

Maximizing Profit of Cloud Brokers under Quantized Billing Cycles: a Dynamic Pricing Strategy based on Ski-Rental Problem

Gourav Saha and Ramkrishna Pasumarthy

Abstract—In cloud computing, users scale their resources (computational) based on their need. There is massive literature dealing with such resource scaling algorithms. These works ignore a fundamental constrain imposed by all Cloud Service Providers (CSP), i.e. one has to pay for a fixed minimum duration irrespective of their usage. Such quantization in billing cycles poses problem for users with sporadic workload. In recent literature, Cloud Broker (CB) has been introduced for the benefit of such users. A CB rents resources from CSP and in turn provides service to users to generate profit. Contract between CB and user is that of pay-what-you-use/pay-per-use. However CB faces the challenge of Quantized Billing Cycles as it negotiates with CSP. We design two algorithms, one fully online and the other partially online, which maximizes the profit of the CB. The key idea is to regulate users demand using dynamic pricing. Our algorithm is inspired by the Ski-Rental problem. We derive competitive ratio of these algorithms and also conduct simulations using real world traces to prove the efficiency of our algorithm.

I. INTRODUCTION

A. Overview

There is no universal definition of cloud computing. However as far as our research is concerned, the most apt definition of cloud computing is found in [1] which can be quoted as: "computing as a *utility*". In our day to day life the most common utilities are electricity, water, gas, heat, postpaid mobile services etc. Similarly in cloud computing, computing resources (like CPU, memory, storage, network domains, virtual desktop) are *rented* to users based on their demand. From user's viewpoint, it eliminates the need of an upfront investment as an user can pay based on the amount of resources it has used. This is termed as "pay-per-use" or "pay-as-you-go" model. Therefore, *resource scaling* is the most fundamental aspect of cloud computing and hence extensive amount of effort has been channeled to explore this area. Cloud Service Providers (or CSP's), like Amazon, ElasticHost etc, rent computing resource to the users in form of *Virtual Machines* (also called *instances*) or VMs. Scaling of VMs revolves around two fundamental questions:

1. *From Users Perspective*: How to scale VMs to optimize a certain objective?
2. *From CSP's Perspective*: How to support the active VMs using minimum number of physical servers?

In cloud computing literature the former is often called *auto scaling* while the latter is called *dynamic provisioning*. These two questions are indeed similar and rely on a common line of research. Various researchers approached these two problems with different objectives and by using different mathematical tools. In the following we will give a brief account of these approaches.

Concerning auto scaling, the most simplest methods are the Static Threshold based policies [2] where the scaling of VMs is triggered when certain CPU parameters (CPU load, response time) crosses a pre-defined threshold. Such methods are too simple and if not designed properly face the problem of limit-cycles. Queuing theory based techniques are also very well developed. Some well cited papers in this regard can be [3], [4]. Theoretic results available in queuing theory uses simplified models and hence cannot be directly applied to complex systems like those encountered in cloud computing framework. Also the results are statistical implying that the analysis is true only over a long run after the system reaches steady-state. Control Theory based approaches aim to solve the latter problem by trying to control the transients [5], [6]. They are generally reactive methods, and hence are not good for workloads which have a lot of sudden spikes. There is a lot of active research concerning proactive methods. They depend on predictive abilities and use tools like Model Predictive Controller [7], Time Series Analysis [8], [9], Machine Learning [10]. Work done in [9] is impressive in the sense that the prediction algorithm has low overhead and also has the ability to distinguish fine workload patterns. There are several tasks which are deadline sensitive. There is a separate line of research based on heuristic algorithms [11], [12] to handle these tasks.

Similar tools have been used to optimize performance of data centers. Queuing theory has been widely used. [13] is a good reference in this regard as it uses much less approximation to model a cloud data center. Control Theoretic approaches are very famous. One popular book dealing with this subject is [14]. [15] is as well cited paper which presents practically viable method to regulate a performance metric of data center using feedback control. A very interesting method to save energy of a cloud data center has been presented in [16] where they use feedback control to regulate CPU clock frequency based on incoming request rate. Optimization based techniques have been directly used to design resource allocation strategies to maximize profit of cloud data centers [17], [18]. Another line of research considers optimizing the electricity and the cooling cost required to run a data center. Such problems are not trivial when electricity [19] and the water cost [20] are time varying.

B. Quantized Billing Cycles and the Cloud Broker

There is a fundamental misconception regarding resource scaling in cloud computing literature. Definitely, the main objective of cloud computing is elasticity, i.e. the user can scale-up or scale-down VMs based their short term needs.

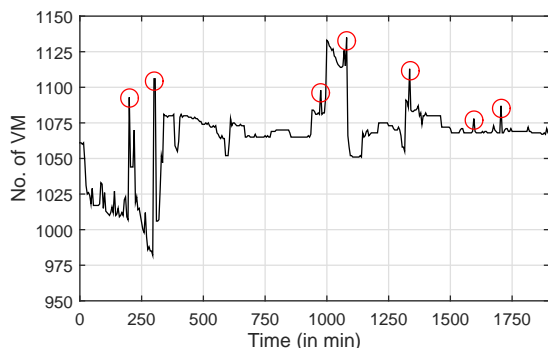


Fig. 1. A part of the google cluster usage traces showing spikes in user demand. Red circles indicates the spikes.

However this “short term” is not infinitely small in the sense that it may not be possible to return the VMs in the very next instant after it was rented.

We will explain this situation using an example. Consider that we are using on-demand instances of Amazon EC2 to satisfy these demands. Billing Cycle of AmazonEC2 on demand instance is 1 hour, i.e. you have to pay the same price if you use the VM for 1 min or 1 hour. We call this phenomenon as *Quantized Billing Cycles* (abbreviated as QBC) in the rest of the paper. This leads to serious problem especially for those users whose demand pattern is sporadic in nature. For e.g. Consider the workload trace shown in Figure 1 which shows the number of VMs required to satisfy user demand. There are few spikes in the demand curve to satisfy which one needs to buy extra VMs. These VMs may be in use for a very small fraction of 1 hour after which it may be idle. However we cannot return these VMs before 1 hour¹. Hence there will be wastage of VMs.

To mitigate this problem to some extent the concept of cloud broker has been proposed in recent literature [21], [22], [23]. A broker forms a middle man between the CSP and the general users as shown in Figure 2. The users send their job requests to the cloud broker. The cloud broker rents VMs from the Cloud Service Provider to service these demands. The cloud broker charges the user based on the fraction of VMs resources used to service the job request. This is called *pay-what-you-use*. It can also be based on *per-request-basis*. This transfers the challenge of QBC from the users to the cloud broker. However this issue is not as critical for the cloud broker as it is for the users. This can be understood as follows:

1. QBC poses problems for those users whose demand is sporadic. Higher the sporadic nature, greater the loss.
2. Cloud broker serves the aggregate demand of many users. In statistical sense, the sporadic/spiky nature in aggregated demand should be lesser compared to individual user’s demand. This is because a summer is a discrete integrator, a low pass filter which removes the noises.

In the rest of the paper we will concentrate on how a cloud

¹The user may choose to return the VM but it won’t get any financial benefit. Therefore it is better to keep the VM for the entire billing cycle even though it remains idle. This is called *smart-kill*.

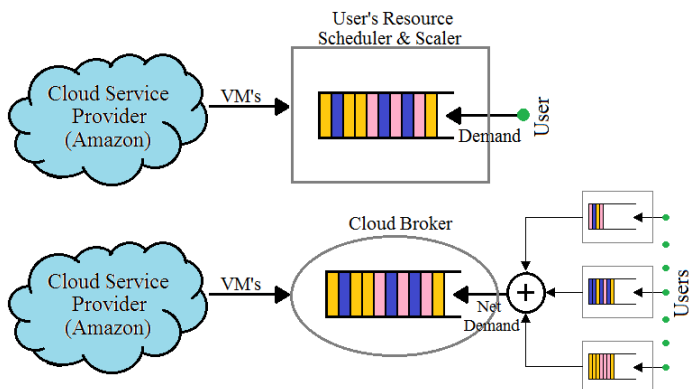


Fig. 2. Schematic showing cloud broker as a middleman between Cloud Service Provider (CSP) and the users. Adapted from [23].

broker can maximize its profit under QBC.

