Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

MANAGING ON-DEMAND COMPUTING SERVICES WITH HETEROGENEOUS CUSTOMERS

Inbal Yahav

Graduate School of Business
Bar Ilan University
ISRAEL 52900

Itir Karaesmen

Kogod School of Business
American University
Washington, DC 20016, USA

Louiqa Raschid

UMIACS and Robert H. Smith School of Business
University of Maryland
College Park, MD 20742, USA

ABSTRACT

Cloud computing allows the practice of allocating shared computational resources to a diverse set of customers in an on demand fashion. The heterogeneity of customers' requirements can be characterized by the frequency of demand, the value associated with the service and delay sensitivity. We consider the following two classes of customers: frequent users (FC) and urgent users (UR). The FC agreement is a proxy for the current business model used by typical providers; all customers are guaranteed the service for a flat, usage-based fee. Any delays are subject to penalties in FC. In the UR, service is guaranteed to be completed without any delay, but the service provider can accept/reject UR job requests with no penalties. We focus on the admissions control decision for the UR customers. Using computational experiments and simulation, we test the effectiveness of heuristic admissions control policies.

1 INTRODUCTION

Virtualized computing has revolutionized the way a company would scale to more powerful computational resources. As demand surges, instead of purchasing and maintaining additional resources which may often be under-utilized, virtual computing provides transparent and scalable access to remote and distributed resources on demand and, typically, at a reduced cost. Virtualized computing covers both the first wave of grid computing and the later wave of cloud computing. Both provide on-demand and scalable access to virtual resources. Today, virtualization is pervasive, and the typical organization that makes use of virtualized computing is no longer the early adopter of cutting edge software or services. Users vary from financial institutions to earthquake and climate forecasting simulations, to government and international organizations that deploy virtualized computing to solve problems like disaster response, to online gaming sites. While grid/cloud resources are typically external to an organization, there are many organizations that deploy an internal (private) grid/cloud or both. For instance, GM relies on both the private cloud and the public cloud (Bongard 2011). It is estimated that the global market for cloud computing will reach $241 billion by 2020 (Valention-Devries 2011).

The current success of both grid- and cloud-based computing reflects the maturity of the underlying enterprise solutions and services. In contrast, business models for virtualized computing are at a nascent stage. For resource providers, without adequate knowledge of consumer needs, the appropriate solutions

appear to lie in fairly simple business models and pricing structures. Our research is inspired by this. We consider a service provider that offers two different service agreements (contracts) to its customers. The first type of contract is intended for customers that have repeated and regular needs of resources and that can tolerate delay to some degree. The second service contract is intended for customers that do not necessarily use the provider's services regularly but who require it on an urgent basis when they have time-sensitive jobs.

The main research question we answer is related to the market-mix for the provider: When does the provider (if at all) benefit from offering different service agreements? To answer this question, we first investigate allocation of system resources and analyze the admissions control problem of the UR customers. We show the properties of the optimal admission control policy and introduce a near-optimal, myopic policy for which a closed form solution exists. We show that, in the presence of the UR segment, the service provider can benefit from limiting/reducing the demand of FC customers, hence reducing the workload of FC. This will allow accommodating a larger UR market share. These findings show that service providers can benefit from offering different agreements, but they have to carefully balance the demand from customers with heterogeneous needs, as demand for virtual computing increases.

The paper is organized as follows. In Section 2 we define the service provider's problem. In Section 3, we review the relevant literature. Next, in Section 4, we solve the admissions control problem to optimality and introduce the myopic policy. We present computational results in Section 5. We summarize our findings and discuss future research directions in Section 6.

2 PROBLEM DEFINITION

Consider a system composed of a single, preemptive single-processor server. Given the type of examples provided in Section 1, the service provider is aware of customers with distinct needs in terms of urgency of the on-demand service. On the other hand, the server is also interested in long-run profitability. Thus, managing a steady set of customers, with regular, frequent needs is important. Given these perspectives of the business, the service providers offers two types of service 'contracts' to *Frequent Customers* (FC) and to *Customers with Urgent Needs* (UR). For FC, the service provider guarantees to provide service upon request, or else it is subject to delay penalties. UR is not guaranteed to receive service and the provider can admit or reject their job requests. However, once an UR job is accepted, it is processed immediately, by preempting service (if needed) to FC jobs.

Currently, many of the cloud service providers are experimenting with different economic models. Still, the fixed fee is the most common form of pricing structures. In this paper, we analyze a business situation where the contract terms and prices for FC are fixed while customers can name their own fee for each job they submit under UR. This business model leads to a natural price segmentation among the customers who have urgent needs. We derive the service providers admissions control decisions for UR service. As a special case, we first study a business model where the service provider has a fixed fee and fixed expected service time for both FC and UR customers. This representative of the current practice but also constitutes a stepping stone for our analysis of the model where UR customers name their own fees.

Here is the notation and assumptions: FC service jobs arrive according to a Poisson process with rate λ^f. FC jobs have identically, independent distributed service times, exponentially distributed with mean μ^f. The service fee is r^f per each FC service, independent of the length of service. FC jobs do not have priorities and are served in a first-in-first-out (FIFO) order. If a job from an FC customer is not processed immediately upon arrivals, linear delay penalties are incurred; $\Pi^f d$ is the total delay penalty when a job from a FC customer is delayed by d time units, Π^f being the delay penalty per time unit. Similarly, UR job arrival is a Poisson process with rate λ^u. Service times of UR jobs are exponential distributed. Service times of incoming jobs in each segment are independent and identically distributed (iid). There are no delay penalties for UR; by definition these jobs <u>must</u> be processed immediately <u>if</u> accepted. Any FC job that is being served at the time an UR job is accepted, is preempted. In the name-your-own-fee (NYOF) model, UR customers differ by their service needs, i.e. their service time distribution, and their bid, i.e.

the fee they are willing to pay for the service. We define K to be the number of classes of UR customers in the NYOF model. A class k is characterized by the bid r_k^u and mean service rate μ_k^u, $k = 1,..,K$. Each incoming UR job has a probability of $p(k)$ ($\sum_{k \in K} p(k) = 1$) of belonging to class k. In the special case where all UR jobs pay the same fee and have the same expected processing time (i.e. $K = 1$), $\bar{r}^u$ denotes the bid of each UR customer and $\overline{\mu}^u$ is the mean service rate for UR customers.

