Gabriel R. Bitran*****

Received (in revised form): 3rd December, 2004

*Rotman School of Management, University of Toronto, 105 St George Street, Toronto, ON, Canada M5S 3E6
Tel: +1 416 978 4164; Fax: +1 416 978 5433; e-mail: Opher.Baron@rotman.utoronto.ca
**Applied Mathematical Modeling, Intelligent Enterprise Technologies Laboratory, Hewlett-Packard Laboratories, 1501 Page Mill Road, MS1140, Palo Alto, CA 94040, USA
Tel: +1 650 236 2711; Fax: +1 650 857 6278; e-mail: Dirk.Beyer@hp.com
***Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 53–355, Cambridge, MA 02139, USA
Tel: +1 617 253 2652; Fax: +1 617 258 7579; e-mail: GBitran@mit.edu

Opher Baron *is an assistant professor of operations management at the Rotman School of Management, the University of Toronto. He has a PhD in Operations Management from the MIT Sloan School of Management along with an MBA and a BSc in Industrial Engineering and Management from the Technion, Israel Institute of Technology. His research interests include applied probability, pricing and revenue management.*

Dirk Beyer *is a principal scientist in the Decision Technologies Department at Hewlett-Packard Laboratories in Palo Alto, CA. He holds a Master's degree in Mathematics and Physics and a PhD in Operations Research from Leipzig University, Germany. His research interests span a wide spectrum of business applications of operational research, including the application of analytical techniques to the design and management of computing infrastructure, supply chain management and customer relationship management.*

Gabriel R. Bitran *is Chair Professor at the MIT Sloan School of Management. He is a member of the Operations Management Group. Professor Bitran's research interests lie in the field of manufacturing, logistics and the service industry. More recently, he has been working on pricing for high-tech services, fashion retail goods and services, and design of bandwidth markets, as well as related revenue management problems.*

Journal of Revenue and Pricing Management, Vol. 4, No. 1, 2005, pp. 54–65
© Henry Stewart Publications, 1476–6930

ABSTRACT

Keywords: pricing of services, token bucket, regulated random walks, admission control

This paper proposes a framework for analysing admission controls as pricing schemes for shared services. Token bucket admission-control mechanisms are considered as pricing schemes. To analyse the buyer's problem of choosing optimal parameters for token bucket schemes, an important performance metric of token bucket mechanisms is considered, the long-run probability of being denied service, which is equivalent to the threshold-crossing probability of two-sided or one-sided regulated, random walks. For the buyer's problem, the paper gives approximations for these hard-to-calculate metrics, shows that the problem is convex under mild assumptions, and provides

a closed-form solution to an approximation of the problem when demand is normally distributed.

INTRODUCTION

Today, many businesses outsource their computing needs (web hosting, record keeping, databases) to external service providers, who sometimes install dedicated computer resources and charge for hardware, software and support for these resources. Leading computer companies, such as IBM, HP and Sun Microsystems, however, also provide computer services in a shared manner. Provisioning of services in a shared manner allows providers to increase usage rates and to deliver a better service at lower costs.

Pricing of shared computer services is challenging. When resources are shared, some of them might be idle but still require costly maintenance, a cost that should be shared between all buyers. Moreover, it makes sense to charge heavier users more than lighter ones, using some usage-based pricing mechanism. Providing service-level guarantees to buyers of shared computer services is an additional challenge, because the service level provided to buyers does not depend only on their individual usage. Thus, users might experience a lower service level than they expect or need. In the case of a dedicated resource, a low service level may be attributed to an insufficient amount (capacity) of a resource the buyer bought (or rented); however, in the case of shared service, such an explanation might be contentious. Therefore, shared service increases the responsibilities of the service provider. In traditional shared service environments, no service-level guarantees are given, but the service level provided is high enough to satisfy buyers. This is achieved by sellers over-provisioning their resources so that they can serve the maximum demand of all their customers. Clearly, such a solution is inefficient, particularly in computer services, where sharing resources is expected to reduce expenses. Sellers can prevent overloading the system by operating an admission control rather than by the over-provisioning of resources.

This paper addresses the pricing and admission-control problems related to the provision of shared services. Combining admission control with pricing could induce buyers to choose their usage levels carefully. The logic behind this claim is that admission controls accept or reject buyers' requests; thus they dictate the performance buyers see. Moreover, relating the service cost to the performance buyers require and experience is only fair. In additional, such a pricing policy could induce buyers to smooth their demand (eg shift demand away from peak periods in aggregate demand). Finally, if sellers face smoother demands, they can provide a higher level of service using few resources and therefore decrease cost.

