Fixed and Market Pricing for Cloud Services

Vineet Abhishek
University of Illinois at Urbana-Champaign
abhishe1@illinois.edu

Ian A. Kash
Microsoft Research Cambridge
iankash@microsoft.com

Peter Key
Microsoft Research Cambridge
peter.key@microsoft.com

Abstract: This paper considers two simple pricing schemes for selling cloud instances and studies the trade-off between them. We characterize the equilibrium for the hybrid system where arriving jobs can choose between fixed or the market based pricing. We provide theoretical and simulation based evidence suggesting that fixed price generates a higher expected revenue than the hybrid system.

1 Introduction

Cloud computing provides on-demand and scalable access to computing resources. Public clouds, such as Windows Azure and Amazon EC2, treat infrastructure computing as a service that can be purchased and delivered over the Internet. A user purchases units of computing time on virtual machines (referred to as *instances*). The most commonly used pricing mechanism for instances is *pay as you go* (henceforth, PAYG), where a user is charged a fixed price per unit time per instance. However, given stochastic demand, such fixed pricing may result in unused resources. Rather than letting resources sit idle, the provider could operate a *spot market*, selling unused resources at a reduced price via using auction to users willing to tolerate delays and interruptions.

This paper examines the tradeoffs for a provider deliberating whether or not to operate a spot market. On one hand, operating a spot market can create price discrimination, as users with low values and low waiting costs compete for spot instances, thereby extracting payments from the users who would balk if PAYG were the only option. On the other hand, the spot market provides a cheaper alternative to users with high value but low waiting cost, causing a loss of revenue from the users who would have paid a higher PAYG price if PAYG were the only option. In consequence, it is not obvious if operating PAYG and the spot market simultaneously provides any net gain in the expected revenue to the cloud service provider.

To quantify the trade-offs we construct a simple model of a cloud computing service with users who are heterogeneous both in their value for service and in their waiting cost. We first analyze PAYG and a spot market in isolation and use the resulting insights to analyze what happens when they operate simultaneously. Our analysis is not tied to any particular pricing rule for the spot market. Instead, we use a characterization similar to the revenue equivalence theorem for auctions [12] to compute the expected payment made by a user in any equilibrium of any pricing rule. Moreover, while the analysis of the queuing system with multiple priority classes and multiple servers is complex (see, e.g., [7], [13]), an application of the revelation principle [12] allows us to circumvent this complexity. We describe a general queuing system for the spot market purely in terms of a waiting time function and exploit its properties for our analysis. Our main contributions are:

- We model a cloud computing service as a queuing system described by a waiting time function and apply techniques from the theory of optimal auctions to analyze it.
- We show that, in equilibrium, users have a waiting cost threshold that determines whether they participate in the spot market or PAYG. Moreover, the expected payments by the users in the spot market are independent of their value for service and increasing in their waiting cost[1].
- Using this equilibrium characterization, we provide theoretical and simulation evidence suggesting that operating PAYG in isolation provides a higher expected revenue to the cloud service provider than operating PAYG and a spot market simultaneously.

Our work is at the nexus of queuing theory and game theory. Hassin and Haviv [9] provide a survey of this area. For observable $M/M/1$ queues with identical customers, Balachandran [4] derives a full information equilibrium strategy. Hassin [8] and Lui [11] consider unobservable $M/M/1$ queues where customers with heterogeneous waiting cost bid for preemptive priority using the first price auction. They characterize an equilibrium where bids are increasing in the waiting cost. Afèche and Mendelson [3] extend this to a more general waiting cost function. Dube and Jain [6] consider a different problem with competing $GI/GI/1$ priority queues; arriving jobs decide which queue to join. They find conditions for the existence of the Nash equilibrium. Perhaps the closest to our work are papers that apply the theory of optimal auction design to optimize pricing and service policies in queuing system. Afèche [2] shows that delaying jobs or choosing orderings that increase processing time can increase revenue. Yahalom et al. [14] generalize [2] by relaxing the distributional assumptions on valuation and working with convex delay cost. Katta and Sethuraman [10] design a pricing scheme that, under some assumptions, is optimal for an M/M/1 queuing system and certain generalizations of it. Cui et al. [5] move beyond admission control through priority pricing. Instead, they consider the problem of joint pricing, scheduling, and admission control policy for revenue maximization for $M/M/1$ queue and find solutions for some special cases. Compared to previous work in this lit-

[1] Throughout this paper, "increasing" means "strictly increasing."

erature, the distinguishing aspects of our work are: (i) we allow for an arbitrary queuing system with multiple servers and arrival process which need not be memoryless; (ii) our analysis is not tied to a specific auction mechanism for the spot market; (iii) we allow PAYG and the spot market to operate simultaneously and are not limited to analyzing a system in isolation.