Remark: We would like to stress that the mathematical formulation, algorithms and analysis presented here can be applied to a single user without any change. However given that a cloud broker has a smoother workload pattern compared to a single user, a broker will benefit more compared to a user.

C. Existing literature and our contribution

In our problem, every time a cloud broker has to buy a VM it faces the risk of under utilization of the VM in the subsequent time slots. The broker has to make a decision without knowledge of future demand. Study of such problems comes under the category of Online Algorithms, more specifically the Ski-Rental problems. There have been a few applications of Ski-Rental literature to solve real world problems. In this section we will discuss the ones pertaining to cloud computing. These works has close resemblance with our problem.

We derive our motivation from the work done in [24], [23]. In these two works the cloud broker has to decide in each time slot whether to reserve instances or to serve the demand using on-demand instances. If the demand persists for a long time then reserving is a better option. However if the demand falls down quickly it is better to use on-demand instances. Cloud broker has to make this decision without any knowledge of future demand or with partial knowledge of future demand. The authors designed both deterministic and randomized on-line algorithms and also derived their competitive ratio. [25] and [26] considers the same problem however the model used in [25] is more general. It tries to optimize the switching cost and electricity cost of a data center. If there is a decrease in demand then the data center has to decide whether to shut down few physical servers or let them run. To run a physical server one needs to bear the electricity bill while switching down the server incurs wear-and-tear cost. Wear-and-tear cost is higher than electricity bill in short run but smaller in the long run. If one decides to switch off the server and immediately later it is needed, then one saves a small electricity bill in the expense of suffering a large switching cost. Therefore taking such decisions without knowledge of future demand is difficult. In [25], author designed only deterministic algorithms

to tackle this problem while [26] considers both deterministic and randomized algorithms. A recent work presented in [27] shows that the competitive ratio of the randomized algorithms can be improved if statistical information (like first and second order moment) of the demand process is available.

To the best of our knowledge the references given above constitutes almost all the major work dealing with the use of ski-rental framework for cloud computing applications. We make two main contributions to this literature:

1. We suggest how *dynamic pricing* can be used to maximize the profit of the cloud broker under QBC. This can be used as an alternative to the work done in [24], [23] specially in the case where the workload is deferrable. One can also use this approach alongside [24], [23]. These two points will be discussed further in Section V.
2. We show that the knowledge of future demand leads to better competitive ratio for the Partial Online algorithm. In [26] such attempts have been made but with a major assumption on the demand graph, i.e. the demand increases or decreases no more than one step in every time slot.

II. PROBLEM FORMULATION

A. Motivating Example

We aim to use dynamic pricing as a control signal to regulate user demand. The motivation to use dynamic pricing in cloud computing setup is derived from the paper [28]. Before proceeding forward with the quantitative formulation of the problem, we will first consider an example to illustrate our idea. Consider the following scenario:

1. Cloud broker buys VMs from Amazon at \$0.132 per VM. This is the cost price of the cloud broker. The billing cycle of each VM is 1 hour = 60 min.
2. Duration of a time slot = 10 min. Hence a VM is active for 6 time slots.
3. Nominal selling rate = \$0.03 per VM per time slot. Nominal selling rate should be such that if the VM is in use for all the time it is active (6 time slots here) then the cloud broker should make a profit. In our case: $0.03 \times 6 = \$0.18 > \0.132 . Hence the condition holds.
4. The cloud broker has the freedom of changing the selling rate² at the beginning of each 10 min interval. The users remains totally aware of the current selling rate.

We will investigate two cases, one with static pricing and the other with dynamic pricing. Relevant graphs are shown in Figure 3. In both the cases we start with the assumption that at time $t = 0$ the cloud broker has no VMs.

Case 1 (Static Pricing):

In static pricing, the selling price remains constant at nominal rate therefore, Demand = Actual Demand. In the 1st interval the demand is of 2 VMs while we have 0 VMs. So we buy 2 VMs incurring a cost of $2 \times \$0.132 = \0.264 . In the 2nd interval, demand is of 10 VMs while we have 2 VMs (bought in the 1st interval). So we buy 8 VM

²In the work the term “VM price” inherently means the selling price of the VM not the cost price.

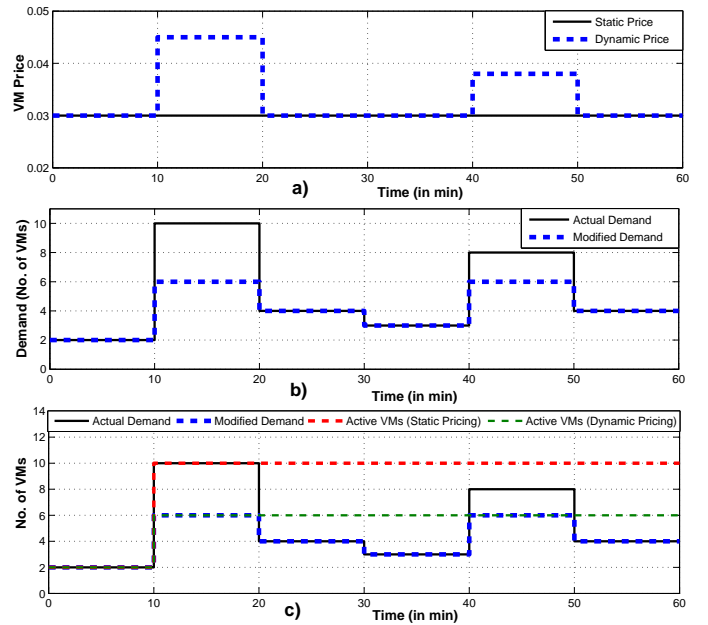


Fig. 3. An example of static and dynamic pricing: a) Graph showing Price vs Time for static and dynamic pricing. b) The graph corresponding to actual demand (for static pricing) and modified demand (for dynamic pricing). c) Graph showing actual/modified demand along with the active number of VMs for both static and dynamic pricing case.

incurring a cost of $8 \times \$0.132 = \1.056 . In the next 4 intervals the demand is less than 10 VM and hence we don't have to buy any more VMs. Therefore the net cost price of the cloud broker is \$1.32. The net selling price of the cloud broker is $(2 + 10 + 4 + 3 + 8 + 4) \times \$0.03 = \$0.93$. Profit = Selling Price – Cost Price = $\$0.93 - \$1.32 = -\$0.39$, i.e. the cloud broker suffered a loss.

Case 2 (Dynamic Pricing):

In this case the selling price varies in response to which the user's demand gets modified. This case is similar to Case 1 except that in the 2nd and the 5th interval the selling price goes up to \$0.045 and \$0.038 respectively. The cloud broker has to buy 2 VMs and 4 VMs in the 1st interval and 2nd interval respectively incurring a cost price of $6 \times \$0.132 = \0.792 . The net selling price of the cloud broker is $(2 + 4 + 3 + 4) \times \$0.03 + 6 \times \$0.045 + 6 \times \$0.038 = \0.888 . Profit = Selling Price – Cost Price = $\$0.888 - \$0.792 = \$0.096$, i.e. the cloud broker makes a profit.