The framework we are proposing consists of the use of a pricing and admission control mechanism to specify rules for coordination between buyers and the seller in settings of a non-cooperative game. Using these rules the buyers' and sellers' problems can be formulated as a leader-follower (Stackelberg) game with a sub-game perfect Nash equilibrium (Gibbons, 1992). The seller moves first and chooses prices and resource level to maximise profits; then buyers react to prices by changing their demand patterns and choosing the parameters of the admission control in order to minimise their expenditure.

Another well-studied application of shared resources is the management of air space and airport runways. In the USA, these resources are managed using the collaborative decision-making (CDM) process (eg Ball *et al.*, 2000). In the CDM, airlines, airports, and the Federal Aviation Admin-

istration share information and develop rules for online traffic flow management using the framework of cooperative game theory (Herve, 1995).

Traditional pricing problems, as surveyed by Bitran and Caldentey (2003) and McGill and van Ryzin (1999), focus on pricing 'tangible' products to a large number of customers. Maglaras and Zeevi (2003) discuss pricing for a service provided to a large number of customers using a shared resource. Their main result is that, for economic reasons, the operation of a shared pool of resources should be managed by a heavy traffic regime.

The pricing problem sellers face when they consider the provisioning of shared computer services focuses on pricing a service for a relatively small number of customers; thus, it differs from traditional pricing problems. A small number of customers allows fine market segmentation by provisioning a highly tailored service. The token bucket (TB) admission controls (Berger, 1991) use two parameters to tailor service level to different customers and are well known in the computer networks and telephony literature.

This paper presents a framework for operating admission controls as pricing schemes for shared resources and demonstrates part of this framework using TB admission controls as pricing schemes.

The paper is divide into sections as follows: an overview of the literature of pricing of computer services; a framework for the analysis of admission controls as pricing schemes; two TB pricing schemes; an analysis of TB pricing schemes; a general solution to the buyer's problem when facing these pricing schemes; special-case closed-form approximations for normal demand cases; and a summary of the paper and future research directions.

For the sake of brevity, proofs of the results are not included and are available from the first author on request.

PRICING OF SHARED COMPUTER RESOURCES

Pricing computer resources in the literature

The vast majority of papers on the pricing of computer resources deals with the pricing of bandwidth usage. This literature is extensive, and only a few relevant examples are mentioned here. Shenkar *et al.* (1995) give an important review of the main ideas in pricing in computer networks.

Congestion pricing

Congestion pricing has attracted a fair amount of research (Kelly and Tan, 1998; MacKie-Mason and Varian, 1994a; and references therein). The main goal of congestion pricing is to reduce the congestion on the internet, and it can probably be generalised to managing usage levels of other resources.

The advantages of congestion pricing are twofold. First, because customers are charged a higher price when resources are busy, they are expected to balance their demand and, by doing so, to help the seller to satisfy the required quality of service using fewer resources. A second advantage results from an economic analysis of congestion pricing schemes, where the 'right price' for the resource usage will be charged and the 'right amount' of investment in increasing resource capacity will be induced. MacKie-Mason and Varian (1994b) justify this logic.

A major disadvantage of dynamic pricing for shared resources, however, is that it prevents buyers and sellers from predicting the cost (or revenue) during a period — predictions that are valuable for both parties. Moreover, it requires buyers to continually monitor their usage and preferences.

Non-congestion pricing

An exception in the literature for internet pricing schemes is the flexible service plan

(Altmann and Chu, 2001). This scheme, commonly used in pricing of cell phone service, suggests charging a fixed price for a fixed amount of bandwidth and allowing end users to purchase higher bandwidth on demand, for an additional cost (according to the number of bits sent or minutes used).

Pricing computer resources in practice

This section presents the results of a field study, reported in Baron, 2003, that took place between November 2001 and January 2002. This study included interviews with practitioners and academics.

Pricing for companies typically includes a setup cost, a fixed monthly fee (for a fixed amount of usage and the operation of the dedicated hardware). For a web connection, which is provided in a shared manner, there is an additional cost component for quantities over the fixed amount of usage agreed on.

Two methods are used to measure the monthly usage level. The less common one, used by 5–15 per cent of sellers, is to measure the total usage during the month. The main drawback of this measure is that it ignores the usage variation and therefore complicates the tasks of giving service-level guarantees and of sellers' resource planning.