2 Model

Consider a cloud computing system where jobs arrive sequentially according to a stationary stochastic process with independent interarrival time. Each job demands one instance and is associated with a distinct user. We will use the terms "users" and "jobs" interchangeably. The service time for each job is independently drawn according to an arbitrary distribution with the expected time of $1/\mu$. Jobs differ in their values for service and the waiting costs. There are two classes of jobs. Each job from class i has the same value v_i for job completion. Assume $v_1 > v_2$. The total arrival rate of potential jobs is $\lambda_1 + \lambda_2$. Each job is independently assigned class i with probability $\lambda_i/(\lambda_1 + \lambda_2)$, hence the total arrival rate of potential jobs from class i in λ_i. Each job from class i incurs a waiting cost per unit time which is an i.i.d. realization of a random variable C_i with the cumulative distribution function (CDF) $F_i(c)$. The exact waiting cost of a job is its private information; however, the probability distributions F_i's are common knowledge. The random variable C_i's are independent of each other. If a job from class i with waiting cost c pays a total price m for using the instance and spends the total time w in the system (the sum of the queuing time and the service time, referred to as the *waiting time*) then its payoff is $v_i - cw - m$. A job wants to maximize its expected payoff from using an instance and competes to acquire an instance only if its expected payoff is nonnegative. Let $f_i(c)$ be the corresponding probability density function (pdf) of $F_i(c)$; $f_i(c)$ is assumed to be strictly positive for $c \in [0, \mu v_i]$[2]. Each job is infinitesimally small and cannot affect the system dynamics on its own.

Modeling PAYG: PAYG is modeled as a $GI/GI/\infty$ system with service rate μ. A job arriving to PAYG joins immediately and is served until completion. Each job is charged a price $p > 0$ per unit time for using a PAYG instance. The price p is known to everyone a priori. The expected payoff of a job from class i with the waiting cost c from using a PAYG instance is thus $v_i - (c + p)/\mu$. If $c > \mu v_i - p$, the job does not participate in PAYG.

Modeling the spot market: The spot market is modeled as a $GI/GI/k$ system with preemption where jobs bid for priority. We will be mostly working with auctions where a job with a higher bid is given priority over a job with a lower bid and can preempt the lowest priority job under service if needed; Section 3.1 provides further details on the assumptions we make on the relationship between bids and priorities. A job which is preempted goes back to the queue and waits to resume from the point it left. The queue state (i.e., the bid vector in the spot market) is unobservable to the arriving jobs. Jobs are not allowed to renege or change their bids. A job is charged based on its own bid and the bids of others according to some spot pricing mechanism. Examples include the first price auction where jobs with k highest bids are served and each pays its bid, and the $(k+1)^{th}$ price auction where the jobs with k highest bids are served and each job pays the $(k+1)^{th}$ highest bid. We do not explicitly assume any specific spot pricing mechanism and abstract away from it by considering the expected payment by a job in a Bayes Nash Equilibrium (henceforth, BNE) using the revenue equivalence theorem for auctions [12].

3 PAYG and Spot Market Analysis

3.1 Strategy, waiting time, and spot pricing

When a spot market is operating, either alone or in conjunction with PAYG, a job that decides to join it participates in an auction and must decide how much to bid based on the payment rules of the auction. The optimal bid may depend in a complicated way on its private information (value for service and cost of waiting). However, we show in this section that this complexity is inessential. Regardless of the auction mechanism, jobs that enter the spot market with higher waiting costs pay more and wait less time and these values are (essentially) independent of the job's class. The job's class does matter in determining whether the job participates in the spot market, but this take the form of a simple cutoff with jobs with waiting costs below the cutoff participating and those with costs above not.

