Fixed and Market Pricing for Cloud Services

Vineet Abhishek · Ian Kash · Peter Key

January 9, 2017

Abstract We study a model of congestible resources, where pricing and scheduling are intertwined. Motivated by the problem of pricing cloud instances, we model a cloud computing service as linked $GI/GI/\cdot$ queuing systems where the provider chooses to offer a fixed pricing service, a dynamic market based service, or a hybrid of both, where jobs can be preempted in the market-based service. Users (jobs), who are heterogeneous in both the value they place on service and their cost for waiting, then choose between the services offered. Combining insights from auction theory with queuing theory we are able to characterize user equilibrium behavior, and show its insensitivity to the precise market design mechanism used. We then provide theoretical and simulation based evidence suggesting that a fixed price typically, though not always, generates a higher expected revenue than the hybrid system for the provider.

Keywords Cloud Services · Spot Market · Pricing · Congestion Resources · Auctions

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 (IaaS) that can be purchased and delivered over the Internet. An agent purchases units of computing time on virtual machines (referred

This work was done while Vineet Abhishek was interning at Microsoft Research

Vineet Abhishek
Walmart Labs, USA
E-mail: vineet.abhishek@gmail.com

Ian Kash
Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, UK,
E-mail: iankash@microsoft.com

Peter Key
Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, UK,
E-mail: peter.key@microsoft.com

to as *instances*). The most commonly used pricing mechanism for instances is *pay as you go* (henceforth, PAYG), where an agent 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 an auction to agents willing to tolerate delays and interruptions; indeed, Amazon EC2 runs such a market for Spot Instances. Ben-Yehuda et al. [4] attempt a retrospective deconstruction of how the Spot Instances are priced.

This paper examines the trade-offs for a provider deciding whether or not to operate a spot market. On one hand, operating a spot market can create price discrimination, as agents with low valuations and low waiting costs compete for spot instances, thereby extracting payments from the agents who would balk if PAYG were the only option. On the other hand, the spot market provides a cheaper alternative for agents with high valuations but low waiting cost, causing a loss of revenue from the agents 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. Nor is it clear how operating a spot market affects social welfare: while adding a spot market causes some new agents to be served, some existing agents that were already being served will now join the spot market, increasing their waiting costs. Furthermore, if the cloud services provider desires to extract a given amount of revenue, while maximizing welfare subject to that constraint, he would set different prices in each system.

To quantify the trade-offs we construct a simple, representative model of a cloud computing service with agents 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 [18] to characterize the expected payment made by an agent 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., [10], [13]) an application of the revelation principle [18] 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. Throughout our paper, by "waiting time"' we mean the waiting time in the *system* or sojourn time.

In summary, we make four contributions in this paper:

(i) We combine insights from auction theory with queuing theory: we model a cloud computing service as a queuing system described by a waiting time function, and then apply techniques from the theory of optimal auctions to analyze it.

(ii) We characterize agent behavior, and show that, in the unique equilibrium, agents have a waiting cost threshold that determines whether they participate in the spot market or PAYG. Moreover, their bids in the spot market are independent of their value for service and increasing in their waiting cost[1].

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

(iii) Using this equilibrium characterization, we provide theoretical and simulation evidence suggesting that operating PAYG in isolation generally, though not always, provides a higher expected revenue to the cloud service provider than operating PAYG and a spot market simultaneously. This is under the assumption that there is no other cost to running a job in the spot market, e.g. that there is not cost in being preempted.
(iv) We prove that, taking the PAYG price as a real cost (not just a transfer), agents make *efficient* decisions about whether to join the spot market; we give simulation evidence describing the tradeoffs between efficiency in revenue under each system.

While our results are based on a stylized model of a cloud computing system, we also discuss how the assumptions of our model can be relaxed and the implications of our results for the decision Amazon has made to run a spot market.

1.1 Related Work

Our work is at the nexus of queuing theory and game theory. Hassin and Haviv [12] provide a survey of this area. For observable $M/M/1$ queues with identical customers, Balachandran [6] derives a full information equilibrium. Hassin [11] and Lui [16] consider unobservable $M/M/1$ queues where customers with heterogeneous waiting costs bid for preemptive priority using the first price auction. They characterize an equilibrium where bids are increasing in the waiting cost. Afèche and Mendelson [2] extend this to more general waiting costs. Dube and Jain [9] 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 a Nash equilibrium.