The more common practice is known as 95/5 pricing. In this method, a month is divided into intervals of five minutes. The usage (number of packets sent) in each interval is recorded and, at the end of the month, the different usage records are ordered from the lowest to the highest. The monthly usage is charged according to the 95th percentile of the usage. The 95/5 scheme gives buyers an incentive to decrease their usage deviation. It still does not bound the usage a buyer can ask for in a period, however, and does not simplify the tasks of providing service-level guarantees and of sellers' resource planning.

Discussion

The pricing schemes in common practice are simple and were adequate to start up the industry. As the industry matures, however, some more sophisticated pricing schemes are needed, especially in the market considered: selling services to large companies. But dynamic pricing seems too cumbersome to be implemented as pricing for shared computer services at this time.

Provisioning of shared computer resources means that the service level provided for each user does not depend only on his or her individual usage. In order to allow sellers to provide the required service level, sellers need to operate an admission control, so that they can prevent one buyer from consuming all of the resources (and thereby blocking other buyers) in any given time period. Moreover, some of the useful theoretical properties of the pricing schemes discussed in the literature are related to the fact that there is an admission control combined with them.

Therefore, an attractive way to induce buyers not to overuse resources would be to combine pricing and admission control.

FRAMEWORK FOR IMPLEMENTING ADMISSION CONTROLS AS PRICING SCHEMES

A framework is presented for coordination between buyers and a seller of a shared resource in a free market using an admission control that is also a pricing scheme. The buyer's demand is considered an exogenous input to the model. It is assumed that the price charged buyer depends on some parameters of the admission control. As buyers increase their payments, they increase the acceptance of their jobs by the admission control, ie the resource availability (or service level) they experience. There are four main sub-problems:

1. *The buyer's problem*: It is assumed that the buyer wishes to minimise expendi-

ture, subject to some service-level constraint. The buyer's demand is stochastic, and therefore the service-level requirement will be probabilistic in nature. Thus, the buyer tries to balance the risk of overpaying for the admission control with the risk of losing or delaying jobs.

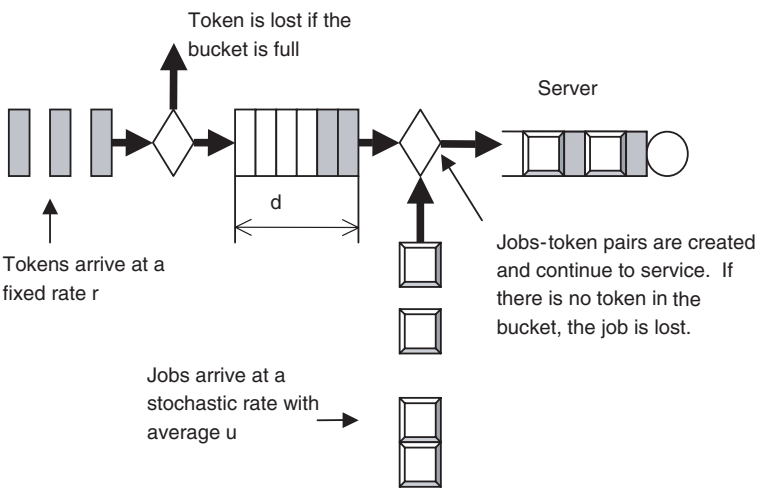
2. *Performance of the admission control*: To solve the buyer's problem, the buyer needs to know the resource availability as a function of his/her demand and choice of admission-control parameters. It is helpful to assume that any demand that is accepted by the admission control will be processed by the seller. (An alternative approach is to assume that the seller can only guarantee a specified service level to any demand that is accepted. Then, the buyer's service-level requirement would already consider this degraded performance.)
3. *The seller's pricing and resource planning problem*: It is assumed that the seller wishes to maximise profits or revenues. Even if the seller's resource planning is ignored (constraining the study to revenue maximisation), the seller's problem is not trivial. While in most revenue management problems the buyers' reactions to prices are simple, with this coordination method, buyers are strategic players (as they minimise their expenditures). Since optimal prices are a function of the demand faced by the seller, an essential input to the seller's problem is the output from the admission control.
4. *The output stream from the admission control*: Based on the seller's pricing problem, the characterisation of the output stream from the admission control is to find the probabilistic nature of this stream in steady state. This problem can be divided into two: characterising demand from an admission control of one buyer and characterising aggregation of buyers' demand (the total demand faced by the seller).