By the revelation principle for BNE [12], it suffices to restrict our consideration to truthful direct revelation mechanisms: mechanisms where jobs report their private information and it is an equilibrium for them to do so truthfully. Any implementable outcome is implementable by such a mechanism. Thus, a job reports a type (v, c); if it participates in the spot market, it has an expected waiting time $\widetilde{w}(v, c)$ and expected payment $\widetilde{m}(v, c)$. In principle, these could depend on the value v of the job's class, however, we show that it is essentially without loss of generality to assume they do not.

LEMMA 1. *For all truthful direct revelation mechanisms for the spot market and all equilibria, there exists an equilibrium with the same expected utility where expected waiting time and payments are independent of class for all values of c where both classes participate in the spot market.*

PROOF. A job of class i with waiting cost c that participates in the spot market chooses a report (v', c') to minimizing the expected total cost $c\widetilde{w}(v', c') + \widetilde{m}(v', c')$. Thus, when both classes participate, the set of optimal reports is class-independent; in particular, both (v_1, c) and (v_2, c) belong to the set of optimal reports. Let s_1 and s_2 be the (randomized) equilibrium strategies for class 1 and class 2 with cost c. Now, suppose that the job of class i with waiting cost c uses

[2] Jobs from class i with waiting cost greater than μv_i will always balk, hence we restrict attention to range $[0, \mu v_i]$.

strategy s_1 with probability $\lambda_1 f_1(c)/(\lambda_1 f_1(c) + \lambda_2 f_2(c))$ and strategy s_2 otherwise. Then the arrival process for the strategies s_1 and s_2 remains identical to the original process, hence the waiting time and the expected payment remain unchanged. This new class-independent randomized strategy is also an equilibrium for both classes. □

Since jobs can undo any tie-breaking the mechanism does on the basis of class, we assume for the remainder of the paper that mechanisms have a class-independent expected waiting time $\widetilde{w}(c)$ and expected waiting cost $\widetilde{m}(c)$. As we are interested in what outcomes are implementable, again by the revelation principle it is without loss of generality to assume that jobs report truthfully and we do so for the remainder of the paper. We now show that jobs with higher waiting costs pay more and spend less time waiting.

LEMMA 2. *In (the truthful) equilibrium, $\widetilde{w}(c)$ is nonincreasing in c and $\widetilde{m}(c)$ is nondecreasing in c for values of c that participate in the spot market for some class.*

PROOF. Consider $\widehat{c} > c$. The optimality of truthful reporting implies:

$$\widehat{c}\widetilde{w}(\widehat{c}) + \widetilde{m}(\widehat{c}) \leq \widehat{c}\widetilde{w}(c) + \widetilde{m}(c), \tag{1}$$

$$c\widetilde{w}(c) + \widetilde{m}(c) \leq c\widetilde{w}(\widehat{c}) + \widetilde{m}(\widehat{c}). \tag{2}$$

Adding (1) and (2) implies $\widetilde{w}(\widehat{c}) \leq \widetilde{w}(c)$. Using this and (2), we get $\widetilde{m}(\widehat{c}) \geq \widetilde{m}(c)$. □

Thus far, our assumptions have been without loss of generality. We now make two assumptions that are not. First, we assume that jobs with no waiting cost are served for free in the spot market, hence $m(0) = 0$. Second, we assume that, in equilibrium in the spot market, jobs with higher waiting costs always have strictly higher priority than jobs with lower waiting costs . Note that this is a stronger condition than assuming that $\widetilde{w}(c)$ is decreasing. Since $\widetilde{w}$ is the expected waiting time, if priorities are assigned randomly it is possible to have a a strictly lower expected waiting time but in some cases a lower priority. All mechanisms that assign a strictly higher priority to the jobs with higher bids in the spot market, admit an equilibrium where the spot market bids are increasing in the waiting cost, and have no reserve price satisfy these restrictions. For example, we show later in this section that the first price auction satisfies these properties.

We now characterize the participation decision facing jobs.