The exact service times of the jobs are not known prior to the start of the service. The provider only knows the distribution of service times and has to make his/her admissions control decision on UR based on the fee proposed by the incoming customer and the distribution of service times. The service provider obtains revenues upon arrival of each FC job and whenever an UR job is accepted. The delay penalties to FC customers are paid as delay occurs.

3 LITERATURE REVIEW

The question of online resource allocation is well studied in the literature with several integrated questions: admissions control, due date quotation, and scheduling (also referred as job sequencing). Keskinocak and Tayur (2004) state that these questions are ideally to be considered simultaneously rather than sequentially. There is a growing body of literature that study pricing and resource allocation decisions using queueing models. This research dates back to Naor's (1969) work that studied static pricing decisions to control the arrival rate of a finite-buffer queueing system. More recently, Maglaras (2006) considers a single-server make-to-order production firm that offers multiple products. Assuming a general demand distribution, the authors study the problem of finding the optimal state-dependent pricing and sequencing strategy. Celik and Maglaras (2008) study a make-to-order manufacturer that offers multiple products to price- and delay-sensitive customers. They focus on dynamic pricing and lead time quotation as well as sourcing (expediting orders from a secondary source) decisions. Focusing on a make-to-order firm again, Webster (2002) examines policies for adjusting price and capacity in response to periodic and unpredictable shifts in how the market values price and lead-time. Pricing decisions are also the focus of Gilland and Warsing (2009) who model a single server that processes jobs from customers who have heterogeneous waiting costs. Gans and Savin (2007) model a car rental problem that provides a fixed-price service to contract customers and shop-for-price service for walk-in customers. The rental firm has to decide when to accept/reject contract customers (rejection is subject to penalty), and what fee to charge to walk-in customers. Walk-in customers are accepted as long as there is available capacity. We refer the reader to Cil et al. (2011) and the references therein for more information on dynamic pricing and scheduling problems in multi-class single-server queueing systems. Among the research papers that focus on virtual computing resources, Das et al. (2011) study the problem of pricing and risk management for an online storage grid.

We briefly review the current best practices for defining business models and service contracts for virtualized computing and services in the cloud. While typical aspects of a service contract include contract type, pricing, admission control, reliability, security, etc., the cloud business model is much simplified. The majority of providers offer one or both of the following service contracts that only differentiate on price: **Pay-as-you-go services** for which consumers are charged a at, usage-based fee; **Long-term services** offered to frequent customers where the providers commonly charge a flat fixed fee and a discounted or scalable usage fee. Until recently, goGrid and Amazon offered both pay-as-you-go and long-term pricing. One of the more sophisticated business models is offered by Amazon EC2. Amazon offers multiple instance types to match varying processor and memory needs. In addition, Amazon provides the following three pricing models, extending on the two mentioned previously: **On-demand instances** which resemble the pay-as-you-go service; **Reserved instances** offered to long-term customers; **Spot instances** to sell excess capacity at a discount. The reserved instance offers reduced costs to long-term customers. The spot instances allow on demand customers to bid on residual capacity. Their jobs will be accepted if their bid exceeds the Spot Price threshold. This threshold fluctuates based on capacity. We note that the three pricing models differentiate consumers based on price sensitivity alone; all the models assume that the customers similar delay sensitivity.

Our work differs from the models studied in the literature and current business models as follows: We model heterogeneity of customers with respect to delay sensitivity. Specifically, we have repeat customers who can be delayed, where the provider incurs a penalty. We also have urgent customers who cannot be delayed but they can be rejected entirely. We study a particular pricing scheme (the name your own fee model) to make admissions control decisions. Current business models focus on price sensitivity and do not differentiate customers based on delay sensitivity.

4 ANALYSIS OF ADMISSIONS CONTROL DECISIONS

We characterize the system based on the number of FC and UR jobs pending. Notice that there cannot be more then one UR job in the system at any point in time while there is no limit to the number of pending FC jobs as all FC jobs are admitted. Therefore, the system states are defined as (n^f, n^u), where n^f is the number of FC jobs, $n^f = 0,1,2,...$, and n^u is the number of UR jobs $n^0 = 0,1$ in the system. We track the evolution of the system at distinct time points, defined by an arrival or a service completion. Clearly, the system state affected by admissions decisions. Admission (accept/reject) decisions are done immediately upon arrival of an UR job. Then the following state transitions take place:

$$(i,0) \rightarrow \begin{cases} (i+1,0) & \text{arrival of FC job} \\ ((i-1)^+,0) & \text{service completion of FC job} \\ (i,1) & \text{arrival of UR job, job is accepted} \\ (i,0) & \text{arrival of UR job, job is rejected} \end{cases} \tag{1}$$

$$(i,1) \rightarrow \begin{cases} (i+1,1) & \text{arrival of FC job} \\ (i,0) & \text{service completion UR job} \end{cases}$$