THE TOKEN BUCKET

Token bucket admission controls

One common admission control in the network and telephony literature is the TB

Figure 1: Diagram of the token bucket control mechanism



method (Berger and Whitt, 1994; Kelly and Tan, 1998). It uses two parameters to define the demand for a network's resources: the token rate, denoted by r, and the bucket depth, denoted by d. Every source gets tokens at a rate r (not necessarily an integer) and has to have a token in order to send a packet. Thus, it is assumed that a packet sent without a token is lost. Unused tokens can be accumulated until level d is reached, and if a token arrives at a full bucket, it is lost. Figure 1 illustrates this control mechanism.

The TB mechanism is interpreted as an admission control for any resource. A *work unit* is defined as analogous to a packet. Each token represents such a unit, and a job requiring processing of a few work units is analogous to a file. A buyer who sends jobs to be processed is analogous to a source that sends files. Finally, a *basic time period* is established in which tokens enter the bucket. It is important to choose this time period such that the processing of any job takes no more than one period.

Another version of the TB mechanism adds a job buffer that allows jobs to wait in it until the arrival of tokens. This practical extension, however, does not change the analysis of the service level, because the probability of loss in a system with job and token buffers depends only on the combined capacities of both buffers, as proved in Berger's (1991) Theorem 1.[1] Extending this idea, consider a control where the job queue is infinite, so when the bucket is empty, jobs are backlogged. This admission control is called a TB with rate control (TBwRC) (Berger, 1991).

Both TB admission controls monitor the mean (or total) demand, using the token-rate parameter, and the variability in demand, using the depth parameter. Therefore, a user who submits large jobs, but only occasionally, can request a low rate but a big bucket, while a user with low demand variability can request a small bucket.

Token bucket admission controls as pricing schemes for shared resources

The application of TB admission controls in the context of pricing shared resources (Baron, 2003) has some desirable properties: they allow for flexibility in the definition of the demand process and for differentiation between customers. Moreover, they are helpful in provisioning service-level guarantees.

To implement TB pricing schemes[2] for shared resources, the seller sets the price per token, denoted by R, and a price per rental of token storage space (bucket depth), denoted by D. Note that there is no point in purchasing any depth if $R \leqslant D$.

The sequence of events is as follows: at the start of each period the bucket level is increased by r tokens; then usage occurs and tokens are consumed up to their number in the bucket. If tokens remain, up to d of them are carried over to the next period. Note that the bucket can hold temporarily up to $r+d$ tokens, but it can carry over at most d tokens to the next period. In the TB case, demand that exceeds the number of tokens in the bucket is lost. In the TBwRC case, excess demand is backlogged, which is represented by a negative bucket level.

ANALYSIS OF TOKEN BUCKET ADMISSION CONTROLS AS PRICING SCHEMES

The remainder of the paper focuses on solving the performance of the admission-control problem and the buyer's problem, ie the problems 1 and 2 above, when using TB admission controls as pricing schemes. Problems 3 and 4 are outside the scope of this paper and are left for future research.

It is assumed that the usage U_i in period i forms an independently and identically distributed (*i.i.d.*) sequence of non-negative random variables with a cumulative distribution function (CDF)[3] $F_U(u) = P(U \leqslant u)$. Let $G_U(s) = E_U(e^{sU})$ be its moment-

generating function, which is assumed to exist in a neighbourhood of $s=0$.

The assumptions on the usage process outlined above are rather restrictive. Nonetheless, for aggregate usage over many users (if the buyer is a large company) and sufficiently large time periods, they may be defendable as an approximation.

Performance of token bucket admission controls

Starting with a full bucket in period 0, the bucket level at the beginning of period $i+1$ in the TB case (excess demand is lost) is

$$\begin{aligned}\tilde{L}_o &= d \text{ and } \tilde{L}_{i+1} \\ &= \min\left\{d, \max[0, \tilde{L}_i - (U_i - r]\right\} \text{for}, i \\ &0, 1, \infty\end{aligned} \tag{1}$$

Similarly, the bucket level at the beginning of period $i+1$ in the TBwRC case (excess demand is backlogged) is

$$\begin{aligned}L_0 &= d \text{ and } L_{i+1} \\ &= \min\{d, L_i - (U_i - r)\} \text{ for } i \\ &= 0, 1, \infty\end{aligned} \tag{2}$$

Define $X_i = U_i - r$, and let $F_X(x)$ and $G_X(s) = E_X(e^{sX})$ be the CDF and moment-generating function of x, respectively. Then it is assumed:

Assumption 1. $E(X_i) < 0$ and $F_X(0) < 1$.
Assumption 2. Let $s^\star \equiv \arg\{G_X(s) = 1 \mid s^\star > 0\}$, which is called the conjugate point of X, and is assumed to exist.