Closest to our work are papers that apply the theory of optimal auction design to optimize pricing and service policies in queuing system. Afèche [1] and Afèche and Pavlin [3] show that delaying jobs or choosing orderings that increase processing time can increase revenue. Yahalom et al. [22] generalize [1] by relaxing the distributional assumptions on valuation and working with convex delay cost. Katta and Sethuraman [14] 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. [7] consider the problem of jointly managing pricing, scheduling, and admission control policy for revenue maximization for $M/M/1$ queues and find solutions for some special cases. Xu and Li [21] examine possibilities to improve revenue in a PAYG market through resource throttling. Doroudi et al.[8] study pricing in an $M/G/1$ queue for a single class, and a continuum of customers in that class. They assume that waiting costs are proportional to valuations, which allows customer types to be unidimensional and hence the theory of optimal (revenue maximizing) auctions that flowed from Myerson's seminal work [18] can be applied. They derive closed form solution for the price functions when the customer values are drawn from specific types, and compare priority pricing with fixed pricing. While their results show something similar in spirit to spot pricing raising more revenue than PAYG, this is driven by an assumption that their PAYG market is oversubscribed while this is not the case for us.

One issue we do not address is competition among cloud providers, an issue studied by Anselmi et al. [5].

Compared to previous work in this literature, 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; and (iv) we examine the tradeoff between efficiency and revenue.

2 Model

Consider a cloud computing system where jobs arrive sequentially according to a stationary stochastic process with independent interarrival times. Each job demands one instance and is associated with a distinct agent; we will use the terms "agents" and "jobs" interchangeably. The system designs the pricing and scheduling mechanism, with the aim of maximizing revenue, while the jobs aim to maximize their expected payoff.

The service time for each job is independently drawn according to an arbitrary distribution with the expected time of $1/\mu$, where we assume that the exact service time is unknown to anyone, including the job itself. Jobs differ in their values for service and their waiting costs. There are n classes of jobs. Each job from class i has the same value v_i for job completion, and we assume $v_i > v_{i+1}$. The total arrival rate of potential jobs is $\lambda = \sum_i \lambda_i$. Each job is independently assigned class i with probability λ_i/λ, hence the total arrival rate of potential jobs from class i is 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) $F_i(c)$, where $f_i(c)$ is the corresponding probability density function (pdf) of $F_i(c)$. The class and exact waiting cost of a job is its private information; however, the class values v_i and probability distributions F_i are common knowledge. The random variable C_i's are independent of each other.

Jobs (agents) choose whether or not to enter the system, are *Individually Rational* and risk neutral with respect to payments and benefits and hence aim to maximize their expected payoff. If a job from class i with waiting cost c pays an amount 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 the full price to the job is the sum of the direct payment and indirect waiting costs, $cw + m$, and then hence its payoff is $v_i - cw - m$. Jobs compete for system resources by submitting a "bid" to the system, where the information contained in the bid will depend on the mechanism design, and is some function of the job's value, v_i and waiting cost c_i.

Under our Individually Rational assumption, each job competes to acquire an instance only if its expected payoff is nonnegative. Hence, $f_i(c)$ is assumed to be strictly positive[2] for $c \in [0, \mu v_i]$, (since jobs from class i with waiting cost greater than μv_i will always balk).

Two pricing and scheduling regimes are allowed, shown schematically in Figure 1 and which we now describe.