We use Φ to denote the set of all feasible, non-anticipative admission control policies. We use ϕ for a generic admissions control policy. Let $TP^{\phi}(t)$ be the expected profit rate at time t when policy ϕ is used. The provider's problem is to find the optimal admissions control policy for UR jobs in order to maximize the long run average profit:

$$TP^* = \max_{\phi \in \Phi} \left(\lim_{t \to \infty} \frac{1}{t} \int_0^t TP^{\phi}(s) ds \right) \tag{2}$$

Notice that when $\lambda^u = 0$ the seller has no decisions to make (making any admissions control policy moot) and the system reduces to an M/M/1/∞ queue. In that case, the seller's maximum profit is

$$TP^* = \lambda^f (r^f - \Pi^f Wq), \tag{3}$$

Wq being the expected waiting time for FC jobs, defined as $Wq = \frac{\rho^f/\mu^f}{1-\rho^f}$ for $\rho^f = \lambda^f/\mu^f$. For the general case, the problem of finding the admissions control policy that yields the optimal profit can be further simplified using the following property (all technical proofs are available in Yahav et al. (2013) unless noted otherwise):

Theorem 1 The optimal admissions control induces a unique steady-state probability p_{ij} of visiting the system state set (i,j), for all $i = 0,1,2,...,\ j = 0,1$.

Based on Theorem 1, we can rewrite the optimal expected profit as the following optimization problem:

$$TP^* = \max_{\phi \in \Phi} \quad TP^{\phi} = \sum_{i=0}^{\infty} \sum_{j=0}^{1} TP_{ij}^{\phi} p_{ij}^{\phi} \tag{4}$$

where TP_{ij}^{ϕ} is the expected profit obtained while in state (i,j) when policy ϕ is used, and p_{ij}^{ϕ} is the steady state probability of being in state (i,j) at any point in time when policy ϕ is used.

4.1 Analysis of the Model with Fixed Fees

We focus on the special case with one class of UR customers. This special case corresponds to the situation where UR customers name their own fees, but happen to have the same willingness-to-pay (WTP) for the service which is exactly $\bar{r}^u$, and the same service needs (service time being exponential distributed with mean $\overline{\mu}^u$). This model is analogous to a practical situation where the service provider sets the non-negotiable fee $\bar{r}^u$ for mean service rate μ^u, and knows the resulting arrival rate λ^u.

We write the provider's expected profit function recursively for this model:

$$TP_{i0} = \frac{\mu^f}{\mu^f+\lambda^f+\lambda^u}(TP_{i-1,0}) + \tag{5}$$

$$\frac{\lambda^f}{\mu^f+\lambda^f+\lambda^u}(r^f - \frac{i\Pi^f}{\mu^f} + TP_{i+1,0}) + \tag{6}$$

$$\frac{\lambda^u}{\mu^f+\lambda^f+\lambda^u}\max_{u\in 0,1}\left\{1_{\{u=0\}}(TP_{i0}), 1_{\{u=1\}}\left(\bar{r}^u - \frac{i\Pi^f}{\overline{\mu}^u} + TP_{i1}\right)\right\} \tag{7}$$

if there are no UR customers in the queue; else it is as follows:

$$TP_{i1} = \frac{\lambda^f}{\lambda^f+\overline{\mu}^u}(r^f - \frac{i\Pi^f}{\mu^f} - \frac{\Pi^f}{\overline{\mu}^u} + TP_{i+1,1}) + \tag{8}$$

$$\frac{\overline{\mu}^u}{\lambda^f+\overline{\mu}^u}(TP_{i0}) \tag{9}$$

. These expressions are interpreted as follows: First of all, the term in Equation (5) corresponds to a service completion of FC job. The probability of this event is $\frac{\mu^f}{\mu^f+\lambda^f+\lambda^u}$ (derived from exponential racing). There is no immediate profit associated with this event. Next, the term in Equation (6) corresponds to an arrival of FC job, that finds the system in state (i, 0). The probability of this event is $\frac{\lambda^f}{\mu^f+\lambda^f+\lambda^u}$. The service provider collects the fee r^f and pays the immediate penalty from placing the FC job in the ith position. Next, the term in Equation (7) corresponds to an UR arrival, that finds the system in state (i, 0) (recall that if an UR job finds the system in state (i, 1) it is immediately rejected). The probability associated with this event is $\frac{\lambda^u}{\mu^f+\lambda^f+\lambda^u}$. In this case the service provider rejects ($1_{\{u=0\}}$) or accepts ($1_{\{u=1\}}$) the job to maximize his profit (where $1_{\{.\}}$ is the indicator function). If the service provider accepts the job, he collects the fixed fee $\bar{r}^u$ and pays the penalty caused by delaying current pending FC jobs. Next, the term in Equation (8) corresponds to an arrival of FC job, that finds the system in state (i, 1). The probability of this event is $\frac{\lambda^f}{\lambda^f+\overline{\mu}^u}$. The service provider collects the fee r^f and pays the immediate penalty from placing the FC job in the ith position and additional delay due to UR processing. Finally, the term in (9) corresponds to service completion of UR job. The probability of this event is $\frac{\overline{\mu}^u}{\lambda^f+\overline{\mu}^u}$. There is no immediate profit associated with this event.

Notice that the admissions control decision is actually embedded within Equation (7), and the optimal control policy is determined by solving

$$\max_{u\in 0,1}\left\{1_{\{u=0\}}(TP_{i0}), 1_{\{u=1\}}\left(\bar{r}^u - \frac{i\Pi^f}{\overline{\mu}^u} + TP_{i1}\right)\right\}. \tag{10}$$

This optimization problem has nice structural properties.