Assumption 1 implies that $E(U) < r$ and that r is not larger than the maximal possible demand per period. Assumption 2 states that X satisfies a large deviation principle and holds for many commonly used distributions (eg normal). During the rest of the paper, these assumptions hold unless otherwise stated.

In what follows, we denote by L and $\tilde{L}$ the steady-state processes of the TB and TBwRC levels, respectively. Substituting x_i into (1) shows that the bucket-level process $\tilde{L}$(TB) is a two-sided, regulated, random walk with a positive drift. A similar analysis shows that L (TBwRC) is a positive-drift, one-sided, regulated random walk.

Define the service level for the TB case as 'percentage of periods with loss' and for the TBwRC case as 'percentage of periods with backlogs' or, more concisely

$$P(L \le 0) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(L_i \le 0) \tag{3}$$

and

$$P(\tilde{L} = 0) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\tilde{L}_i = 0) \tag{4}$$

where $I\{\cdot\}$ is the indicator function with value one if the event happens and zero otherwise.

If the demand in each period is a continuous random variable, the service-level definitions are equivalent to the traditional ones. Thus, one can write the service-level requirement for the TB case as $P(\tilde{L} = 0) = 1 - SL^{TB} \le 1 - \alpha$, where α is the minimum required service level, and the service-level requirement for TBwRC as $P(L \le 0) = 1 - SL^{TBwRC} \le 1 - \alpha$.

Note that from (1) and (2) it is seen that; $P\{\tilde{L}=0\} \le P\{L \le 0\}$ this is also discussed by Paschalidis and Liu (2003).

Using (1), SL^{TB} can be interpreted as the event that a two-sided, regulated, random walk reaches a threshold; using (2) SL^{TBwRC} can be interpreted as the event that a one-sided, regulated, random walk crosses a threshold. There is extensive literature providing bounds and asymptotics to the threshold-crossing probability of regulated random walks under Assumptions 1 and 2 (eg on SL^{TB} see Glasserman, 1997; Ross, 1974; on SL^{TBwRC} see Baron

2005). Thus, these bounds will be used as surrogates for the performance of TB admission controls.

THE BUYER'S PROBLEM

This section analyses the buyer's problem when facing TB pricing schemes. It is assumed the buyer wishes to minimise his/her expenditures, subject to a constraint on the percentage of periods with shortages. The buyer's problem addresses a traditional problem in implementing TB admission controls: choosing parameters that will result in satisfactory performance. Moreover, it is assumed buyers want to minimise the cost of this rate and depth choice; thus, this problem requires the prices of the depth and rate as inputs.

The buyer's problem (BP) in the TB case is

$$\min_{d,r}(Dd + Rr) \qquad \text{(BP)}$$

s.t.

$P(\tilde{L} = 0) \leq 1 - \alpha$ (service-level constraint)
$r \geq 0$ (initial rate constraint)
$d \geq 0$ (depth constraint)

Note that, for a moment, Assumption 1 that $E(U) < r$ is ignored. Replacing the service-level constraint with $P(L \leq 0) \leq 1 - \alpha$ gives the buyer's problem for the TBwRC case.

The buyer's problem is an example of a 'random walk optimal control problem', ie there is a constraint on a performance measure of a random walk, and controls are the drift of the random walk and its range (its regulators); thus, it is applicable in additional settings.

To solve the BP problem, the buyer needs to know the percentage of periods with shortages as a function of demand and choice of TB parameters. Given the discussion earlier, one can use known results to express the steady-state loss (or backlog) probability of TB controls. An additional source of complexity in solving the buyer's problem is that the buyer's controls are both the token rate and the bucket depth (ie the drift and the threshold not to be crossed by the random walks). Glasserman (1997) also investigated a random walk optimal control problem. He considered additional performance measures but only used the threshold as a control. Moreover, he confined his work to the backlog case (the TBwRC).

Two trivial solutions to the buyer's problem follow.

1. *Depth non-negativity constraint is active ($^{d} = 0$)*: this happens when the ratio between the depth price and the rate price, D/R, is large (ie $D/R \uparrow 1$); thus, there is no point in purchasing any depth. In such a case, any period with demand higher than r results in a loss, and r should be chosen as the α percentile of the usage-level distribution, ie, $r^* = F_U^{-1}(\alpha)$, with a cost $Z^* = RF_U^{-1}(\alpha)$.
2. *Rate non-negativity constraint is active ($^{r} = 0$)*: this might happen when the ratio D/R is small. For the TBwRC, however, the rate constraint is $E(U) < r$ in order for a steady-state distribution to exist (or else from some point of time all jobs will be backlogged). Moreover, it is claimed that for a 'reasonable' pair of service-level requirement and demand, the real constraint on r for the TB case is, as in the case of a TBwRC, $E(U) < r$. Thus, the rate constraint is replaced by $r \geq E(U)$ (final rate constraint)

Analysis of the buyer's problem

Here, a constrained version of the buyer's problem is solved in both the TB and TBwRC cases. Theorem 1 states that this constrained version is convex, thus it is easy to solve numerically.