LEMMA 3. *For each class i there is a cutoff c_i below which jobs participate in the spot market and above which they do not.*

PROOF. A job participates in the spot market if the payoff is better than its alternative (0 if the spot market is operated in isolation or $\max\{0, v_i - (p + c)/\mu\}$ if PAYG with price p is available). The payoff from participation is $v_i - c\widetilde{w}(c) - \widetilde{m}(c)$. Let c be any type that participates. Taking the case of the spot market in isolation first, if $v_i - c\widetilde{w}(c) - \widetilde{m}(c) \geq 0$ then $v_i - \widehat{c}\widetilde{w}(c) - \widetilde{m}(c) > 0$ for all $\widehat{c} < c$. Thus, if a job of class i with cost c participates, all lower cost jobs do as well. This argument also implies that if a job with waiting cost c does not participate, then neither does any job with waiting cost $\widehat{c} > c$. Thus, there is some threshold c_i below which jobs participate and above which they do not. The argument with PAYG as an option is essentially the same because the minimum possible value of $\widetilde{w}(c)$ is $1/\mu$, the same as the waiting time under PAYG. □

In order to characterize an equilibrium where jobs use cutoffs (c_1, c_2), we need to analyze the expected waiting time for a job with waiting cost c in the spot market with cutoffs (c_1, c_2). It suffices to characterize some properties of the waiting times for arbitrary choices of cutoffs. Given a queuing system for the spot market, define the waiting time function $w(c; c_1, c_2)$ as the expected waiting time of a job with cost c when jobs of class i use cutoff c_i. Note that we are defining w for arbitrary cutoffs, not just equilibrium ones. The following lemma gives the relevant properties of w.

LEMMA 4. *The waiting-time function $w(c; c_1, c_2)$ is well defined whenever $(\sum_{i=1,2} \lambda_i F_i(c_i))/(k\mu) < 1$. It is an increasing function of $\sum_{i=1,2} \lambda_i \left[F_i(c_i) - F_i(c)\right]^+$. In particular, this implies:*

(i) $w(c; c_1, c_2)$ is decreasing in c for $c \in [0, c_1 \vee c_2]$[3]*, $w(c; c_1, c_2) > 1/\mu$ if $c < c_1 \vee c_2$, and $w(c; c_1, c_2) = 1/\mu$ if $c \geq c_1 \vee c_2$.*

(ii) $w(c; c_1, c_2)$ is increasing in c_1 and c_2 for $c_i \in [0, \mu v_i]$.

(iii) For any $c_1 > \widehat{c}_2 > c_2$ and $t \in [\widehat{c}_2, c_1]$, $w(t; c_1, c_2) = w(t; c_1, \widehat{c}_2)$.

PROOF. The condition $(\sum_{i=1,2} \lambda_i F_i(c_i))/(k\mu) < 1$ ensures the queue is stable so that the expected waiting time is finite. This must be true in equilibrium. Since priority is given to the job with a higher waiting cost, the expected waiting time of a job with waiting cost c increasing with the total arrival rate of the jobs with waiting cost higher than c, which is equal to $\sum_{i=1,2} \lambda_i \left[F_i(c_i) - F_i(c)\right]^+$. The job with waiting cost greater than or equal to $c_1 \vee c_2$ gets the highest priority and is served immediately with no interruptions. The enumerated properties follow easily. □

Next, we use a characterization similar to the revenue equivalence theorem for auctions [12] and show that the expected payment by any job with waiting cost c is uniquely determined by the waiting time function w; in particular, it is the same for any spot pricing mechanism.

Suppose that the truthful reporting with cutoffs (c_1, c_2) constitutes a BNE for the given spot pricing mechanism. Let $m(c)$ be the expected payment made by a job with waiting cost c (the expected payment is independent of its class). For a BNE to exist, the following *incentive compatibility* (henceforth, IC) constraint must hold: for any $\widehat{c}, c \leq c_1 \vee c_2$, and

[3] Here, $a \vee b = \max\{a, b\}$.

any i,

$$v_i - cw(c; c_1, c_2) - m(c) \geq v_i - cw(\widehat{c}; c_1, c_2) - m(\widehat{c}). \quad (3)$$

By analogy with [12], the next lemma relates the expected payment with the waiting time function w and shows that the properties of the waiting time function along with the expected payment given by (4) ensure that the IC constraint (3) is satisfied. The proof can be found in the full version of this paper [1].

LEMMA 5. *A necessary condition for* (3) *to hold is:*

$$m(c) = \int_0^c w(t; c_1, c_2)dt - cw(c; c_1, c_2). \qquad (4)$$

Hence, the expected payment by a job with waiting cost c is uniquely determined by the function w and is same for all spot pricing mechanisms that satisfy our assumption that $m(0) = 0$. Moreover, Lemma 4 and (4) *together satisfy the IC constraint* (3)*.*

Since $w(c; c_1, c_2)$ is decreasing in c for $c \in [0, c_1 \vee c_2]$, the proof of Lemma 2 can be used to establish a stronger monotonicity of the expected payment m.