Modeling PAYG: We assume that the overall system has enough capacity to serve the exogenous demand λ/μ with negligible buffering or rejection of individual

[2] This is not a restrictive assumption in practice since we can approximate any distribution arbitrarily well with such a distribution. See, e.g., Figures 3 and 4 and associated discussion

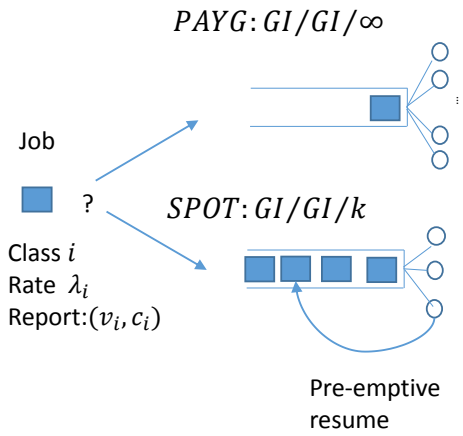

Fig. 1 System model for PAYG and spot market

jobs, and hence PAYG is modeled as a $GI/GI/\infty$ system with service rate μ (this is discussed further in Section 7.1). 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 common knowledge. The only information contained in a "bid" from a job is a request to enter. 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, where a bid can be thought of (but does not have to be) a single number that represents a willingness to pay. We will be 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 4.1 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 is unobservable to the arriving jobs. Jobs are not allowed to renege or change their bids. A job is charged based its own bid and the bids of others according to some spot pricing mechanism. Examples include the first price auction where at each time the jobs with k highest bids currently in the system 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 current $(k+1)^{th}$ highest bid per-unit time. 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 [18]. A consequence is that only the expected payment matters, and hence any payment implementation that achieves

the correct payments can be used. In practice, charging via a per-unit time price is a natural implementation.

3 Mechanism Design

We now recap certain fundamental concepts from mechanism design, which we make use of in our analysis, and which may be unfamiliar to the queueing systems community.

Mechanism design is a branch of game theory that describes rules governing the allowed interactions of agents. Given some allowed set of *bids* or reports from the buyers, which will depend on their underlying "true" reports or values, a mechanism prescribes both an allocation rule (who should get what) and a payment rule, how much they should pay, where the allocation and payment rules are functions of the submitted bids. A standard notion of a stable solution relevant to this setting is that of a Bayes Nash Equilibrium (henceforth BNE). In a BNE there is a prior distribution over the values of other agents and the strategy adopted by each agent is optimal in expectation given this prior, his value, and the strategies other agents of each type would adopt.

The *revelation principle*, states that if there is a BNE for some mechanism, then an equivalent direct *truthful* mechanism exists, one where it is an equilibrium for the the buyers to bid their true values, equivalent in the sense of outcomes (allocation and payments). This result, proved by Myerson [19], makes it sufficient to focus on truthful, direct mechanisms.

How to choose amongst competing (truthful direct) mechanisms? In a seminal work, Myerson [18] showed that for mechanisms that are truthful or *incentive compatible* in the BNE sense, then first, the allocation rule leads to one and only one payment rule, and secondly that the payoff formula bounds the payoff of any feasible mechanism. The first result is a *Revenue Equivalence* result: different mechanisms which lead to the same allocation rule provide the same revenue; the second allows the design of optimal mechanisms, where revenue is maximized. In this paper, we use these techniques to characterize the equilibrium outcome once a PAYG price has been fixed.

One notable truthful mechanism is the Vickrey-Clarke-Groves or VCG mechanism, which has the property that it is *efficient*, i.e. maximizes social welfare (see, for example [15] for a description of VCG). It also has the property that the payment rule for an agent depends on the bids of other agents and represents the externality that agent imposes by being part of the system. For example, if k identical goods are being sold, then the VCG mechanism is the Vickrey auction [20] that allocates the k items to the highest k bidders, and an agent wining pays the $k+1$st highest bid (i.e. the first losing bid). In this paper, we make use of this characterization of the VCG mechanism when analyzing the social welfare created by a hybrid market in Section 6.

4 PAYG and Spot Market Analysis

4.1 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 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). As a result, previous work has typically focused on analyzing a particular mechanism such as a first price auction. However, we show in this section that this complexity is dispensable. 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 a job participates in the spot market, but this takes the form of a simple cutoff: jobs with waiting costs below the cutoff participate and those above do not.