Upper bound on the buyer's cost

For token bucket with rate control

With SLTB $\geqslant$ SLTBwRC, and using $1 -$ SLTBwRC $\leqslant \exp[-s^{\star}(r)d]$ (see Gallager, 1996, Chap. 7), the service-level constraint is replaced with

$$\exp[-s^{\star}(r)d] \leqslant 1-\alpha \qquad \text{(BCP)}$$

Based on (BCP), one has

$$d^{*} = -\frac{1n(1-\alpha)}{s^{*}(r^{*})}$$

which can be substituted into the objective function, creating an optimisation problem that is named the BCP problem with r as its single control.

Theorem 1. The BCP problem

$$\min_{r} Rr - D\frac{1n(1-\alpha)}{s^{*}(r^{*})}$$

is convex with respect to r.

Using Theorem 1 the BCP can be easily solved numerically.

Notice that the BCP depends on the ratio D/R and not on their actual values and on the service level via a multiplication by the log of one minus the shortage probability. Therefore, any optimal solution to the BCP would have the same dependencies.

Observe that the solution to the buyer's problem will be given by the solution to the BCP problem as long as its resulting cost is lower than the cost of the trivial solution given by setting $d^{*} = 0$, $r^{*} = F_{U}^{-1}(\alpha)$, with a cost $Z^{*} = RF_{U}^{-1}(\alpha)$. In these cases, that trivial solution is the *optimal* solution to the buyer's problem, since it uses no approximations.

Despite the convenient form of the BCP problem, when the demand distribution is known, better bounds for the buyer's cost can be found, as the next section demonstrates for cases of normal demand.

For the token bucket

From the results of Baron (2005), one can bound the loss probability

$$P(\tilde{L}=0) \leq \left(\bar{F}_{U}(r) + \frac{1}{E(T)}\right)P(L \leq 0) = A \cdot P(L \leq 0)$$

where $E(T)$ can be approximated when r is known, and A is introduced for notational convenience. One can combine this with the bounds on the service level in the TBwRC, as given in (BCP), to replace the service-level constraint. The constant A, however, depends on the tokens' rate in a complicated manner. Therefore, the TB approximation procedure is proposed. First, find the optimal rate to BP by solving the BCP problem and approximate the constant A given this rate. Secondly, if A is larger than one, take the solution of the BCP problem; otherwise, solve the BCP problem again with a relaxed service-level constraint of $P(L=0) \leq (1-\alpha)/A$.

SPECIAL-CASE SOLUTION FOR THE BUYER'S PROBLEM FOR NORMAL DEMAND CASES

Closed-form solutions are presented for the values of r and d in the case of a TBwRC when buyer's demand is Normal distributed. These solutions are better than the ones achieved solving the BCP problem, as they use more information regarding the demand process. A numerical simulation confirmed that costs resulting from the solutions are close to optimal costs. Detailed simulation results are available from the first author on request.[4] Consider normal demand when the ratio between the mean demand and its standard deviation is such that the probability of negative usage is negligible.

For the token bucket with rate control

The TBwRC approximation

If the buyer's demand in each period is *i.i.d.* and normally distributed $U \sim \text{Normal}(\mu, \sigma^2)$, an approximate solution to the buyer's problem is

$$r_A^* = \mu + \frac{\sigma\sqrt{2}C}{2}$$

$$d_A^* = -\frac{\sigma 1n(1-\alpha)}{\sqrt{2}C} - 0.583\sigma$$

and

$$Z_A^* = R(\mu + \sigma\sqrt{2}C) - 0.583D\sigma$$

where $C = -D\ln(1-\alpha)/R$.