LEMMA 6. *Given cutoffs (c_1, c_2), the expected payment $m(c)$ is increasing in c for $c \in [0, c_1 \vee c_2]$.*

3.2 Revenue and equilibria for isolated markets

First consider PAYG in isolation. If PAYG price is p, a job from class i with waiting cost c obtains an expected payoff $v_i - (p + c)/\mu$ by using a PAYG instance. A job will participate in PAYG if this payoff is nonnegative. Thus, a job from class i participates in PAYG if its waiting cost $c \leq \mu v_i - p$. The effective arrival rate of class i jobs is then $\lambda_i F_i(\mu v_i - p)$ where $F_i(\mu v_i - p) = 0$ if $p \geq \mu v_i$. Each such job uses a PAYG instance for an expected duration of $1/\mu$ and pays p per unit time. Hence, the expected revenue to the cloud service provider per unit time, denoted by $R^{payg}(p)$, is:

$$R^{payg}(p) \triangleq \frac{p}{\mu}\Bigg(\sum_{i=1,2} \lambda_i F_i(\mu v_i - p)\Bigg), \qquad (5)$$

and the optimum revenue is $\max_p R^{payg}(p)$.

Next, consider the spot market in isolation. Given the cutoffs (c_1, c_2), the expected payment by a job with waiting cost c in any BNE is given by (4). Thus, we need to compute the cutoff for each class i when the spot market is operated in isolation; denote the cutoffs in this case by $\mathbf{c}^s \triangleq (c_1^s, c_2^s)$. From (4), the expected payoff of a job from class i with waiting cost c is $v_i - \int_0^c w(t; \mathbf{c}^s)dt$. A job will participate in the spot market as long as its expected payoff is nonnegative. Hence, the cutoff vector $\mathbf{c}^s$ must satisfy:

$$v_i - \int_0^c w(t; \mathbf{c}^s)dt \begin{cases} \geq 0 & \text{if } c < c_i^s, \\ = 0 & \text{if } c = c_i^s. \end{cases} \qquad (6)$$

Theorem 1 below shows that there is an unique cutoff vector $\mathbf{c}^s$ satisfying (6) and uses it to characterize the BNE for the spot market in isolation. The proof can be found in the full version of this paper [1].

THEOREM 1. *The following holds:*

(i) There is a unique solution to the following system of equations in (x_1, x_2):

$$\begin{aligned} \int_0^{x_1} w(t; x_1, x_2)dt = v_1, \\ \int_0^{x_2} w(t; x_1, x_2)dt = v_2. \end{aligned} \qquad (7)$$

(ii) Choose the cutoff vector $\mathbf{c}^s$ as the unique solution of (7)*. Then $\mathbf{c}^s$ satisfies* (6)*, $c_1^s \geq \overline{c}$, and $c_2^s \leq \overline{c}$. Here $\overline{c}$ uniquely satisfies $\int_0^{\overline{c}} w(t; \overline{c}, \overline{c})dt = v_2$.*

(iii) In all BNE, a job from class i with waiting cost c participates in the spot market if and only if $c \leq c_i^s$.

To highlight the explicit dependence of the expected payment on the cutoffs vector $\mathbf{c}^s$, we use $m(c; \mathbf{c}^s)$; i.e,

$$m(c; \mathbf{c}^s) = \int_0^c w(t; \mathbf{c}^s)dt - cw(c; \mathbf{c}^s). \qquad (8)$$

Using Theorem 1, the expected revenue to the cloud service provider per unit time when the spot market is operated in isolation, denoted by R^s, is:

$$R^s \triangleq \sum_{i=1,2} \lambda_i \int_0^{c_i^s} m(t; \mathbf{c}^s) f_i(t)dt. \qquad (9)$$

3.3 Revenue and equilibria in the hybrid market

We now leverage the insights gained from analyzing PAYG and the spot market each in isolation and move to analyzing the hybrid system where both are operated simultaneously. As mentioned in Section 3.1, for a given PAYG price p, we look for a cutoff vector $\mathbf{c}(p) \triangleq (c_1(p), c_2(p))$ such that a job from class i with waiting cost c joins the spot market if and only if $c < c_i(p)$, and if so, it reports its waiting cost truthfully; otherwise it joins PAYG as long as $c \leq \mu v_i - p$ (the cutoff for class i if PAYG is operating in isolation).