By the revelation principle for BNE [19], 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) and if it participates in the spot market 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 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 multiple 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') minimizing the expected total cost $c\widetilde{w}(v', c') + \widetilde{m}(v', c')$. Thus, when multiple classes participate, the set of optimal reports is class-independent; in particular, if classes i and j participate, both (v_i, c) and (v_j, c) belong to the set of optimal reports. Let $s_i(c)$ be the (randomized) equilibrium strategy for class i with cost c. Now, suppose instead that every job from a class that participates with positive probability with waiting cost c uses strategy $s_i(c)$ with probability $\lambda_i f_i(c)/(\sum_{j \in SPOT(c)} \lambda_j f_j(c))$ where $SPOT(c)$ is the set of all classes that participate in the spot market with positive probability with cost c. Then the arrival process for all strategies $s_i(c)$ remains identical to the original process. That is, the probability that the next job to arrive reports a pair (v, c) is unchanged. Hence the waiting time and the expected payment remain unchanged. This new class-independent randomized strategy is also an equilibrium for all classes. □

Since jobs can undo any tie-breaking a class-based mechanism does, we assume for the remainder of the paper that mechanisms have a class-independent expected waiting time $\widetilde{w}(c)$ and expected payment $\widetilde{m}(c)$. This allows us to concentrate on one-dimensional reports (bids) from the agents, related to their waiting cost. 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:

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

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

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.

Assumption 1 *We assume that jobs with no waiting cost are served for free in the spot market, hence $\widetilde{m}(0) = 0$.*

We revisit this assumption at the end of the paper.

Assumption 2 *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 assumption 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 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.

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 class 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 $\mathbf{c} = (c_1, .., c_n)$, we need to analyze the expected waiting time for a job with waiting cost c in the spot market with cutoffs $\mathbf{c}$. 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 $w(c; \mathbf{c})$ 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. We define (c_i', c_{-i}) to be the vector obtained from $\mathbf{c}$ by replacing the ith component by c_i'.

The following lemma gives the relevant properties of w. Intuitively, they capture that jobs with higher waiting costs get lower waiting times, if more jobs decide to enter the spot market then waiting times increase, and that entering jobs do not affect the waiting time of jobs with higher waiting costs.

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

(i) *$w(c;\mathbf{c})$ is decreasing in c for $c \in [0, \max_i c_i]$, $w(c;\mathbf{c}) > 1/\mu$ if $c < \max_i c_i$, and $w(c;\mathbf{c}) = 1/\mu$ if $c \geq \max_i c_i$.*
(ii) *$w(c;\mathbf{c})$ is increasing in c_i for $c_i \in [0, \mu v_i]$.*
(iii) *For any $t \geq \widehat{c}_j > c_j$ $w(t;\mathbf{c}) = w(t;(\widehat{c}_j, \mathbf{c}_{-j}))$.*

Proof The condition $(\sum_i \lambda_i F_i(c_i))/(k\mu) < 1$ ensures the queue is stable and hence expected waiting time is finite. Since priority is given to jobs with higher waiting cost, the expected waiting time of a job with waiting cost c increases with the total arrival rate of the jobs whose waiting cost is higher than c, which is equal to $\sum_i \lambda_i \left[F_i(c_i) - F_i(c)\right]^+$. The job with waiting cost greater than or equal to $\max_i c_i$ 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 [18] to 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 truthful reporting with cutoffs $\mathbf{c}$ constitutes a BNE for the given spot pricing mechanism. Let $m(c)$ be the expected payment made by a job with waiting cost c (which is independent of its class). For a BNE to exist, the following *incentive compatibility* (henceforth, IC) constraint must hold: for all $\widehat{c}, c \leq \max_i c_i$, and any i,

$$v_i - cw(c;\mathbf{c}) - m(c) \geq v_i - cw(\widehat{c};\mathbf{c}) - m(\widehat{c}). \tag{3}$$

By analogy with [18], 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.

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

$$m(c) = \int_0^c w(t;\mathbf{c})dt - cw(c;\mathbf{c}). \tag{4}$$

Hence, the expected payment by a job with waiting cost c is uniquely determined by the function w. Moreover, Lemma 4 and (4) *together satisfy the IC constraint* (3).