We will encapsulate the idea behind dynamic pricing by making the following key observations:

1. The idea of increasing the selling price is to decrease the demand and *not* to increase the revenue. To understand this point consider the 2nd interval. The actual demand was 10 VMs which could have lead to a revenue of \$0.3 if the rate was nominal. As the selling price increased to \$0.045 the demand reduced to 6 VMs leading to an revenue of \$0.27. Therefore the net revenue in the 2nd interval decreased even though the selling price increased. Similar argument is true for the 5th interval. Therefore whenever there is an increase in price there is a decrease

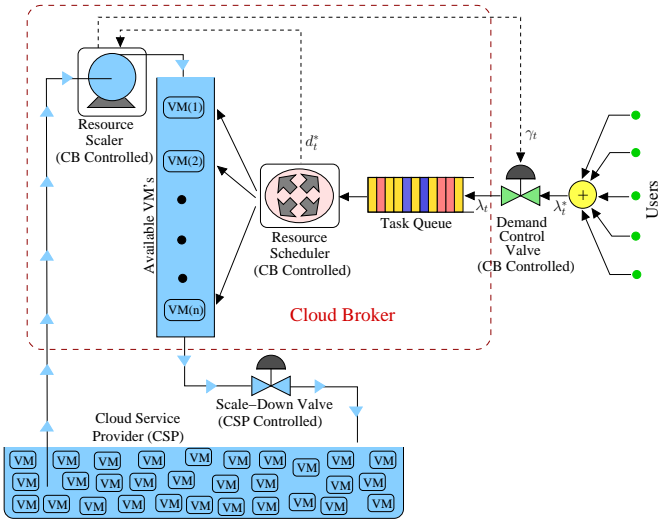


Fig. 4. Cartoonistic representation of Cloud Brokerage mechanism.

in revenue (and hence in profit) in that interval. Pictorially speaking, the revenue loss suffered in dynamic pricing can be captured by the area between the *solid black* curve and the *dashed blue* curve in Figure 3b.

2. Yet dynamic pricing makes more profit than static pricing. This is because in static pricing, many VMs are underutilized and hence does not contribute to the revenue. This is clearly shown in Figure 3c. Underutilized VMs for static case corresponds to the area between the *solid black* curve and the *dashed red* curve while that in dynamic case corresponds to the area between the *dashed green* curve and the *dashed blue* curve. Definitely the area corresponding to static case is more.
3. The idea behind dynamic pricing is: “Suffer a small loss in one interval by decreasing the demand (Referred as *Demand Loss* later) rather than buying a VM and then suffering a major loss in the subsequent intervals due to low demand (Referred as *VM Loss* later)”.

B. Quantitative Modeling

Cloud brokerage mechanism basically consist of two blocks: 1) Resource Scheduler 2) Resource Scaler. This is shown in Figure 4. Both resource scheduler and resource scaler are controlled by the cloud broker. The job of the Resource Scheduler is to schedule the incoming tasks onto the available VMs. While doing so it has to consider service level agreement (or SLAs). Resource Scheduler can be designed using well established theoretic tools as discussed in Section I-A. The role of the resource scaler is to rent (scale-up) VMs from CSP and also to perform dynamic pricing in order to maximize the profit of the cloud broker. In this paper, we assume that resource scheduler is already designed and concentrate on designing algorithms for Resource Scaler. Note that the design of Resource Scheduler and Resource Scaler can be decoupled. The Resource Scheduler should just update the Resource Scaler regarding the number of VMs required to complete the tasks in a given time slot. Before moving forward, please note

the scale-down valve which is controlled by CSP. As shown in Figure 4, this is not in control of the cloud broker. This aptly captures the notion of QBC and the process of *smart-kill*. As mentioned in Section I-B, due to the presence of QBC there is no point of giving back a VM before its billing cycle ends. This will be equivalent to the CSP automatically scaling down the VMs after its billing cycle. Therefore from cloud broker’s perspective: “Only scaling-up of VMs is controllable but scaling-down is not”.

We will now pose our problem mathematically. We consider that the user pays the cloud broker based on per-request/pay-what-you-use basis. In such a scenario the resource scaler has to solve the following profit optimization problem:

$$\begin{aligned}
 \text{OP1} & : \max_{\{\gamma_t, v_t\}} P = \sum_{t=1}^T (\gamma_t d_t - v_t) \\
 \text{subject to} & : \sum_{i=t-\tau+1}^t v_i \geq d_t ; \forall t = 1, 2, \dots, T \\
 & d_t = f(d_t^*, \gamma_t) ; \forall t = 1, 2, \dots, T
 \end{aligned}$$

In optimization problem **OP1**, P is the profit to be maximized. The term $(\gamma_t d_t - v_t)$ is the profit at t^{th} interval where, γ_t is the selling price per VM per time slot, d_t is the number of VMs required to service the incoming job request and v_t is the number of VMs bought at t^{th} interval. Without any loss of generality we normalize the cost price of a VM to 1 unit. τ is the period of the billing cycle and hence $\sum_{i=t-\tau+1}^t v_i$ is the number of active VMs in the t^{th} interval. d_t is the modified VM demand when the selling price is γ_t . The relation between the actual demand d_t^* and the modified demand d_t is captured by the *price-demand function* $f(\cdot)$. If $\gamma_t = \gamma^*$, the nominal price, then $d_t = d_t^*$. The revenue earned by selling a VM at the nominal price of γ^* for one complete billing cycle is $\gamma^* \tau$. For the cloud broker to make profit, $\gamma^* \tau > 1$, 1 being the cost price of a VM. It should be noted that all the variables associated with **OP1** lies in the set \mathbb{R}_+ .

In conventional sense, the optimization problems dealt in ski-rental framework are minimization problems. Therefore we are interested in formulating **OP1** as an equivalent minimization problem. In this regard note that

$$\begin{aligned}
 P & = \sum_{t=1}^T (\gamma_t d_t - v_t) \\
 & = \sum_{t=1}^T [\gamma^* d_t^* - \{(\gamma^* d_t^* - \gamma_t d_t) + v_t\}] \\
 & = \sum_{t=1}^T [\gamma^* d_t^* - \{(\gamma^* d_t^* - \gamma_t f(d_t^*, \gamma_t)) + v_t\}] \quad (1)
 \end{aligned}$$

In equation (1) the first term $\gamma^* d_t^*$ is not controllable and hence the maximization of P becomes equivalent to minimization of $\sum_{t=1}^T [(\gamma^* d_t^* - \gamma_t f(d_t^*, \gamma_t)) + v_t]$. We thus pose the following optimization problem which is equivalent to **OP1**:

$$\begin{aligned}
\text{OP2} : \quad & \min_{\{\gamma_t, v_t\}} L = \sum_{t=1}^T \left[\underbrace{(\gamma^* d_t^* - \gamma_t d_t)}_{\text{Demand Loss}} + \underbrace{v_t}_{\text{VM Loss}} \right] \\
\text{subject to} : \quad & \sum_{i=t-\tau+1}^t v_i \geq d_t; \forall t = 1, 2, \dots, T \\
& d_t = f(d_t^*, \gamma_t); \forall t = 1, 2, \dots, T
\end{aligned}$$

Intuitively speaking, **OP2** does the following: Consider that there is a hike in demand d_t^* which decays soon. In such a case **OP2** will increase the selling price γ_t to reduce the demand. In this way the cloud broker will suffer a small ‘‘Demand Loss’’. Buying enough VMs to support the demand hike is not a good option in such scenario as the cloud broker may suffer a huge ‘‘VM Loss’’ in subsequent intervals due to underutilized VMs. But if the hike in demand persists for a long time it is better to buy VMs to support this hike. However **OP2** is an offline optimization problem, i.e. to solve **OP2** we need d_t^* for all $t = 1, 2, \dots, T$. It is not possible to know in advance if an increase in demand is going to persist or will decay soon. The challenge is to design algorithms which can make such decisions online based on present and past data. Such algorithms are called online algorithms.

We now define the concept of Competitive Ratio which will be used later in Section III to compare the performance of the online algorithm with its optimal offline counterpart. Say that an online algorithm A and the optimal algorithm OPT (**OP2** here) suffers a loss $L_A(d^*)$ and $L_{OPT}(d^*)$ respectively for a given demand sequence $d^* = [d_1^* \ d_2^* \ \dots \ d_T^*]^T$. Then algorithm A is called c -competitive if:

$$L_A(d^*) \leq c \cdot L_{OPT}(d^*) \quad \forall d^* \in \mathbb{R}_+^T \quad (2)$$

Indeed $c \geq 1$. In inequality (2) we have slightly misused the notation. \mathbb{R}_+^T is a T dimensional vector of non negative real numbers *not* transpose of a non negative real number.