Theorem 2 $TP_{i1} + \bar{r}^u - \frac{i\Pi^f}{\mu^f} - TP_{i0}$ is non increasing in $i, i = 0,1,2,\ldots$.

Theorem 3 The optimal admissions control policy for UR jobs is threshold based, i.e., an incoming UR job is accepted if no other UR job is being processed and the number of FC jobs in the system is less then a critical number.

We define $i^*(\bar{r}^u)$ to be the optimal threshold, i.e., an incoming job is always rejected when the number of FC jobs in the system is greater than $i^*(\bar{r}^u)$. Consequently the system threshold state is $(i^*(\bar{r}^u), 0)$. While we can analytically show that $i^*(\bar{r}^u)$ is monotonic in $\bar{r}^u$, numerical experiments indicate that $i^*(\bar{r}^u)$ could be convex or concave in $\bar{r}^u$; see Yahav et al. (2013).

4.2 Analysis of the Name-your-own-fee (NYOF) model

In this section we examine the general case in which UR customers differ by processing time and willingness to pay. We consider K classes of UR customers where class k is characterized by the parameter pair (r_k^u, μ_k^u) for $k = 1,..,K$. The aggregated UR arrival rate remains λ^u, but each incoming customer has a probability $p(k)$ of belonging to class k ($\sum_{k\in K} p(k) = 1$). Suppose a class k customer arrives at time t. Then the service provider can accept or reject this customer after observing the customers' bid (r_k^u) and identifying his service needs (which is characterized by the mean service time μ_k^u).

The provider's expected profit in this NYOF model is given by the following recursive function:

$$TP_{i0} = \frac{\mu^f}{\mu^f+\lambda^f+\lambda^u}(TP_{i-1,0}) + \tag{11}$$

$$\frac{\lambda^f}{\mu^f+\lambda^f+\lambda^u}(r^f - \frac{i\Pi^f}{\mu^f} + TP_{i+1,0}) + \tag{12}$$

$$\frac{\lambda^u}{\mu^f+\lambda^f+\lambda^u}\sum_{k\in K}\left(p(k)\max_{u\in 0,1}\left\{1_{\{u=0\}}(TP_{i0}), 1_{\{u=1\}}\left(r_k^u - \frac{i\Pi^f}{\mu_k^u} + TP_{i1_k}\right)\right\}\right) \tag{13}$$

for $i = 0,1,2,..$ if there are no UR customer in the queue; else it is as follows:

$$TP_{i1_k} = \frac{\lambda^f}{\lambda^f+\mu_k^u}(r^f - \frac{i\Pi^f}{\mu^f} - \frac{\Pi^f}{\mu_k^u} + TP_{i+1,1_k}) + \tag{14}$$

$$\frac{\mu_k^u}{\lambda^f+\mu_k^u}(TP_{i0}) \tag{15}$$

In comparison to the provider's profit in the fixed price case, we have: First, the terms in Equations 11-12 remain unchanged (see Equations 5-6). Next, the term in Equation (13) corresponds to an UR arrival, that finds the system in state (i, 0) (recall that if an UR job finds the system in state (i, 1) it is immediately rejected). The probability associated with this event is $\frac{\lambda^u}{\mu^f+\lambda^f+\lambda^u}$. Given that an incoming UR request is of class k with probability $p(k)$, the service provider makes admissions control decision to maximize his expected profit. Finally, the terms in Equations 14-15 correspond to the provider's profit at state (i, 1) when the currently processed UR job is of class k. These equations are generalization of Equations 8-9 that account for different job classes.

We next explore the structure of the optimal admission control in the NYOF model.

Theorem 4 For each UR job in class k with bid and mean service time (r_k^u, μ_k^u), $TP_{i1_k} + r_k^u - \frac{i\Pi^f}{\mu^f} - TP_{i0}$ is nonincreasing in $i, i = 0,1,2,....$

Theorem 4 leads us to the following result, which is expected from our analysis of the fixed-fee model.

Theorem 5 The optimal admissions control policy for for each UR job class is threshold based, i.e., an incoming UR job is accepted if no other UR job is being processed and the number of FC jobs in the system is less then a critical number $i^*(k) = i^*(r_k^u, \mu_k^u)$, which varies with fee and service need.

4.3 A Myopic Admissions Control Policy

The NYOF model allows the decision-maker to exploit the differences in willingness-to-pay and service needs of the UR customers, while capturing the trade-off between the UR and FC segments. While the service provider can have very good information on FC - these are the customers who sign long-term contracts, - eliciting information on UR segment can be difficult. In particular, knowing the arrival rate λ^u with certainty and identifying the probability terms $p(k)$, together with r_k^u and μ_k^u, is almost impossible if the provider has no prior experience with the UR segment. Without any historical data, the service provider may resort to a myopic admission control policy. We next analyze the myopic policy and show that the myopic threshold (represented as the minimum bid to be accepted) can be obtained in closed form.

The myopic policy works as follows: The service provider disregards all future UR arrivals. Essentially, the service provider treats the incoming UR customer as a one-time opportunity in making his admission control decision. In this case we analyze only the tradeoff between the fee r_k^u that the customer pays and the additional penalty costs induced by accepting this job, based on the service need μ_k^u of the customer and the current system state.

Theorem 6 When the system is in state $(i,0)$, an incoming UR customer with (r_k^u, μ_k^u) is accepted by the myopic optimal policy if the bid r_k^u exceeds a certain threshold, specifically, the job is accepted if

$$r_k^u \geq \Pi^f \sum_{N=0}^{\infty} \left\{ \frac{i+N}{\mu_k^u} + (\lambda^f + \mu^f) BP \sum_{k=1}^{N} (BP + i + k - Wq) \right\} \tag{16}$$

and rejected otherwise. In this expression, BP is the busy period in an $M/M/1/\infty$ queue, and is given by $BP = \frac{1/\mu^f}{1-\rho^f}$ (Adan and Resing, 2001).