When demand is $U \sim \text{Normal}(\mu, \sigma^2)$ the conjugate point is $s^* = -2(r-\mu)/\sigma^2$, and one can use Siegmund's (1985) technique to approximate the threshold-crossing probability of a one-sided, regulated, random walk by

$$P(L \leq 0) \approx e^{-s^* d} V(\mu - r, \sigma)$$

where $V(\mu, \sigma)$ can be approximated as $e^{-0.583\frac{2\mu}{\sigma}}$, an approximation that becomes more accurate as μ/σ decreases. From (5), the service-level constraint that is always active in the BCP is expressed

$$\exp - \left[2\frac{(r-\mu)}{\sigma^2} d \; + 0.583\left(2\frac{r-\mu}{\sigma}\right)\right] = 1 - \alpha$$

Solving this for d and substituting it in the objective function of BCP yields, one gets

$$Z(r) = Rr + \frac{RC\sigma^2}{2(r-\mu)} - 0.583\sigma$$

It is easily verified that $r = \mu + \frac{\sigma\sqrt{2}C}{2}$ satisfies the required first and second-order conditions for optimality. Then the optimal depth and cost can be calculated.

An *upper bound* based on replacing the service level constrained with (BCP) can be obtained in a similar manner, and it is omitted. Note, however, that the difference between the solution based on the upper bound and the solution based on the TBwRC approximation is the correction factor -0.583σ in the depth parameter. Thus, the depth value of the TBwRC approximation is smaller and might be negative (when depth is much more expensive than rate), which is infeasible for the original problem. In such cases, the TBwRC approximation to the buyer's problem is the one for the case $d = 0$.

The insights provided from both the TBwRC approximation and the upper bound are: First, when service level α increases, the rate and depth parameters (and the cost) are increasing, proportionally to $\ln(1-\alpha)$. Secondly, when the cost ratio D/R increases, so does the optimal token rate, whereas the depth decreases.

A final insight is that the optimal rate linearly increases with both the mean and the standard deviation of demand, and the optimal depth linearly increases with the standard deviation of demand. Technically, these dependencies can be used to update the rate and depth parameters when demand changes, and can be helpful when solving the BCP problem for general demands. Conceptually, both costs and revenue increase linearly with the mean and the standard deviation of demand. The rate of increase due to the standard deviation is higher, however. This representation of costs and revenues is prevailing because it is fair and simple (it might be simpler than the 95/5 mechanism used in practice); thus it can help in the dissimilation of TB pricing schemes.

For the token bucket

As show in Baron (2005) the loss probability when demand is normal can be approximated by

$$P(\tilde{L}=0) \approx \frac{1+\bar{F}_U(r)\left[\frac{\sigma}{(r-\mu)^2\sqrt{2}\exp\left(-\frac{1.166(r-\mu)}{\sigma}\right)}-1\right]}{\frac{\sigma}{(r-\mu)^2\sqrt{2}\exp\left(-\frac{1.166(r-\mu)}{\sigma}\right)}-1}P(L\leq 0) = A\cdot P(L\leq 0)$$

where A is introduced for notational convenience. When demand is normal, the value of A can only be found numerically; however, one can use the approximation of the TBwRC, as given in (5), to express, in closed-forms, A and $P(L\leqslant 0)$ as functions of the service level (and the problem's parameters). A tailored TB approximation is proposed. First, solve $1-\alpha = AP(L\leqslant 0)$ for the service level in the TBwRC case, and denote this solution by β. This service level is the guess for the worst performance one can allow in the TBwRC while maintaining the required performance in the TB case. Secondly, if this service level is higher than the requested one (this can happen if the approximation for A is not between zero and one), take the solution of the TBwRC approximation; otherwise, use the solution obtained from the TBwRC approximation for a problem with a relaxed service-level constraint of $P(L=0)\leqslant(1-\beta)$.[5]

CONCLUSION

Provisioning of shared computer services receives much attention in industry. The practice of managing shared resources is not systematic, however, partially because the pricing of shared resources is not used to influence buyers' demand. A possible improvement, which might be accepted by both sellers and buyers, would be to use admission controls as pricing schemes.

The implementation of TB admission controls as pricing schemes is still challenging; it requires attention to additional problems, such as the seller's problem, the characterisation of the output process from a TB, and pricing of multiple resources, etc. Another important problem related to the implementation of TB pricing schemes is a secondary token market operated by the seller. In such a case, buyers could purchase tokens (to fill their bucket) at a spot price, whenever their bucket level is low or when they predict a high usage level. The authors believe that operation of such a token market is necessary. A spot market, however, adds a portfolio management aspect to the buyer's problem, and therefore complicates it.

Despite these shortcomings, the use of TB admission controls as pricing schemes has a number of advantages. The main ones are the conceptual simplicity of these pricing schemes, their fairness, and that they provide the seller with a mechanism to induce buyers to smooth demand. The discussion in the second section claims that most of these advantages are related to the fact that the TB pricing schemes are based on admission controls. Therefore, the authors believe that there is a potential in using admission controls as pricing schemes and that further work on such usage would help improve the resource management of shared services.