C. Properties and Assumptions

Properties of Demand and Revenue Function

Most of the price demand function $f(d^*, \gamma)$ (also called demand function) found in real world must satisfy the following conditions:

1. $f(d^*, \gamma) \geq 0$.
2. $f(d^*, \gamma) \geq 0$ is monotonically increasing in d^* .
3. $f(d^*, \gamma)$ is monotonically decreasing in γ in the range $[\gamma^*, \infty)$ while it is constant³ at d^* in the range $[0, \gamma^*]$.
4. The revenue function $\gamma f(d^*, \gamma)$, is monotonically decreasing in γ in the range $[\gamma^*, \infty)$. This captures the idea that increasing the selling price is to decrease the demand and *not* to increase the revenue. As $f(d^*, \gamma)$ is constant

³We assume that the demand cannot increase above the actual demand d^* even if the price decreases below γ^*

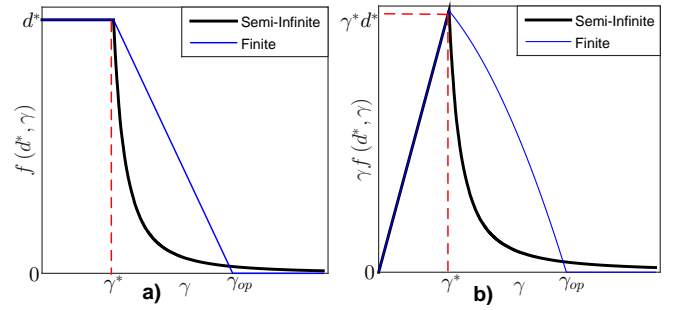


Fig. 5. Graphs showing typical: a) Demand Functions b) Revenue Functions. Two demand functions and the corresponding revenue functions are shown. For the Semi-Infinite case $[\gamma^*, \infty)$ is the operating zone while for Finite case $[\gamma^*, \gamma_{op})$ is the operating zone.

in $[0, \gamma^*]$, it is obvious that $\gamma f(d^*, \gamma)$ will linearly increase in this range. Revenue function is maximum at $\gamma = \gamma^*$ and hence this is set as the nominal price.

Let $d = f(d^*, \gamma)$. Then according to *property 4* we have

$$-\infty < \frac{\partial}{\partial \gamma} (\gamma d) \leq 0 \quad (3)$$

Differentiating inequality (3) using chain rule we get

$$-\infty < d + \gamma \frac{\partial d}{\partial \gamma} \leq 0 \Leftrightarrow -\infty < \frac{\partial d}{\partial \gamma} \leq -\frac{d}{\gamma} \quad (4)$$

A demand function satisfying inequality (4) will be *strictly* monotonically decreasing if $d > 0$. Two cases may arise:

Case 1 (Semi-Infinite Operating Zone): $d > 0$ for all $\gamma \in [\gamma^*, \infty)$ and hence the demand function is strictly monotonically decreasing in this range.

Case 2 (Finite Operating Zone): $d > 0$ for all $\gamma \in [\gamma^*, \gamma_{op})$. For $\gamma = \gamma_{op}$, $d = 0$. Due to *property 1*, $d = 0$ for all $\gamma \geq \gamma_{op}$. Therefore the demand function is strictly monotonically decreasing in the range $[\gamma^*, \gamma_{op})$.

The range of γ where $f(d^*, \gamma)$ is *strictly* monotonically decreasing is called the *operating zone*. Figure 5 illustrates the concept of operating zone for the above two cases. Strict monotonic nature of $f(d^*, \gamma)$ in the operating zone implies that the function is invertible in this range. Mathematically, the following function exist in the operating zone,

$$\gamma = g(d^*, d) \quad (5)$$

The function $g(\cdot)$ returns the imposed price γ_t given the actual demand d_t^* and the modified demand d_t .

Remark: The role of dynamic pricing is to regulate the demand. In the range $[0, \gamma^*]$, price has no effect on the demand and yet the cloud broker will incur a demand loss. Therefore to minimize **OP2** we work in the operating zone.

Assumptions in Problem Formulation

1. We ignore the effect of *reputation* while formulating our optimization problem. By increasing the price we force some tasks to exit the queue. By doing this we earn

negative reputation of the users. A *static* demand function of the form $d_t = f(d_t^*, \gamma_t)$ does not capture the effect of pricing history and hence the role of reputation.

2. The partial online algorithm, discussed later in Section III, relies on demand prediction for future window w . In reality, only an estimate of future demand is possible however we assume perfect knowledge of future demand.
3. In real-life scenario, the relation between the modified demand d_t and the actual demand d_t^* is not governed by a deterministic function $f(\cdot)$. Rather the demand d_t has a probability distribution⁴ in the range $[0, d_t^*]$ for a given γ_t . However many works dealing with social welfare maximization using dynamic pricing (like [29], [28]) consider such deterministic demand function. We also assume the knowledge of $f(\cdot)$.
4. Define the following function in the operating zone,

$$p(d^*, d) = \frac{\partial}{\partial d} [d \cdot g(d^*, d)] \quad (6)$$

Also define two more variables,

$$\begin{aligned} p_m &= \min_{d^* \in \mathbb{R}_+, 0 \leq d \leq d^*} p(d^*, d) \\ p_M &= \max_{d^* \in \mathbb{R}_+, 0 \leq d \leq d^*} p(d^*, d) \end{aligned} \quad (7)$$

We impose the following constraints on p_m and p_M ,

$$\frac{1}{\tau} < p_m \leq p_M < 1 \quad (8)$$

Inequality $p_M < 1$ implies that renting is cheaper than buying in short run while the inequality $\frac{1}{\tau} < p_m$ implies that buying is cheaper in the long run. We will further elaborate this in Section III-A.

Let $\gamma = g(d^*, d)$. Then according to inequality (8) we have

$$p_m \leq \frac{\partial}{\partial d} (\gamma d) \leq p_M \quad (9)$$

Differentiating inequality (9) using chain rule we get

$$p_m \leq \gamma + d \frac{\partial \gamma}{\partial d} \leq p_M \Leftrightarrow \frac{p_m - \gamma}{d} \leq \frac{\partial \gamma}{\partial d} \leq \frac{p_M - \gamma}{d} \quad (10)$$

According to inequality (4), the following inequality is true in the operating zone

$$-\frac{\gamma}{d} \leq \frac{\partial \gamma}{\partial d} < 0 \quad (11)$$

Inequality (10) and (11) is simultaneously satisfied if

$$\frac{p_m - \gamma}{d} \leq \frac{\partial \gamma}{\partial d} < \min \left(0, \frac{p_M - \gamma}{d} \right) \quad (12)$$

This is because $\frac{p_m - \gamma}{d} > -\frac{\gamma}{d}$. However $\frac{p_M - \gamma}{d} \geq 0$ for $\gamma \leq p_M$ while $\frac{p_M - \gamma}{d} < 0$ for $\gamma > p_M$. It is trivial to observe that there exist a $\frac{\partial \gamma}{\partial d}$ satisfying inequality 12 *if and only if* $\gamma > p_m$. Also $\frac{\partial d}{\partial \gamma}$ is the slope of the demand function $f(\cdot)$ with respect to γ . Hence inequality 12 upper and lower bounds the slope of the demand function for a given d and γ .

⁴One may consider that the function $f(\cdot)$ is the mean of this probability distribution.

A. Ski-Rental Problem

The ski-rental problem abstracts a class of problem in which a player has to decide whether to buy or rent a resource without a priori knowledge of the period of usage. Renting is cheaper if the period of usage is short while buying is cheaper in the long run. In the original problem a skier is faced with the option of either buying or renting a set of skis without knowing in advance the number of days she will be skiing.

Cost of buying skis is \$1 while renting cost $\$P$ per day where $P < 1$. If the skier knows in advance that she will be skiing for y days then the choice of buying or renting is simple. If $y \geq \frac{1}{P}$ then the skier will buy the skis in the very first day. Otherwise she will keep renting the skis for y days.