A job from class i with waiting cost c gets the expected payoff $v_i - \int_0^c w(t; \mathbf{c}(p))dt$ from using a spot instance and reporting its waiting cost truthfully, while its expected payoff from using a PAYG instance is $v_i - (p + c)/\mu$. It will pick the one which offers a higher expected payoff. If the PAYG price is too high for a class, then no jobs from that class goes to PAYG. Theorem 2 below finds the unique cutoff vector $\mathbf{c}(p)$ and uses it to characterizes the BNE of the hybrid system. The proof can be found in the full version of this paper [1].

THEOREM 2. *Let $\overline{c}$ and $\mathbf{c}^s$ be as given by Theorem 1 and p be a PAYG price. Choose the cutoff vector $\mathbf{c}(p)$ as follows:*

(i) If $p \in (0, \mu v_2 - \overline{c}]$, then there is a unique $x \in [0, \overline{c}]$ satisfying $(p + x)/\mu = \int_0^x w(t; x, x)dt$. Choose $c_1(p) = c_2(p) = x$. Each $c_i(p) \in (0, \overline{c}]$ and is increasing in p.

(ii) If $p \in (\mu v_2 - \overline{c}, \mu v_1 - c_1^s]$, then there is a unique (x_1, x_2) such that $x_1 \geq x_2$ that satisfies the following system of equations:

$$\begin{aligned} \int_0^{x_1} w(t; x_1, x_2)dt = \tfrac{p + x_1}{\mu}, \\ \int_0^{x_2} w(t; x_1, x_2)dt = v_2. \end{aligned} \qquad (10)$$

Choose $c_1(p) = x_1$ and $c_2(p) = x_2$. $c_1(p) \in (\bar{c}, c_1^s]$ and is increasing in p, $c_2(p) \in [c_2^s, \bar{c})$ and is decreasing in p, and $\sum_{i=1,2} \lambda_i F_i(c_i(p))$ is increasing in p.

(iii) If $p > \mu v_1 - c_1^s$, choose $c_1(p) = c_1^s$ and $c_2(p) = c_2^s$.

Then in any BNE, a job from class i with waiting cost c participates in the spot market if and only if $c < c_i(p)$, it participates in PAYG if $c_i(p) \leq c \leq \mu v_i - p$. If $\mu v_i - p < c_i(p)$ then no class i job participates in PAYG[4].

Our analysis so far characterizes a truthful BNE for the system where PAYG and the spot market are operating simultaneously. This equilibrium can be implemented by assigning higher priority to the jobs with the higher waiting cost and collecting the payment according to (4). In the first price auction, the bid is same as the payment; a byproduct of our analysis is that the payment rule (4) and cutoffs given by Theorem 2 characterize the bidding strategy if the first price auction is used for the spot market.

The expected revenue to the cloud service provider per unit time is the sum of expected revenue from the spot market and PAYG. From (5), (9), and Theorem 2, given a PAYG price p, the expected revenue per unit time for the hybrid system, denoted by $R^h(p)$, is:

$$R^h(p) \triangleq \sum_{i=1,2} \lambda_i \bigg(\frac{p}{\mu} \left[F_i(\mu v_i - p) - F_i(c_i(p))\right]^+ + \int_0^{c_i(p)} m(t; \mathbf{c}(p)) f_i(t) dt \bigg), \quad (11)$$

and the optimum revenue is $\max_p R^h(p)$.

The next theorem provides theoretical evidence suggesting that PAYG in isolation can provide a higher expected revenue to the cloud service provider than operating PAYG and the spot market simultaneously.

THEOREM 3. *Suppose the optimal price p^h of the hybrid system is such that $p^h \leq \mu v_2 - \bar{c}$, i.e., case (i) of Theorem 2 holds. Then the optimum expected revenue per unit time from PAYG in isolation is higher than the optimum expected revenue per unit time from the hybrid system; i.e., $\max_p R^h(p) = R^h(p^h) < \max_p R^{payg}(p)$.*

PROOF. It suffices to show that $R^{payg}(p^h) > R^h(p^h)$.