5 COMPUTATIONAL RESULTS

We use computational experiments to investigate different business models for the service provider, and to test effectiveness of different admissions control policies. We also use numerical examples to illustrate the trade-off between UR and FC segments. A list of the parameters used in the examples throughout the paper is given in Table 1.

Parameter(s)	Example 1	Examples 2a,2b	Example 3
$\{\lambda^f, \mu^f, \rho^f\}$	$\{varies, 1, varies\}$	$\{1,3,1/3\}$	$\{1,3,1/3\}$
$\{\Pi^f, r^f\}$	$\{1,10\}$	$\{1,10\}$	$\{1,10\}$
K	1	{10, 20, 30, 40, 50}	3
p_k	-	1/K	1/K
λ^u	$\{0.2, 0.5, 1\}$	2	2
μ^u	1	1	$\mu_k^u \in \{1,2,3\}$
ρ^u	-	2	-
r^u	$\{5,10,20\}$	$r_k^u = k$	15

Table 1: Values of the parameters.

5.1 Market Analysis (Optimal rate of FC arrivals)

As we mentioned before, the FC segment represents the long-term potential for the service provider. However, if the UR segment is lucrative enough, the service provider can choose the optimal rate of FC customers by limiting the total number of contract signed in order to maximize the profits from both segments. Clearly, controlling the rate of FC would be beneficial even without the existence of UR customers. In the absence of UR, the optimal rate of FC, denoted $\{\lambda^{f*} | \lambda^u = 0\}$ can be determined by

solving:

$$\max_{\lambda^f} \left(\lambda^f \left[r^f - \Pi^f W q\right]\right) \tag{17}$$

and is equal to

$$\{\lambda^{f*}|\lambda^u = 0\} = \mu^f \left(1 - \mu^f \sqrt{\frac{\Pi^f}{r^f}}\right). \tag{18}$$

We denote the service provider's profit in this case by $\{TP^*|\lambda^u = 0\}$.

On the other extreme, the profit rate of providing service to UR customers alone when $\lambda^f = 0$, denoted by $\{TP^*|\lambda^f = 0\}$, is

$$TP^* = TP^{\phi^0} = \frac{\lambda^u}{1+\rho^u} E[r^u]. \tag{19}$$

and is non- decreasing with λ^u. Hence, if $\{TP^*|\lambda^f = 0\} > \{TP^*|\lambda^u = 0\}$ then the optimal rate of FC customers, denoted λ^{f*} satisfies $\lambda^{f*} < \{\lambda^{f*}|\lambda^u = 0\}$. We next use a numerical example to show the value of serving UR customers in addition to the FC segment, and what is the optimal market-mix for the service provider.

Example-1. In this experiment, we show how the optimal rate of FC arrivals decreases in the presence of UR for a specific example in Figure 1. We use $\mu^f = 1, \overline{\mu}^u = 1, r^f = 10, \Pi^f = 1$. We examine two cases of UR market price: (1) the price is higher than the price for FC customers ($r^u = 20$, left panel), and (2) the price is equal to the price for FC customers ($r^u = 10$, right panel). We vary the rate of UR requests λ^u in the range $\{0.2, 0.5, 1\}$. When $\lambda^u = 0.2$, the availability of UR requests is low, and the provider has to mainly rely on the FC segment to ensure high, steady profit. On the other extreme, when $\lambda^u = 1$, UR requests are constantly available, allowing the service provider to accept more UR customers, and thereby to decrease the market share of the FC segment and reduce delay costs.

The figure presents the optimal profit curve (on the y-axis) as a function of FC arrival rate (on the x-axis) under the optimal admission thresholds for different UR arrival rates. The solid black curve shows how the total profit varies with λ^f in the null policy. In this example, the optimal FC rate in the absence of UR is $\{\lambda^{f*}|\lambda^f = 0\} = 0.68$ and the optimal profit rate is $\{TP^*|\lambda^f = 0\} = 5.36$. In the presence of UR, the optimal arrival rate of FC customers decreases significantly both when the UR market price is greater than FC fixed price, <u>and</u> when the UR market price equals to the FC fixed price and their arrival rate is sufficient (grater than $\lambda^u = 0.2$). For example, when $r^u = 20$, and $\lambda^u = 1$, the optimal arrival rate of FC customers decreases to $\lambda^{f*} = 0.13$. The profit rate increases to approximately $TP^* = 12.3$, resulting in a profit difference of approximately 230%.

5.2 System Performance under Different Models and Policies

In this section we provide numerical comparison of the performance of the optimal admissions control under fixed-fee model compared and the optimal admissions control for NYOF model. In all the experiments (in this and the next section), we use $\lambda^f = 1$, $\mu^f = 3$, $r^f = 10$ and $\Pi^f = 1$.

Example-2a. In this experiment we consider K classes of UR customers differing in bids, but sharing a common processing time $\overline{\mu}^u$. The fee is $r_k^u = k$ for class k. The probability that an incoming UR customer belongs to class k is $\frac{1}{K}$ for any $k = 1, ..., K$. We set $\lambda^u = 2$ and $\overline{\mu}^u = 1$. We vary the number of UR job classes K; K takes values in the set $\{10, 20, 30, 40, 50\}$. For the fixed-fee model, we choose the unit revenue $\overline{r}^u$ in the range $[1, 60]$. The arrival rate of UR in the fixed-fee model is adjusted such that $\widetilde{\lambda}^u = \lambda^u P(r \geq \overline{r}^u)$ where r is the random variable denoting the (random) fee offered by an incoming UR customer in the NYOF model.