NOTES

1. His theorem requires that the arrival of jobs be a Markovian process and that the arrival of tokens be an independent renewal process (both requirements are reasonable and are assumed throughout this paper).
2. A more sophisticated method of pricing token and depth (say, with quantity discounts) can be implemented as well, but for ease of exposition the focus is on a linear pricing mechanism.
3. In what follows, the subscript on a random variable is dropped when it will not lead to confusion.
4. A simulation was used with 500,000 periods, normal demand with mean 10 and standard deviation of 1, 2 or 3, D/R cost ratios of 0.9, 0.5, 0.2 or 0.1, and service-level requirements of 80 per cent, 90 per cent, 95 per cent and 99 per cent. The costs for both the TB and TBwRC approximations were typically within 1–2 per cent of optimal costs.

5. The practical differences between the solutions based on the tailored TB approximation and those obtained from the non-tailored one are that the former results in using the service level β in more cases.

References

Altmann, J. and Chu, K. (2001) 'A proposal for a flexible service plan that is attractive to users and Internet service providers', paper presented at the IEEE Infocom 2001, Conference on Computer Communications, Anchorage, Alaska, April.

Ball, M. O, Hoffman, R. and Chen, C. (2000) 'Collaborative decision making in air traffic management: current and future research directions', Technical Research Report, NEXTOR T.R. 2000–3; http://www.isr.umd.edu/NEXTOR/

Baron, O. (2003) 'Pricing and Admission-control for Shared Computer Services Using the Token Bucket Mechanism', PhD thesis, Sloan School of Management, MIT.

Baron, O. (2005) 'Bounds for regulated random walks and their application to inventory control', Working Paper, Rotman School of Management, the University of Toronto.

Berger, A. W. (1991) 'Performance analysis of a rate-control throttle when tokens and jobs queue', *IEEE Journal on Selected Areas in Communications,* **9**, 2.

Berger, A. W. and Whitt, W. (1994) 'The pros and cons of a job buffer in a token-bank rate-control throttle', *IEEE Transactions on Communications,* **42**, 2/3/4, 857–861.

Bitran, G. R. and Caldentey, R. A. (2003) 'An overview of pricing models for revenue management', *MSOM*, **5**, 203–229.

Gallager, R. G. (1996) *Discrete Stochastic Process,* Kluwer Academic, Boston.

Gibbons, R. (1992) *Game Theory for Applied Economists*, Princeton University Press, Princeton, NJ.

Glasserman, P. (1997) 'Bounds and asymptotics for planning critical safety stock', *Operations Research*, **45**, 2, 244–257.

Herve, M. (1995) *Cooperative Microeconomics: A Game-Theoretic Introduction*, Princeton University Press, Princeton, NJ.

Kelly, P. F. and Tan, D. (1998) *Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability*, University of Cambridge, Cambridge, UK.

MacKie-Mason, J. K. and Varian, H. R. (1994a) 'Some economics of the Internet', http://www.sims.berkeley.edu/hal/people/hal/papers.html.

MacKie-Mason, J. K. and Varian, H. R. (1994b) 'Some FAQs about usage-based pricing', http:/www.sims.berkeley.edu/hal/people/hal/papers.html.

Maglaras, C. and Zeevi, A. (2003) 'Pricing and capacity sizing for systems with shared resources: approximate solutions and scaling relations', *Management Science*, **49**, 8, 1018–1038.

McGill, J. I. and van Ryzin, G. J. (1999) 'Revenue management: research overview and prospects', *Transportation Science,* **33**, 2, 233–256.

Paschalidis, I. C. and Liu, Y. (2002) 'Large deviations-based asymptotics for inventory control in supply chains', *Operations Research*, **51**, 3, 437–460.

Ross, S. M. (1974) 'Bounds on the delay distribution in GI/G/1 queues', *Journal of Applied Probability,* **11**, 417–421.

Shenkar, S., Clark, D., Estrin, D. and Herzug, S. (1995) 'Pricing in computer networks: reshaping the research agenda', in *Proc. TPRC 1995*; http://citeseer.nj.nec.com/shenker95pricing.html.

Siegmund, D. (1985) *Sequential Analysis: Tests and Confidence Intervals*, Springer, New York.

Copyright of Journal of Revenue & Pricing Management is the property of Henry Stewart Publications and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.