The online case is more challenging. In ski-rental literature, the concept of *breakeven point* is used to design online algorithms. Such algorithms suggest that the skier should keep on renting the skis till the n^{th} day when the cost of renting nP , is more than the cost of buying, i.e. $nP > 1$. On the n^{th} day she should buy the skis. These is shown to be the most optimal *deterministic* online algorithm and has a competitive ratio of 2.

Ski-Rental problem has been used to solve real life problems like TCP acknowledgement problem [30], Bahncard problem [31] etc. Its application in cloud computing has already been discussed in Section I-C. The key step towards using ski-rental literature for our problem would be to map the following four entities in our context: **1) Renting** **2) Buying** **3) Buying Cost** **4) Renting Cost**. In remaining of this section we will define these entities.

1. Renting: It is the process of decreasing the demand by increasing the selling price of the VMs. To decrease actual demand d_t^* to modified demand d_t we impose selling price of $\gamma_t = g(d_t^*, d_t)$.

2. Buying: It is the process of buying v_t VMs to support the modified demand d_t . If the number of active VMs in the beginning of time slot t is x_t then $v_t = \max(0, d_t - x_t)$.

Figure 6 illustrates the renting and buying process. In the 3^{rd} time slot renting is equivalent to reducing the demand from 12 to 9 while buying is the process of purchasing 4 VMs. Similarly in the 1^{st} , 6^{th} and the 7^{th} time slot there is both renting and buying. There is no renting in the 2^{nd} time slot, we only buy 2 VMs. On contrary there is no buying in the 8^{th} time slot, we only rent the demand from 9 to 7. In the 4^{th} and the 5^{th} time slot there is neither renting nor buying because the number of active VMs in the beginning of these time slots is more than the actual demand.

3. Buying Cost: It is the cost of buying n VMs. As the cost price of VM is assumed to be 1 unit, the cost of buying n VMs is equal to n units.

4. Renting Cost: It is the demand loss in a given time slot associated with reducing a demand from d to $d - n$, where $n \leq d$, when the actual demand is d^* . Mathematically

$$R(d^*, d, n) = g(d^*, d) \cdot d - g(d^*, d - n) \cdot (d - n) \quad (13)$$

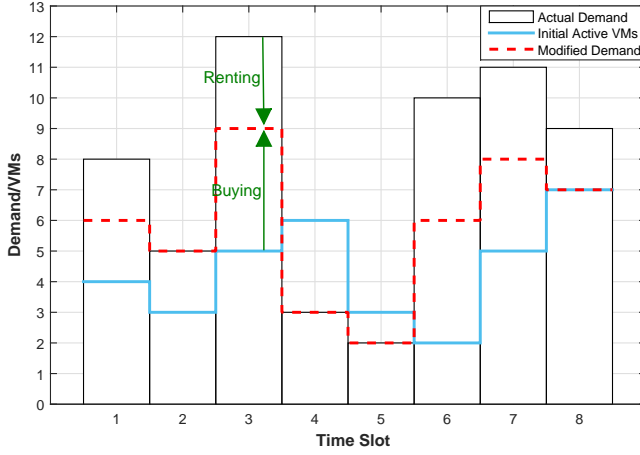


Fig. 6. Figure showing the actual demand d_t^* (solid black curve), modified demand d_t (dashed red curve) and the number of active VMs x_t in the beginning of every time slot (solid blue curve).

It should be noted that unlike the renting cost found in other ski-rental literature, for our case, R is not a constant. It is a function of d^* , d and n . However we will not mention these parameters explicitly for notational simplicity. *Please Note:* “Renting Cost” and “demand loss” means the same and will be used interchangeably.

Now we will explain the importance of inequality constrain (8). As mentioned before renting cost should be more than the buying cost in the long run. Given that the billing cycle is of τ period, the cost of renting for τ period should be greater than the cost of buying. Otherwise buying of VMs will never be required. The cost of buying n VMs is n while the cost of Renting of n demands for τ period is $R\tau$. Hence, $R\tau \geq n$. Similarly renting cost should be lesser than the buying cost in the short run. Therefore the cost of renting n demands for 1 period should be less than the cost of buying n VMs. Hence, $R \leq n$. Therefore to formulate our problem in Ski-Rental framework, it is necessary that the following inequality holds

$$\frac{n}{\tau} \leq R \leq n \quad (14)$$

Proposition 1: If inequality (8) is satisfied then inequality 14 will hold true.

Proof: Consider the following,

$$\begin{aligned} R &= g(d^*, d) \cdot d - g(d^*, d - n) \cdot (d - n) \\ &= \int_{d-n}^d \frac{\partial}{\partial \theta} [\theta \cdot g(d^*, \theta)] d\theta \\ &= \int_{d-n}^d p(d^*, \theta) d\theta \\ &\geq \int_{d-n}^d p_m d\theta = p_m n \end{aligned} \quad (15)$$

Equation (15) and inequality (16) comes from the definition of $p(d^*, x)$ and p_m (refer equation (6) and (7)). Similarly,

$$R \leq p_M n \quad (17)$$

The qualitative interpretation of inequality (16) and inequality (17) is that the minimum and maximum renting cost of 1 demand is p_m and p_M respectively. If inequality (8) holds, then $\frac{1}{\tau} \leq p_m$. Substituting this in inequality (16) we get

$$\frac{n}{\tau} \leq R \quad (18)$$

Similarly if inequality (8) holds, then $p_M \leq 1$. Substituting this in inequality (17) we get

$$R \leq n \quad (19)$$

Combining inequality (18) and (19) we get inequality (14). This completes the proof.

B. Online Algorithm

As discussed earlier in the previous subsection the concept of breakeven point has been widely used to design online algorithms in ski-rental literature. The key idea is to keep renting till the time renting cost equals buying cost. Here buying cost is the breakeven point. If renting cost exceeds the buying cost, we buy the resource. In this section we will apply these concepts to design two online algorithm: 1) Fully Online Algorithm which has no knowledge of future demand 2) Partial Online Algorithm, which assumes perfect future demand information for future window $w < \tau$. Intuitively speaking, both Fully Online Algorithm and Partial Online Algorithm works in *pessimistic sense*. It always assumes that an increase in demand is not going to persist and hence it reduces the demand by increasing the selling price of the VMs. This reduction in demand incurs a renting cost. However at every time interval it calculates the Net Renting Cost in the past and the future intervals. If the Net Renting Cost exceeds 1 unit (the buying cost of 1 VM), it buys a new VM.

The psuedocode of partial online algorithm with future window of w is given in Algorithm 1. The same psuedocode is applicable for the fully online algorithm if we substitute $w = 0$. Both fully/partial online algorithm consist of three basic steps. In the following we will explain these steps in relation to the fully online algorithm.

Step 1. Calculating Net Renting Cost

We calculate the net renting cost l which we could have been saved if we rented 1 less demand in each slot for the past τ period. Let x_i be the number of active VMs at time i . Then the net renting cost is

$$l = \sum_{i=t-\tau+1}^t R(d^*, x_i + 1, 1) \quad (20)$$

Equation (20) is valid only if $x_i + 1 \leq d_i^*$; $\forall i \in [t - \tau + 1, t]$. If for a given interval $x_i + 1 > d_i^*$, the renting cost for that interval is 0. This is done in Steps 4 to 10 of Algorithm 1.