Proof (Sketch- following Myerson[18]) Let

$$\pi(\widehat{c}, c) \triangleq v_i - cw(\widehat{c};\mathbf{c}) - m(\widehat{c}).$$

Then for the IC constraint to hold, the maximum of $\max_{\widehat{c}} \pi(\widehat{c}, c)$ must be achieved, and is achieved at $\widehat{c} = c$. Since π as a function of c is affine, it follows that $\pi(c, c) = \max_{\widehat{c}} \pi(\widehat{c}, c)$ is convex, and hence is differentiable almost everywhere, with (right) derivative:

$$\frac{\partial}{\partial c}\pi(c, c) = -w(c; \mathbf{c}). \tag{5}$$

Integrating between 0 and c, substituting for π and rearranging give the result under our assumption that $m(0) = 0$ (the job with zero waiting cost won't pay anything in the spot market because waiting is costless for it). □

Since $w(c; \mathbf{c})$ is decreasing in c for $c \in [0, \max_i c_i]$, the proof of Lemma 2 can be used to establish a stronger monotonicity result for the expected payment m.

Lemma 6 *Given cutoffs* $\mathbf{c}$*, the expected payment* $m(c)$ *is increasing in* c *for* $c \in [0, \max_i c_i]$.

Proof Suppose that the waiting cost of a job in class i is c and it instead misreports some $\widehat{c} \neq c$. The expected payoff under truthful reporting is $\pi(c, c) \triangleq v_i - cw(c; \mathbf{c}) - m(c)$ and the expected payoff in case of misreport is $\pi(\widehat{c}, c) \triangleq v_i - cw(\widehat{c}; \mathbf{c}) - m(\widehat{c})$. Considering the cases $\widehat{c} < c$ and $\widehat{c} > c$ separately, using the property that $w(t; \mathbf{c})$ is decreasing in t, and using (4), we can show that $\pi(c, c) - \pi(\widehat{c}, c) > 0$. □

4.2 Revenue and equilibria for isolated markets

Next, we analyze PAYG and the spot market each in isolation.

4.2.1 PAYG

First consider PAYG in isolation. If the 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:

$$R^{PAYG}(p) \triangleq \frac{p}{\mu}\left(\sum_i \lambda_i F_i(\mu v_i - p)\right), \tag{6}$$

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

4.2.2 Spot Market in Isolation

For the spot market in isolation, denote the cutoffs in this case by $\mathbf{c}^S$. 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:

$$v_i - \int_0^c w(t; \mathbf{c}^S)dt \begin{cases} \geq 0 & \text{if } c < c_i^S, \\ = 0 & \text{if } c = c_i^S. \end{cases} \tag{7}$$

Theorem 1 below shows that there is an unique cutoff vector $\mathbf{c}^S$ satisfying (7) which characterizes the BNE for the spot market in isolation.

Theorem 1 *The following holds:*

(i) There is a unique solution $\mathbf{c}^S$ to the following system of equations in $\mathbf{x} = (x_1, \ldots, x_n)$:

$$\int_0^{x_i} w(t; \mathbf{x})dt = v_i. \tag{8}$$

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

The proof follows by straightforward inductive argument on the number of agent classes, and is given in Appendix A. 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,

$$m(c; \mathbf{c}^S) = \int_0^c w(t; \mathbf{c}^S)dt - cw(c; \mathbf{c}^S). \tag{9}$$

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:

$$R^s \triangleq \sum_i \lambda_i \int_0^{c_i^S} m(t; \mathbf{c}^S) f_i(t)dt. \tag{10}$$

4.3 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 4.1, for a given PAYG price p, we look for a cutoff vector $\mathbf{c}(p)$ 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 go 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 proceeds along the same lines as that of Theorem 1 and is given in Appendix B. We also derive additional cutoff parameters $\bar{\mathbf{c}}$ which we require for analyzing the hybrid market, and which are useful as starting points for determining the optimal $\mathbf{c}^S$.

Theorem 2 *The following holds:*

(i) For each $i \in \{1, 2, ..., n\}$ there is a unique vector of the form

$$\mathbf{x} = (x_i, x_i, \ldots, x_i, x_{i+1}, x_{i+2}, \ldots, x_n)$$

that satisfies $\int_0^{x_j} w(t; \mathbf{x})dt = v_j$ for all $j \geq i$. Let $\bar{c}_i$ be the value of x_i in this vector.