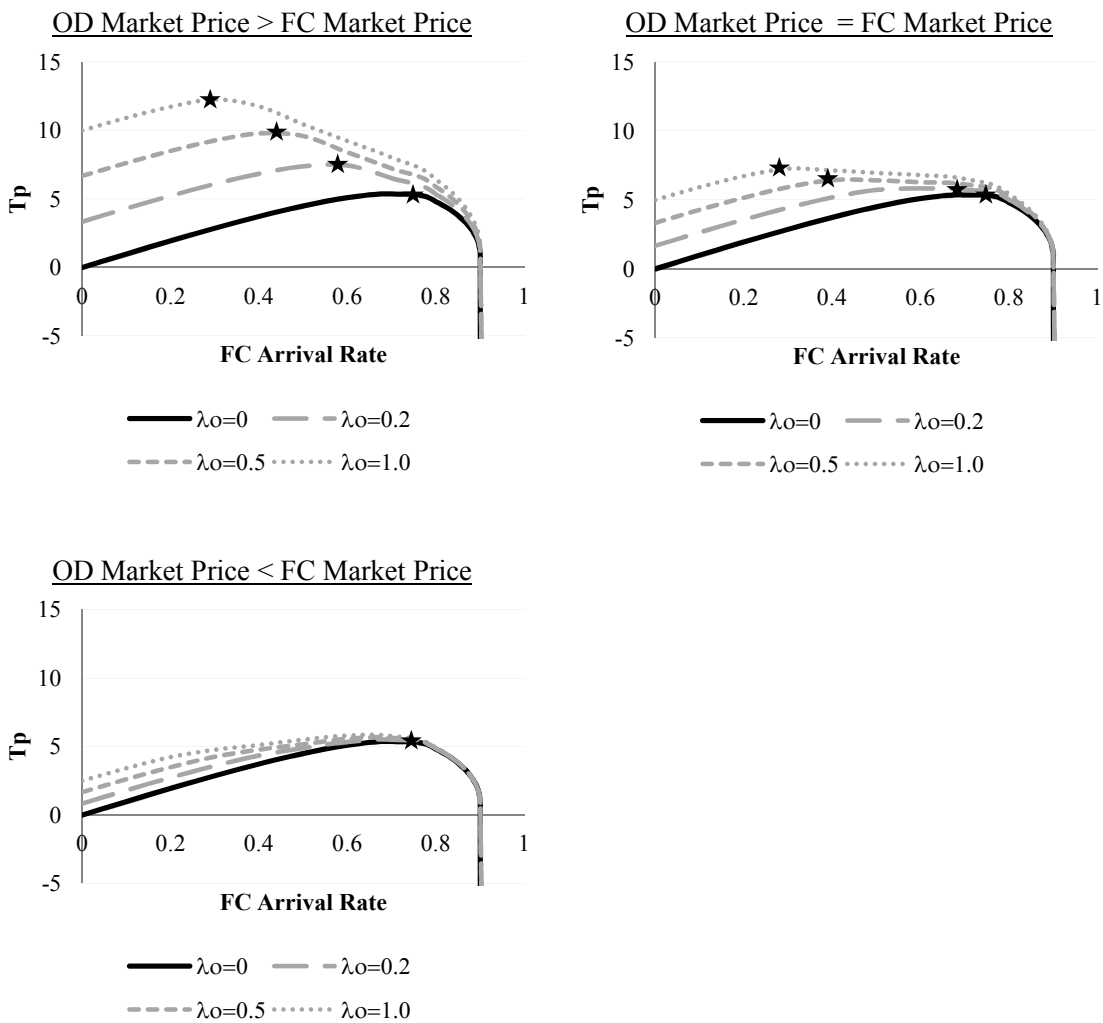


Figure 1: Optimal rate of FC jobs.

We depict the numerical results in Figure 2. In the left column we specify the number of different UR classes in each experiment. Then, in the second column we depict the revenue of the service provider under the fixed price model as a function of fixed price. The revenue of the provider increases up to an optimal price (under the limitation of same price for all jobs, regardless of the system state), then decreases when the rice becomes too high for the vast majority of UR customers. Eventually the price flattens, when it reaches the highest price UR customers are willing to pay. At this point, only the FC segment is being processed. In the last column in Figure 2, we compare the best attainable profit under the fixed price model to that of the optimal, dynamic policy. We show that NYOF induces a higher profit than the fixed-price model. This result is consistent across different number of classes of UR customers. The difference among the average profits of the three models becomes more significant as the number of job classes increases. This observation is consistent for different UR arrival rates.

Next, we evaluate the loss in profit when the myopic policy for NYOF is used rather than the optimal solution. In the next set of experiments, we use simulation. For each simulation run, we generate a sequence