Step 2. Buying new VM

This is done in Steps 11 to 16 of Algorithm 1. In our case the cost of a VM is 1. If $l \geq 1$ then the corresponding demands should not have been reduced. Rather we should have bought a VM to serve them. To compensate for this mistake we buy

Algorithm 1 Partial Online Algorithm with future prediction window w

Let x_i be the number of VM's at time i

1. Set $x_i = 0$; $i = 1, 2, \dots, T$
 2. Predict actual demand d_i^* for $i = t, t + 1, \dots, t + w$.
 3. *do*
 4. Set Net Renting Cost $l = 0$. Also set $i = t + w - \tau + 1$.
 5. *while* ($i \leq t + w$)
 6. *if* ($x_i + 1 \leq d_i^*$)
 7. $l = l + [g(d_i^*, x_i + 1) \cdot (x_i + 1) - g(d_i^*, x_i) \cdot x_i]$
 8. *end if*
 9. $i = i + 1$
 10. *end while*
 11. *if* ($l \geq 1$)
 12. Buy a new VM: $v_t = v_t + 1$.
 13. Update the number of VM's that can be used in future: $x_i = x_i + 1$; $i = t, t + 1, \dots, t + \tau - 1$
 14. Update the number of VM's in the history indicating that previous mistakes have been corrected: $x_i = x_i + 1$; $i = t + w - \tau + 1, \dots, t - 1$
 15. *end if*
 16. *while* ($l \geq 1$)
 17. *if* ($x_t \leq d_t^*$)
 18. $\gamma_t = g(d_t^*, x_t)$
 19. *else*
 20. $\gamma = \gamma^*$
 21. *end if*
 22. Jump to Step 2.
-

a VM in the current time slot. We increase the current and the future x_i by 1 to indicate that an extra VM is available. We also increase the past x_i by 1 to indicate that a corrective measure was taken. We then jump back to Step 1. However if $l < 1$, we jump to Step 3.

Step 3. Setting the Selling Price

Let the number of active VMs in the current time slot be x_t after performing Step 2 and Step 3. Then x_t is the modified demand. So we set our selling price as $\gamma_t = g(d_t^*, x_t)$.

The partial online algorithm is almost same as the fully online algorithm. The difference lies in the calculation of Net Renting Cost in **Step 1**. In case of fully online algorithm it is calculated for the period $t - \tau + 1$ to t while for partial online algorithm it is calculated for the period $t + w - \tau + 1$ to $t + w$.

Theorem 1: Competitive Ratio of partial online algorithm is

$$c(\alpha) = 1 + \min(1, p_M \tau (1 - \alpha)) \quad (21)$$

where $\alpha = \frac{w}{\tau}$ and $w < \tau$.

Proof: Please refer appendix for the proof.

Corollary 1: Fully online algorithm is 2-competitive.

Proof: For fully online algorithm $\alpha = 0$. Hence $c = 1 + \min(1, p_M \tau)$. According to inequality (8) we have

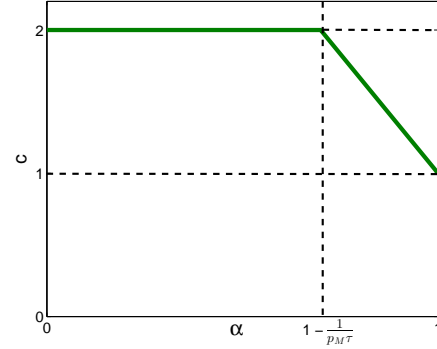


Fig. 7. Graph showing the competitive ratio of partial online algorithm.

$p_m \tau \geq 1$. Also $p_M \geq p_m$ implying $p_M \tau \geq 1$. Therefore $\min(1, p_M \tau) = 1$ and hence $c = 2$.

Note that the competitive ratio of the partial online algorithm can be more explicitly written as

$$c = \begin{cases} 2 & ; \alpha \leq 1 - \frac{1}{p_M \tau} \\ 1 + p_M \tau (1 - \alpha) & ; \alpha > 1 - \frac{1}{p_M \tau} \end{cases}$$

As $p_M \tau > 1$, there always exist an $\alpha_M \in (0, 1)$ such that $1 + p_M \tau (1 - \alpha) < 2$; $\forall \alpha > \alpha_M$. This gives the theoretical guarantee that future demand information indeed improves the performance of the online algorithm. Figure 7 shows a typical plot of equation (21).

IV. SIMULATION RESULTS

We performed simulations driven by real world traces to validate the online dynamic pricing algorithm proposed in the paper.

The first step is to generate the actual demand curve. To do this we have used google cluster usage traces available in [32]. The actual demand curve is shown in Figure 7b and 7d. The curve spans 1 day and is slotted in 5 min interval. Cost price of a VM is taken to be 1 unit. A VM has a billing cycle of 1 hour and hence $\tau = 12$. The next step is to generate the demand function. A real world demand function can only be inferred by doing a market survey. But for the sake of simulation, we have synthesized our own demand function. This can be explained in steps:

1. We consider a demand function of the form $\frac{d}{d^*} = f(\gamma)$.
2. A value of p_m and p_M satisfying inequality (8) is chosen.
3. A finite interval of price $[\gamma^*, \gamma_o]$ is uniformly divided into small parts.
4. For each part we substitute the corresponding value of γ and d in inequality (12). The value of $\frac{\partial \gamma}{\partial d}$ for this part is chosen in random such that it satisfies inequality (12). Using this value of $\frac{\partial \gamma}{\partial d}$, d for the next part is calculated.

As part of this simulation we will conduct two comparative studies, first to study the effect of demand prediction and secondly the effect of p_m .

To study the effect of demand prediction we first synthesized a demand function with $p_m = \frac{1}{12}$ and $p_M = 0.8$. We then

simulated *Algorithm 1* for $w = 0$ and $w = 4$. The results of the simulation is clearly shown in Figure 7b and 7c. Compared to $w = 4$, the reduction in demand is more in $w = 0$. This shows the pessimistic nature of our online algorithm, i.e. it prefers reducing the demand compared to buying new VMs. With increase in future window w the algorithm tends towards the optimal counterpart. The net profit $P = \sum_{t=1}^T (\gamma_t d_t - v_t)$ for $w = 0$ is 781 units while for $w = 4$ it is 937 units. The net profit P is more for $w = 4$ and hence the net loss L (refer **OP2**) is less. This is in consensus with *Theorem 1*.

We next studied the effect of p_m on our algorithm. To do this we constructed another demand function with $p_m = \frac{3}{12}$ and $p_M = 0.8$. From Figure 7d and 7e we can observe that if p_m is high the effect of pricing on demand is less. This is an obvious consequence of the definition of p_m .

V. DISCUSSION AND EXTENSIONS

We discussed the unique challenge posed by the presence of quantized billing cycles. The dynamic pricing strategy proposed in this paper can be considered as an alternative to the work done in [24] to maximize the profit of the cloud broker. Merging our algorithm with that of [24] should be very interesting. Such a merging will lead to a very interesting class of problems where the cloud broker has to decide whether to **a)** reduce the demand by increasing the VM price **b)** buy *on-demand* VMs to support the demand **c)** reserve VMs. This problem is similar to multislope ski-rental problem.

Two *deterministic* online algorithms were designed to increase the profit of cloud brokers in the presence of QBC. The competitive ratio of both the algorithms were derived. We showed the importance of demand prediction by deriving a better competitive ratio for the partial online algorithm than those found in ski-rental literature. In similar lines we would like to explicitly point out that the competitive ratio of *Algorithm 3* of [24], i.e. deterministic algorithm with demand prediction window w , has a competitive ratio of $1 + p\tau \min(1, \frac{w}{\tau})$ which is better than that reported in [24] (refer *Proposition 5*). This result is new in ski-rental literature.

It has been widely reported in ski-rental literature that *randomized* algorithms has better competitive ratio than its *deterministic* counterpart. Extending our algorithm to its randomized counterpart should be trivial. However deriving a better competitive ratio for the randomized algorithm with demand prediction may be challenging.

The key idea of our algorithms is to use pricing signal to regulate user demand. One may argue that such an algorithm gives poor service to the user as it pushes tasks out of the queue in order to maximize cloud broker's profit. We would like to make few comments in this regard:

1. Those tasks which gets pushed out of the queue can enter it again at a later instant. So we are not rejecting the tasks. Rather we are deferring it. Our algorithm is specifically good if p_M is low. A low p_M implies that even with a small increase in price there will be significant change in demand. Qualitatively, a low p_M correspond to tasks with low priority. Such tasks can be deferred.

2. Some literature considers penalizing the cloud provider when it fails to meet SLA's (refer [33]). Such a practice is not encouraged in cloud computing as it is a service oriented computing paradigm. If the cloud provider accepts the task, it must meet the SLAs. If it cannot it is better that it rejects the task. We are following similar practice.
3. We can penalize cloud broker to push tasks out of the queue by introducing *reputation factor*. As discussed in section II-C we need to consider dynamic price-demand function to include reputation factor. Such an extension of our work will be challenging but a very fruitful research.