(ii) *Define* $v_{n+1} = \overline{c}_{n+1} = \overline{c}_0 = 0$ *and* $v_0 = \infty$. *Then there is a unique* $i \in \{0, 1, ..., n\}$, *denoted* i^* *that is the class such that* $p \in [\mu v_{i^*+1} - \overline{c}_{i^*+1}, \mu v_i - \overline{c}_{i^*})$.

(iii) *There is a unique solution* $\mathbf{c}(p)$ *to the system of equations that*

$$\int_0^{x_j} w(t; \mathbf{x}) dt = \begin{cases} \frac{p+x_j}{\mu} & \text{if } j \leq i^* \\ v_j & \text{otherwise.} \end{cases} \tag{11}$$

(iv) *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*[3]. *In particular, this is true exactly for the classes such that* $i > i^*$.

This theorem also provides insight into the structure of the outcome. For example, if follows from parts (ii), (iii) and (iv) that

Corollary 1 *All classes that participate in PAYG have the same cutoff,* $c_{i^*}(p)$, *and if the price is set higher than* $\mu v_1 - \overline{c}_1 = \mu v_1 - c_1^S$ *then no class participates in PAYG and the outcome is the same as if only the spot market existed.*

Our analysis 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).

The expected revenue to the cloud service provider per unit time is the sum of expected revenue from the spot market and PAYG. From (6), (10), and Theorem 2, given a PAYG price p, the expected revenue per unit time for the hybrid system, denoted by $R^H(p)$, is:

$$R^H(p) \triangleq \sum_i \lambda_i \left(\frac{p}{\mu} \left[F_i(\mu v_i - p) - F_i(c_i(p)) \right]^+ + \int_0^{c_i(p)} m(t; \mathbf{c}(p)) f_i(t) dt \right), \tag{12}$$

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

5 Revenue Comparisons

In the previous section, we characterized the equilibrium outcomes and resulting revenue for three different market types. Now we compare their performance. Since just having a spot market is a special case of the hybrid market, we focus on whether a cloud service provider should prefer PAYG or a hybrid market. Perhaps the simplest question we can ask is whether PAYG or hybrid raises more revenue. The next theorem shows that if the optimal price for the hybrid system is sufficiently small, 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* $p^H \leq \mu v_n - \overline{c}_n$, *i.e., all classes participate in PAYG. 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.,* $\max_p R^H(p) = R^H(p^H) < \max_p R^{PAYG}(p)$.

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

Proof Since $\max_p R^{PAYG}(p) \geq R^{PAYG}(p^H)$, it suffices to show that $R^{PAYG}(p^H) > R^H(p^H)$.

Since $p^H \leq \mu v_n - \overline{c}_n$, then for all i and j, $c_i(p^H) = c_j(p^H) \leq \overline{c}_n$, implying $\mu v_i - p^H \geq \overline{c}_n \geq c_i(p)$. Then from (6) and (12),

$$R^{PAYG}(p^H) - R^H(p^H) = \sum_i \lambda_i \left(\frac{p^H}{\mu} F_i(c_i(p^H)) - \int_0^{c_i(p^H)} m(t; \mathbf{c}(p^H)) f_i(t) dt \right). \tag{13}$$

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

$$c_i(p^H) w(c_i(p^H); \mathbf{c}(p^H)) + m(c_i(p^H); \mathbf{c}(p^H)) = \frac{c_i(p^H) + p^H}{\mu}. \tag{14}$$

Since $c_i(p^H) = c_j(p^H)$, $w(c_i(p^H); \mathbf{c}(p^H)) = 1/\mu$. 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 (13) imply:

$$R^{PAYG}(p^H) - R^H(p^H) > \sum_i \lambda_i \left(\frac{p^H}{\mu} F_i(c_i(p^H)) - \int_0^{c_i(p)} \frac{p^H}{\mu} f_i(t) dt \right) = 0. \tag{15}$$

□

The intuition behind Theorem 3 is that, in a class that participates in PAYG, all jobs of that class that instead choose the spot market would prefer PAYG to balking. Since they pay less money (but more waiting time) to use the spot market, we could make more money if we could prevent them from entering the spot market. When this is true of every class, we can actually prevent them, by simply eliminating the spot market.