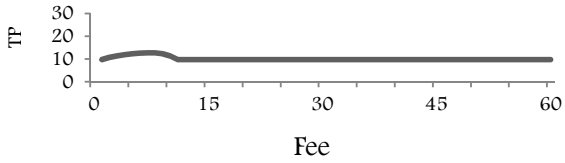
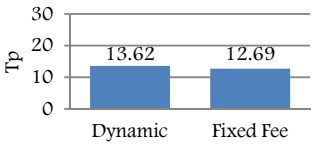
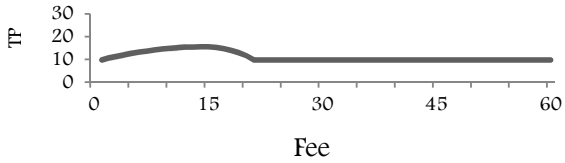
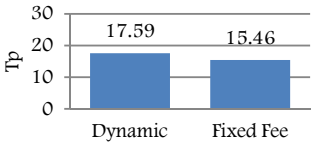
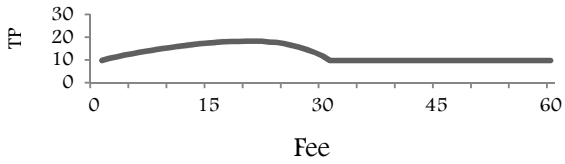
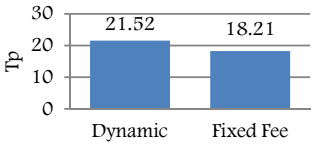
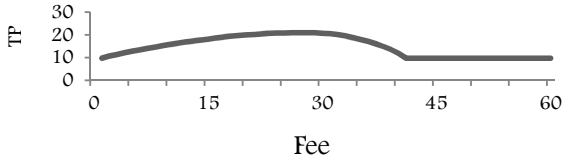
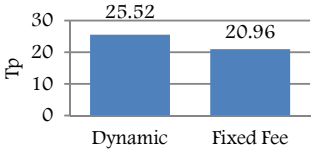
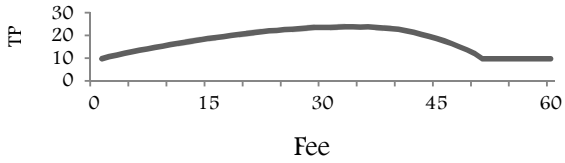
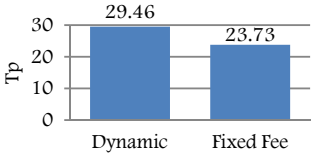
#OD classes (K)	Fixed price with varying fees	Method Comparison
10		
20		
30		
40		
50		

Figure 2: Service provider's expected profit (TP) under different policies as a function of UR job classes, for constant $\overline{\mu}^u$.

of random inter-arrival times and service times for each job of each type. The myopic and the optimal admissions control policies are then applied and the average profit is computed for each policy.

Example-2b. We reconsider Example-2a above. We compare the average profit of the myopic policy and that of the optimal admissions control policy. The results are given in Figure 3. This example suggests that for fewer classes of UR jobs, the myopic policy yields profits close to that of the optimal policy of NYOF. The performance deteriorates as the number of UR classes increases.

Example-3. We evaluate the policies when the fees of UR job classes are fixed but expected service times differ. We consider the $K = 3$ classes, with $\mu_k^u \in \{1,2,3\}$ and common $\bar{r}^u = r_k^u$. We compare the average profit of the myopic policy and that of the optimal admissions control policy for 500 runs. We compare the average profit of the myopic and optimal admissions control policies in Figure 4. Here again, the myopic policy performs nearly as good as the optimal admissions control, yielding a profit of at least 94% of what the optimal policy achieves.

We conclude that the optimal policy is significantly better than the myopic policy only if the arrival rate of UR is very high (or) the bids variance is high and predictable. In practical settings in cloud computing,

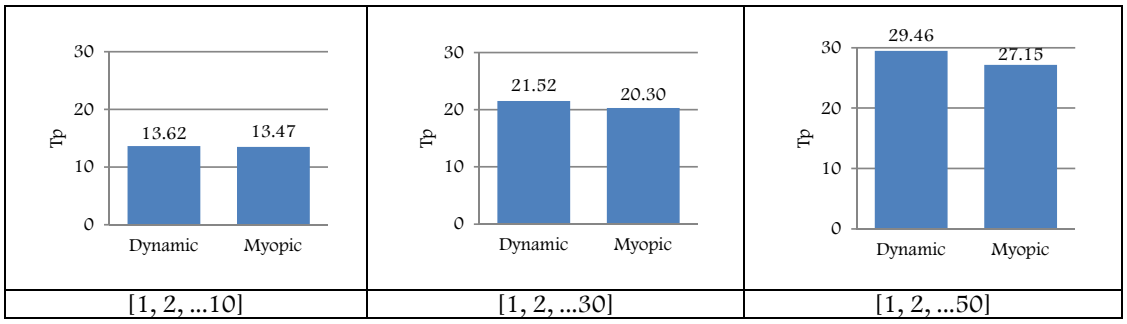


Figure 3: Service provider's optimal profit (TP) under different policies as a function of UR job classes, for constant $\overline{\mu}^u$.

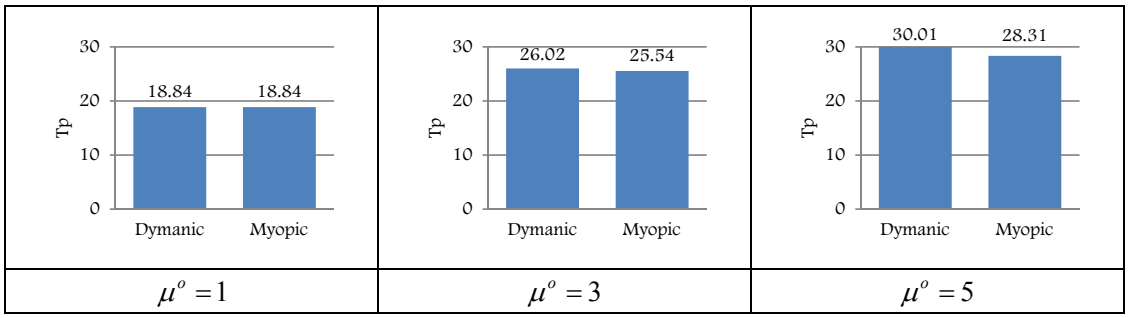


Figure 4: Service provider's under different policies as a function of UR job classes, for constant WTP r^u and three classes of expected completion time.

one can say that the UR arrival rate is low, making the myopic policy very effective: The profit under the myopic policy is less than 1% lower than that obtained by the optimal policy when UR arrival rate is low.