If $p^h \leq \mu v_2 - \bar{c}$, then $c_1(p^h) = c_2(p^h) \leq \bar{c}$, implying $\mu v_i - p^h \geq \bar{c} \geq c_i(p)$. Then from (5) and (11),

$$R^{payg}(p^h) - R^h(p^h) = \sum_{i=1,2} \lambda_i \bigg(\frac{p^h}{\mu} F_i(c_i(p^h)) - \int_0^{c_i(p^h)} m(t; \mathbf{c}(p^h)) f_i(t) dt \bigg). \quad (12)$$

At $c = c_i(p^h)$, a job is indifferent between PAYG and the spot market. Hence,

$$c_i(p^h) w(c_i(p^h); \mathbf{c}(p^h)) - m(c_i(p^h); \mathbf{c}(p^h)) = \frac{c_i(p^h) + p^h}{\mu}.$$

Since $c_1(p^h) = c_2(p^h)$, $w(c_i(p^h); \mathbf{c}(p^h)) = 1/\mu$. Hence, $m(c_i(p^h); \mathbf{c}(p^h)) = p^h/\mu$. From Lemma 6, $m(t; \mathbf{c}(p^h))$ is increasing in t for $t \in [0, c_i(p)]$. This and (12) imply:

$$R^{payg}(p^h) - R^{hybrid}(p^h) > \sum_{i=1,2} \lambda_i \bigg(\frac{p^h}{\mu} F_i(c_i(p^h)) - \int_0^{c_i(p)} \frac{p^h}{\mu} f_i(t) dt \bigg) = 0. \quad (13)$$

This completes the proof. □

4 Simulations

The revenue ranking result of Theorem 3 is for the case when the optimal price p^h of the hybrid system is such that $p^h \leq \mu v_2 - \bar{c}$. However, we conjecture that the revenue ranking result holds in general and present simulation evidence.

We model the spot market as k parallel $M/M/1$ queues. Jobs bid for preemptive priorities using the first price auction. An arriving job is randomly and uniformly sent to one of the k queues where it is served according to its priority order, determined by its bid, in that queue. We extend the results from [11] to compute the waiting time function:

$$w(c; c_1, c_2) = \frac{1}{\mu \big(1 - \sum_{i=1,2} \rho_i \left[F_i(c_i) - F_i(c)\right]^+ \big)^2}, \quad (14)$$

where $\rho_i \triangleq \lambda_i/(k\mu)$. Recall that the payment rule (4) and cutoffs given by Theorem 2 characterize the bidding strategy for the first price auction for the spot market. The proof of Theorem 2 provides a recipe for numerically computing the cutoff vector $\mathbf{c}(p)$ as a function of PAYG price p.

Simulations are carried out by randomly generating the values of v_i's, λ_i's, and k. The service rate μ is kept constant at one and F_i is uniform in the interval $[0, \mu v_i]$. We generate over a hundred random configurations (v_i's, λ_i's, and k). For each realized configuration, we observe that the optimal revenue from PAYG in isolation is always higher than the optimal revenue from the hybrid system where PAYG and the spot market are operating simultaneously, even for the case where the optimal price p^h of the hybrid system is greater than $\mu v_2 - \bar{c}$. An example plot where $p^h > \mu v_2 - \bar{c}$ is shown in Figure 1. Observe that if PAYG price is low, most of the jobs in the hybrid system use PAYG and pay a small price, leading to a small expected revenue. As PAYG price increases, jobs move to the spot market, reaching a point where all jobs use the spot market. At $p = \mu v_2$, the entire class 2 jobs balk from PAYG leading to a kink in the plot for PAYG in isolation. Simulations with exponentially distributed waiting costs are also consistent the revenue ranking that we conjecture.

[4] It is assumed that jobs break ties between the spot market and PAYG in favor of PAYG.