Obviously Theorem 3 has significantly more bite with a small number of classes, since it requires that all participate in PAYG. However, we note that for Amazon only a small percentage of jobs are submitted to the spot market, so this may well be the relevant case. Further, we conjecture that the revenue ranking result holds much more broadly. As an example, we simulate the performance of a spot market that consists of k parallel $M/M/1$ queues with two classes, where jobs bid for preemptive priorities. 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. As shown in prior work [16], the waiting time is given by

$$w(c; c_1, c_2) = \frac{1}{\mu \big(1 - \sum_{i=1,2} \rho_i \left[F_i(c_i) - F_i(c)\right]^+\big)^2}, \tag{16}$$

where $\rho_i \triangleq \lambda_i/(k\mu)$. The proof of Theorem 2 provides a recipe for numerically computing the cutoff vector $\mathbf{c}(p)$ as a function of PAYG price p.

We randomly generated one hundred random configurations of the values of v_i's, λ_i's, and k. All were chosen uniformly at random from $[0, 20]$, with k a random integer from this range. The service rate μ was kept constant at one and F_i was uniform in the interval $[0, \mu v_i]$.

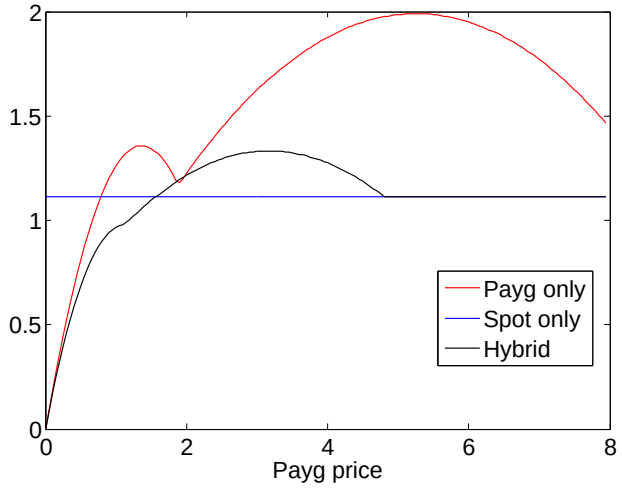

Fig. 2 Expected revenue $R^{PAYG}(p), R^H(p), R^S(p)$ for PAYG, Hybrid and Spot respectively as a function of PAYG price p for fixed $\lambda_1 \approx 0.76, \lambda_2 \approx 1.23, v_1 \approx 10.5, v_2 \approx 1.88$ with $k = 1$ and uniform waitings cost distributions.

For each realized configuration, we observed that the optimal revenue from PAYG in isolation was 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 $\mu v_2 - \bar{c}_2$. An example plot where $p^H > \mu v_2 - \bar{c}_2$ is shown in Figure 2[4]. 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 $p = \mu v_2$, all class 2 jobs balk from PAYG leading to a kink in the plot for PAYG in isolation. However, it is not the case that PAYG raises more revenue at all prices p, as near this kink the hybrid market raises slightly more revenue. Additional simulations with exponential and beta distributed waiting costs also failed to generate instances where the hybrid market generated more revenue. This suggest that for a broad range of parameters not operating a spot market is optimal.

We now explore a situation where intuitively a hybrid scheme should outperform PAYG, specifically where there are two types of jobs: high value jobs with high waiting time costs, and low value jobs having low waiting time costs. A hybrid mechanism where high value jobs use PAYG, and low value use the spot market might be expected to generate more revenue than setting a single PAYG price. We shall see that this is still often not the case, essentially for two reasons:

1. In any hybrid scheme, if the price is such that only the high class participates in PAYG ($i^* = 1$) some high value traffic will always use the Spot Market. This is because it follows from (11) in Theorem 2 that $c_1(p) > 0$, hence $F_1(c_1(p)) > 0$.