As an immediate extension of our work we are interested in two areas. *Assumption 4* discussed in Section II-C may not be true for all real world demand functions. We are exploring the implication of relaxing it. Our intuition is that if we make a small modification in *Algorithm 1*, the competitive ratio will remain the same. We are also actively working on the randomized algorithm for this problem. More specifically we are exploring the possibility of improving the competitive ratio of the randomized algorithm in the presence of statistical information about the actual demand. This is motivated by the work done in [27].

APPENDIX

We will denote the partial online algorithm with future window of w by A_w and the offline optimal algorithm **OP2** by OPT . We will first consider two lemmas which are important for the proof.

Lemma 1: Let A_w buy N_w VMs while OPT buy N_{OPT} VMs throughout the time duration of $t = 1$ to $t = T$. Then for any demand sequence d_t^* , $N_{OPT} \geq N_w$.

Lemma 1 is an obvious consequence of the pessimistic nature of the fully/partial online algorithm as discussed in Section III-B. Both these algorithms assumes that a rise in demand is not going to persist and hence has the tendency to reduce the demand by increasing the selling price rather than buying VMs to support the rise in demand. Therefore $N_{OPT} \geq N_w$. Please refer Appendix A of [24] for the proof.

Lemma 2: The net renting cost of those demands that were served by the same VM in OPT should be greater than or equal to 1 (the breakeven point).

Proof: The proof directly follows from the definition of OPT . We prove this lemma by contradiction. Consider that in OPT a VM was bought to serve demands whose net renting cost is less than 1. The loss suffered to buy a VM to support these demands is more than the loss suffered to rent these demands. Therefore there can be a better algorithm to reduce the loss which contradicts the definition of OPT .

Let A_w and OPT rent n_t and N_t demands respectively at time t . Let $\mathcal{R}(A_w)$ and $\mathcal{R}(OPT)$ be the net renting cost of A_w and OPT respectively. We have,

$$\mathcal{R}(A_w) = \sum_{t=1}^T R(d_t^*, d_t^*, n_t)$$

$$\mathcal{R}(OPT) = \sum_{t=1}^T R(d_t^*, d_t^*, N_t)$$

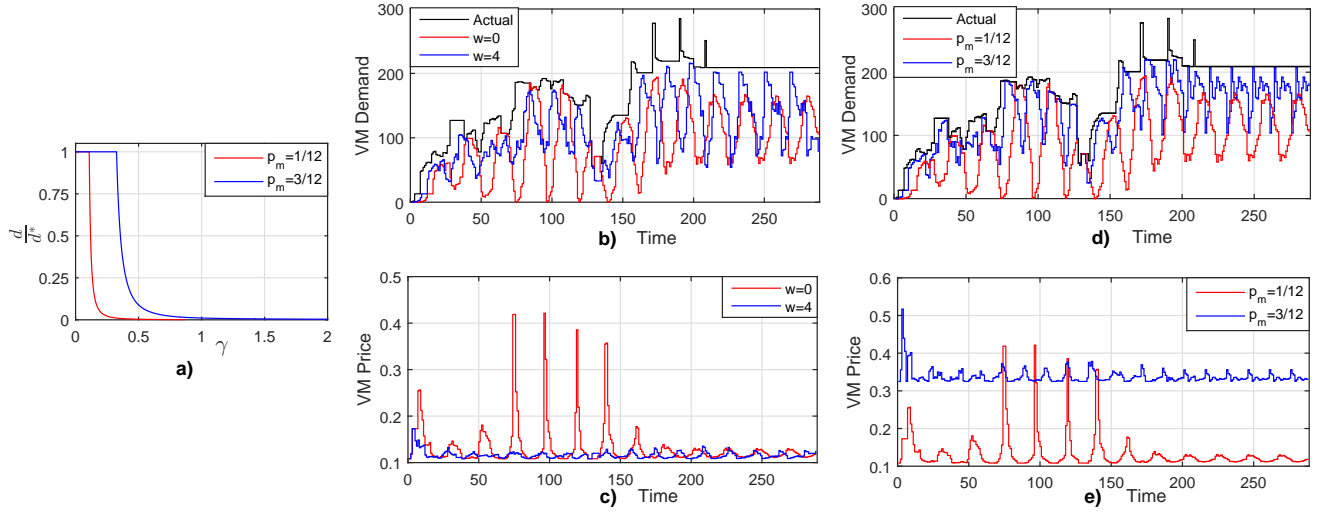


Fig. 8. a) Demand function for two different value of p_m . b), c) Plots of actual and modified demands and the corresponding pricing to compare the effect of prediction window w . d), e) Plots of actual and modified demands and the corresponding pricing to compare the effect of p_m .

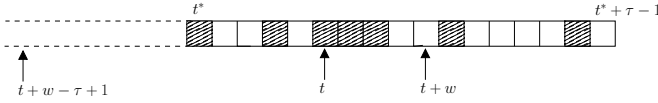


Fig. 9. Figure showing the demands served by a VM in OPT .

Let $\mathcal{R}(A_w \setminus OPT)$ denote the net renting cost incurred in A_w which is not incurred in OPT . Mathematically

$$\mathcal{R}(A_w \setminus OPT) = \sum_{t=1}^T [R(d_t^*, d_t^*, n_t) - R(d_t^*, d_t^*, N_t)]^+$$

where $(x)^+ = \max(0, x)$. We are interested in upper bounding $\mathcal{R}(A_w \setminus OPT)$. To do this, consider the demands that were rented in A_w but not in OPT . These demands were served by not more than N_{OPT} VMs in OPT . We focus on the demands served by one of these N_{OPT} VMs. This is shown in the following figure

Figure 9 shows the demands served by a VM in OPT which is bought at time t^* and last till $t^* + \tau - 1$. The shaded areas shows the demands which are served by this VM. If the area is not shaded then in that time slot the VM does not serve any demand. The shaded areas can be scattered anywhere in the time interval $[t^*, t^* + \tau - 1]$. According to *Lemma 1*, the net renting cost of the demands depicted by these shaded areas should be greater than or equal to 1.

Now we investigate how these demands will be treated by A_w . While calculating the net renting cost, A_w considers both future demand and past demand. This is shown in figure 9 by the three arrows. A_w calculates the net renting cost in the interval $[t+w-\tau+1, t+w]$ where t is the current time, the interval $[t+w-\tau+1, t]$ corresponds to the past window while $[t+1, t+w]$ is the future window. If the net renting cost in the interval $[t+w-\tau+1, t+w]$ becomes 1, then A_w will buy a VM. As mentioned before the net renting cost in the interval $[t^*, t^* + \tau - 1]$ is definitely greater than or

equal to 1 (due to *Lemma 1*). Therefore there is definitely a time t satisfying $t^* \leq t \leq t^* + \tau - w$ when A_w will buy a VM. The demands in the interval $[t^*, t]$ will be rented while that in the interval $[t+1, t^* + \tau - 1]$ will be served by the VM. The maximum renting cost of the demands in the interval $[t^*, t]$ is $\min(1, p_M(t - t^*))$. This can be understood as follows:

1. The maximum renting cost of one demand is p_M (refer inequality (17)). If all the intervals in the past window $[t^*, t]$ has a demand, then the maximum renting cost in the past window is $p_M(t - t^*)$. However the renting cost in the interval $[t^*, t+w]$ is equal to 1 and hence the renting cost in the interval $[t^*, t]$ cannot exceed 1. Therefore the maximum renting cost in the interval $[t^*, t]$ is $\min(1, p_M(t - t^*))$.
2. Due to inequality (8), there *may be* a t satisfying $t^* \leq t \leq t^* + \tau - w$ such that $p_M(t - t^*) > 1$.