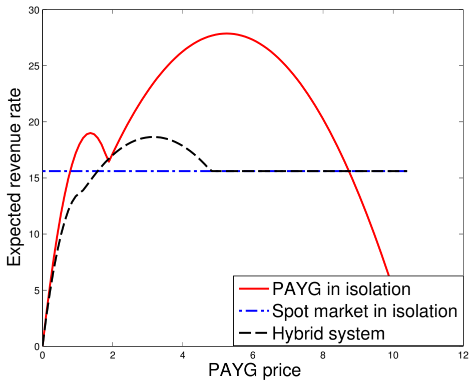

Figure 1: *Expected revenue from PAYG in isolation, the spot market in isolation, and the hybrid system as a function of PAYG price.*

5 Discussion and Future Work

Our analysis in Section 3 characterizes a truthful BNE for the system where PAYG and the spot market are operating simultaneously. Our theoretical results show that in many cases the revenue raised by a PAYG system in isolation with a well chosen price p dominates that of this hybrid system. Simulations suggest that this may be true in general. However, this analysis was based on a number of assumptions. We conclude by discussing how relaxing them affects our results, which points to several areas for future work.

- We assumed that the PAYG system has infinite capacity, which we believe is reasonable given that capacity is endogenous and PAYG jobs are more profitable than spot market jobs. However, it would also be good to understand what happens in situations where this is not the case. In cases with excess demand for PAYG instances, jobs with high value and high delay cost can compete for the spot instances, possibly paying a price higher than the PAYG price. However, this can populate the spot instances and increase the waiting time, possibly causing some low value jobs to balk all together.
- We assumed that the arrival process is independent of job type. This may not be true if both arrival pattern and value depend on underlying characteristics of the job. In this case, it is possible that there are equilibria where jobs of different classes but the same cost have different outcomes. However, as both classes have the same set of optimal outcomes, this requires an amount of coordination on tiebreaking that may be unreasonable in practice.
- Because jobs can get interrupted in the spot market, programmers may need to write more robust code and interruption may be unsuitable for tasks that require high availability. This can be modeled as an upfront cost of participating in the spot market. Are there reasonable scenarios where this makes a hybrid system optimal?
- We assumed that $m(0) = 0$. Choosing a larger value amounts to setting a reserve price. The equilibrium structure would be similar, although the cutoffs would change and there are additional cases. Our theoretical revenue analysis still holds despite a reserve price.
- We assumed that the higher priority is given to the jobs with higher waiting cost and use this to derive properties of the waiting time function w. This excludes systems where, in equilibrium, a variety of types pay the same expected price and receive the same expected waiting time (PAYG could be viewed as an example of this). While this would require a more general equilibrium characterization, our theoretical revenue analysis still applies.
- Our analysis is for a monopolistic provider. The effect of competitive pressures needs to be investigated.

6 References

[1] V. Abhishek, I. A. Kash, and P. Key. Fixed and market pricing for cloud services. Working paper. Manuscript at http://arxiv.org/abs/1201.5621.

[2] P. Afche. Incentive-compatible revenue management in queueing systems: optimal strategic delay and other delay tactics. Working paper., August 2004.

[3] P. Afche and H. Mendelson. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science*, 50(7):pp. 869–882, 2004.

[4] K. R. Balachandran. Purchasing priorities in queues. *Management Science*, 18(5):319–326, 1972.

[5] T. Cui, Y.-J. Chen, and Z.-J. M. Shen. Pricing, scheduling, and admission control in queueing systems: A mechanism design approach. Submitted to Operations Research.

[6] P. Dube and R. Jain. Queueing game models for differentiated services. In *Game Theory for Networks, 2009. GameNets '09. International Conference on*, pages 523 –532, may 2009.

[7] M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman. Multi-server queueing systems with multiple priority classes. *Queueing Systems*, 51:331–360, 2005. 10.1007/s11134-005-2898-7.

[8] R. Hassin. Decentralized regulation of a queue. *Management Science*, 41(1):163–173, 1995.

[9] R. Hassin and M. Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Springer, 1 edition, November 2002.

[10] A.-K. Katta and J. Sethuraman. Pricing strategies and service differentiation in queues a profit maximization perspective. Working paper., March 2005.

[11] F. T. Lui. An equilibrium queuing model of bribery. *Journal of Political Economy*, 93(4):760–781, 1985.

[12] R. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

[13] M. van der Heijden, A. van Harten, and A. Sleptchenko. Approximations for markovian multi-class queues with preemptive priorities. *Operations Research Letters*, 32(3):273 – 282, 2004.

[14] T. Yahalom, J. M. Harrison, and S. Kumar. Designing and pricing incentive compatible grades of service in queueing systems. Working paper, January 2006.