6 CONCLUSIONS

We presented a business model for virtual computing. The service provider has two types of customers, frequent use users (FC), with guaranteed service, possibly with a delay, and on demand users (UR). UR customers do not have a contract guaranteeing service, and therefore can be either accepted or rejected by the provider. We focused on the admissions control for the UR customers. We analyzed a model where there are $K \geq 1$ classes of UR customers, differing in price sensitivity and expected service completion time. We find that for each job class k the optimal admissions control is threshold based in the number of pending FC jobs in the system. Using computational experiments, we show that offering services to both FC and UR segments induces higher profits compared to other commonly used business models that either do not use admissions control or charge each job a fixed fee.

This model can be extended to multiple classes of FC customers, differing in their arrival rate, service time and unit revenue; this would expand the state space and pose computational challenges. In our model, we assume that UR customers do not wait in queue. Consequently, we do not have any scheduling decisions. However, by introducing a slack time for UR jobs, the system has to keep a queue for these customers and the provider has to make scheduling decisions for FC and UR customers in addition to admission control for UR. Another extension of our model would allow customers to choose their service type (FC vs. UR), based on their urgency and price sensitivities. In practice, an FC customer is not limited to submitting jobs through the contract channel, and can in fact choose the UR channel, if the latter is more beneficial for him. In this case one can study the game between the provider and the customers. This problem is richer

as customers utility functions have to be modeled, and the seller's pricing scheme and admission control policies have to be studied in a game-theoretic setting.

REFERENCES

Adan, I., J., Resing. 2001. *Queuing Theory*. Department of Mathematics and Computing Science, Eindhoven University of Technology.

Bongard, A. 2011. "GM CIO in Interview: Protecting GM Electronically is my Top Priority". *AutomotiveIT.com*, published online on September 5, 2011; available at http://www.automotiveit.com/gm-cio-in-interview-protecting-gm-electronically-is-top-priority/news/id-003692

Celik, S., and C. Maglaras. 2008. "Dynamic Pricing and Lead-Time Quotation for a Multiclass Make-to-Order Queue". *Management Science*, 54(6), pp. 1132-1146.

Cil, E.B., F. Karaesmen, and L. Ormeci. 2011. "Dynamic Pricing and Scheduling in a Multi-class Single-server Queueing System". *Queueing Systems,* 61(4), pp. 305-331.

Das, S., A.Y. Du, R. Gopal, and R. Ramesh. 2011. "Risk Management and Optimal Pricing in Online Storage Grids". *Information Systems Research*, 22 (4), p. 756-773.

Gans, N., and S. Savin. 2007. "Pricing and Capacity Rationing for Rentals with Uncertain Durations". *Management Science*, 53(3), pp. 390-407.

Gilland, W.G., and D.P. Warsing. 2009. "The Impact of Revenue-Maximizing Priority Pricing on Customer Delay Costs". *Decision Sciences*, 40(1), pp. 89-120.

Keskinocak, P., and S. Tayur. 2004. "Due Date Management Policies". *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*, pp. 485-553.

Maglaras, C. 2006. "Revenue Management for a Multi-class Single-server Queue via a Fluid Model Analysis". *Operational Research*, 54(5), pp. 914-932.

Naor, P. 1969. "The Regulation of Queue Size by Levying Tolls". *Econometrica,* pp. 37, 15-24.

Valentino-DeVries, J. 2011. "More Predictions on the Huge Growth of Cloud Computing". *The Wall Street Journal,* Digits Blog, published online on April 11, 2011; accessible at http://blogs.wsj.com/digits/2011/04/21/more-predictions-on-the-huge-growth-of-cloud-computing.

Webster, S. 2002. "Dynamic Pricing and Lead-Time Policies for Make-to-Order Systems". *Decision Sciences*, 33(4), pp. 579-599.

Yahav, I., I. Karaesmen, and L. Raschid. 2013. "Managing On-Demand Computing Services with Heterogeneous Customers - Unabridged Version". Working Paper, Bar Ilan University, ISRAEL.

AUTHOR BIOGRAPHIES

INBAL YAHAV is an Assistant Professor at Bar Ilan University. She holds a Ph.D. in Operations Research and Data Mining from University of Maryland. Her research lies at the interface between operations research and statistical data modeling, in the context of health care applications and online auctions. Her email address is inbal.yahav@biu.ac.il and her web page is http://faculty.biu.ac.il/~yahavi1/.

ITIR KARAESMEN is an Associate Professor at the Kogod School of Business at American University. She received her Ph.D. in Management Science from Columbia University. Her research includes pricing and revenue management, stochastic modeling, and supply chain optimization of perishable products. She is a member of INFORMS, and is Associate Editor for *Manufacturing and Service Operations Management*. Her e-mail is karaesme@american.edu and web page is http://auapps.american.edu/~karaesme/index.html.

LOUIQA RASCHID is a Professor at the Smith School of Business and UMIACS at the University of Maryland. She holds a Ph.D. in Electrical Engineering from University of Florida. Her research includes the challenges of information management and data science, with applications in the biomedical sciences, social media, and finance. Her e-mail is louiqa@umiacs.umd.edu and web page is http://www.umiacs.umd.edu/~louiqa/.