Therefore if A_w buys a VM at time t then the demands in the interval $[t^*, t]$ was rented in A_w while it is served by a VM in OPT . The maximum renting cost of these demands is $\min(1, p_M(t - t^*))$. As mentioned before, $t \leq t^* + \tau - w$. Therefore the maximum renting cost of those demands which are served by a VM in OPT but rented in A_w is upper bounded by $\min(1, p_M(\tau - w))$. Given that at most N_{OPT} VMs serve such demands we have

$$\begin{aligned} \mathcal{R}(A_w \setminus OPT) &\leq \min(1, p_M(\tau - w)) N_{OPT} \\ &= \min(1, p_M \tau (1 - \alpha)) N_{OPT} \end{aligned} \quad (22)$$

where $\alpha = \frac{w}{\tau}$. Let $\delta_M = \min(1, p_M(\tau - w))$. We will use inequality (22) to upper bound $\mathcal{R}(A_w)$:

$$\begin{aligned} \mathcal{R}(A_w) &\leq \mathcal{R}(OPT) + \mathcal{R}(A_w \setminus OPT) \\ &\leq \mathcal{R}(OPT) + \delta_M N_{OPT} \end{aligned} \quad (23)$$

Let L_{OPT} be the net loss incurred in OPT . Then,

$$L_{OPT} = \mathcal{R}(OPT) + N_{OPT} \geq N_{OPT} \quad (24)$$

Similarly let L_{A_w} be the net loss incurred in A_w . We have,

$$L_{A_w} = \mathcal{R}(A_w) + N_w \quad (25)$$

$$\leq \mathcal{R}(OPT) + \delta_M N_{OPT} + N_w \quad (26)$$

$$\leq \min(1, p_{M\tau}(1 - \alpha)) N_{OPT} + N_w \quad (27)$$

Inequality (26) is obtained by substituting inequality (23) in equation (25). We now use *Lemma 2* followed by inequality (24) in inequality (27) to get

$$L_{A_w} = \delta_M N_{OPT} + N_w \quad (28)$$

$$\leq \delta_M N_{OPT} + N_{OPT} \quad (29)$$

$$\leq (1 + \delta_M) L_{OPT} \quad (30)$$

Therefore the competitive ratio of A_w is $c = 1 + \delta_M = 1 + \min(1, p_{M\tau}(1 - \alpha))$. This proves *Theorem 1*.

REFERENCES

- [1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A Berkeley view of cloud computing," EECS Department, University of California, Berkeley, Tech. Rep., 2009. [Online]. Available: <http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html>
- [2] M. Z. Hasan, E. Magana, A. Clemm, L. Tucker, and S. L. D. Gudreddi, "Integrated and autonomic cloud resource scaling," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*. IEEE, 2012, pp. 1327–1334.
- [3] B. Urgaonkar, P. Shenoy, A. Chandra, P. Goyal, and T. Wood, "Agile dynamic provisioning of multi-tier internet applications," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 3, no. 1, p. 1, 2008.
- [4] A. Ali-Eldin, J. Tordsson, and E. Elmroth, "An adaptive hybrid elasticity controller for cloud infrastructures," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*. IEEE, 2012, pp. 204–212.
- [5] X. Dutreilh, N. Rivierre, A. Moreau, J. Malenfant, and I. Truck, "From data center resource allocation to control theory and back," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. IEEE, 2010, pp. 410–417.
- [6] P. Padala, K.-Y. Hou, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, and A. Merchant, "Automated control of multiple virtualized resources," in *Proceedings of the 4th ACM European conference on Computer systems*. ACM, 2009, pp. 13–26.
- [7] L. Wang, J. Xu, M. Zhao, and J. Fortes, "Adaptive virtual resource management with fuzzy model predictive control," in *Proceedings of the 8th ACM international conference on Autonomic computing*. ACM, 2011, pp. 191–192.
- [8] E. Caron, F. Desprez, and A. Muresan, "Pattern matching based forecast of non-periodic repetitive behavior for cloud clients," *Journal of Grid Computing*, vol. 9, no. 1, pp. 49–64, 2011.
- [9] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*. IEEE, 2010, pp. 9–16.
- [10] P. Bodik, "Automating datacenter operations using machine learning," Ph.D. dissertation, University of California, Berkeley, 2010.
- [11] R. N. Calheiros and R. Buyya, "Meeting deadlines of scientific workflows in public clouds with tasks replication," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 25, no. 7, pp. 1787–1796, 2014.
- [12] M. A. Rodriguez and R. Buyya, "Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds," *Cloud Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 222–235, 2014.
- [13] H. Khazaee, J. Mistic, and V. B. Mistic, "Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 23, no. 5, pp. 936–943, 2012.
- [14] J. L. Hellerstein, Y. Diao, S. Parekh, and D. M. Tilbury, *Feedback control of computing systems*. John Wiley & Sons, 2004.
- [15] T. F. Abdelzaher, K. G. Shin, and N. Bhatti, "Performance guarantees for web server end-systems: A control-theoretical approach," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 80–96, 2002.
- [16] W. Qin and Q. Wang, "Modeling and control design for performance management of web servers via an lpv approach," *Control Systems Technology, IEEE Transactions on*, vol. 15, no. 2, pp. 259–275, 2007.
- [17] D. Ardagna, M. Trubian, and L. Zhang, "Sla based resource allocation policies in autonomic environments," *Journal of Parallel and Distributed Computing*, vol. 67, no. 3, pp. 259–270, 2007.
- [18] Z. Liu, M. S. Squillante, and J. L. Wolf, "On maximizing service-level-agreement profits," in *Proceedings of the 3rd ACM conference on Electronic Commerce*. ACM, 2001, pp. 213–223.
- [19] M. Polverini, A. Cianfrani, S. Ren, and A. V. Vasilakos, "Thermal-aware scheduling of batch jobs in geographically distributed data centers," *Cloud Computing, IEEE Transactions on*, vol. 2, no. 1, pp. 71–84, 2014.
- [20] S. Ren, "Batch job scheduling for reducing water footprints in data center," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*. IEEE, 2013, pp. 747–754.
- [21] K. Vermeersch, "A broker for cost-efficient qos aware resource allocation in ec2," Ph.D. dissertation, Master's thesis, University of Antwerp, 2011.
- [22] S. Nesmachnow, S. Iturriaga, and B. Dorransoro, "Efficient heuristics for profit optimization of virtual cloud brokers," *Computational Intelligence Magazine, IEEE*, vol. 10, no. 1, pp. 33–43, 2015.
- [23] W. Wang, D. Niu, B. Li, and B. Liang, "Dynamic cloud resource reservation via cloud brokerage," in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 400–409.
- [24] W. Wang, B. Li, and B. Liang, "To reserve or not to reserve: Optimal online multi-instance acquisition in iaas clouds," in *Proc. USENIX Intl. Conf. Autonomic Computing (ICAC)*, 2013.
- [25] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1378–1391, 2013.
- [26] T. Lu, M. Chen, and L. L. Andrew, "Simple and effective dynamic provisioning for power-proportional data centers," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 6, pp. 1161–1171, 2013.
- [27] A. Khanafer, M. Kodialam, and K. P. N. Puttaswamy, "The constrained ski-rental problem and its application to online cloud cost optimization," in *Proc. IEEE INFOCOM*, 2013, pp. 1492–1500.
- [28] M. Polverini, S. Ren, and A. Cianfrani, "Capacity provisioning and pricing for cloud computing with energy capping," in *Proc. Allerton Conference*, 2013, pp. 413–420.
- [29] I. Menache, A. Ozdaglar, and N. Shimkin, "Socially optimal pricing of cloud computing resources," in *VALUETOOLS*, 2011.
- [30] A. Karlin, C. Kenyon, and D. Randall, "Dynamic tcp acknowledgment and other stories about e/(e-1)," *Algorithmica*, vol. 36, no. 3, pp. 209–224, 2003.
- [31] R. Fleischer, "On the bahncard problem," *Theoretical Computer Science*, vol. 268, no. 1, pp. 161–174, 2001.
- [32] C. Reiss, J. Wilkes, and J. Hellerstein. (2011) Google cluster-usage traces: format+schema. [Online]. Available: <https://code.google.com/p/googleclusterdata/>
- [33] S. Di and C.-L. Wang, "Error-tolerant resource allocation and payment minimization for cloud system," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 6, pp. 1097–1106, 2013.