[4] The exact parameter values are $v_1 = 10.508088077186715$, $v_2 = 1.876400535497311$, $\lambda_1 = 0.7576345977040905$, $\lambda_2 = 1.2301997305619036$, and $k = 1$.

2. Under the same setting, not all the low value traffic can use the spot market: by the proof of Lemma 7 in the appendix $c_2(p)$ is decreasing in p in this range. Thus $c_2(p) < v_2$, and hence $F_2(c_2(p)) < 1$.

In other words we cannot achieve perfect separation of the two classes by price alone. We use Beta distributions to examine this example in more detail. Consider a setting where there is one type with high value ($v_1 = 10$) and high waiting costs while another with low value v_2 and low waiting costs. Figure 3 shows such an example where the waiting cost for the low value type is essentially all between 0 and 1 and the waiting costs for the high value type are essentially all slightly less than 5. The other parameters are $k = 10$ and $\lambda_1 = \lambda_2 = 5$. We now explore what happens as we vary v_2 between 3 and 4, settings where the optimal PAYG price excludes type 2 traffic, and hence Theorem 3 does not apply. Intuitively, we could set a PAYG price close to 5 to try and capture as much revenue from the high values as possible while getting some revenue from the low values. Figure 4 shows that when $v_2 = 3$ even in this setting PAYG alone is still optimal (though only very slightly). The essential problem is that the high PAYG price gives class 1 jobs an incentive to drop down into the spot market, wiping out some of the gains from serving the class 2 jobs. It does suggest that the hybrid system can be more robust in some cases if there is uncertainty about how best to set prices. However, this is an example created to make the hybrid system look as good as possible; in Figure 2, which is much more representative of the examples generated in our simulations, the hybrid system does not add such robustness.

If we increase v_2, putting $v_2 = 3.5$, then as shown in Figure 5, the hybrid market does indeed generate (slightly) more revenue that PAYG. But if v_2 is increased still further to $v_2 = 4$, then once again PAYG generates more revenue. For this parameterized model, the hybrid model is only better if approximately $3.1 \leq v_2 \leq 3.8$, whereas PAYG generates more revenue if approximately $v_2 \leq 3$ or $v_2 \geq 3.9$. Similarly, if we fix all parameters except λ_2, there is a small interval $\mathcal{I}$, when hybrid is optimal for $\lambda_2 \in \mathcal{I}$, and PAYG optimal for all other λ_2. Thus, even in an example designed to make the hybrid mechanism look as good as possible, there is a relatively small range of parameters where it is superior.

6 Welfare Analysis

The two systems also provide different social welfare (i.e. the value of served jobs minus their waiting costs). For classes that participate in both PAYG and spot, agents that send their job to the spot market incur a higher waiting cost, which reduces welfare. On the other hand, classes that do not participate in PAYG receive some service in a spot market, which increases welfare. If the service provider wishes to raise a given amount of revenue, Figure 2 shows that he would pick a different (lower) price, which increases welfare.

To help characterize social welfare, we now show that the outcome of the hybrid system is, in a sense, efficient. Economic efficiency is the property of maximizing social welfare, the total utility in the system. Since payments are just a transfer of utility from one agent (the job) to another (the owner of the system), they are irrelevant. It has previously been observed for restricted cases that spot market outcomes are efficient (e.g. [11]). We show that this is true in general, but only if we treat the PAYG price as a real cost. Hence we say that the outcome is

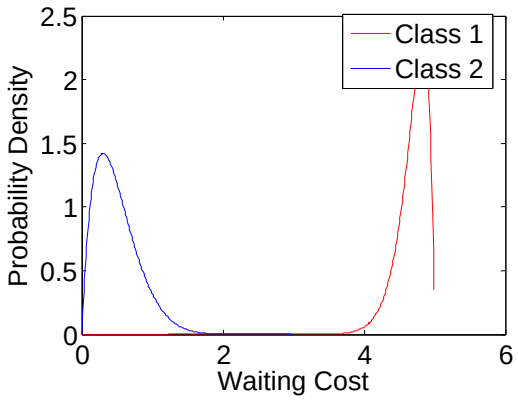

Fig. 3 Exanple pdfs $f_i(c)$ of well-separated waiting costs.

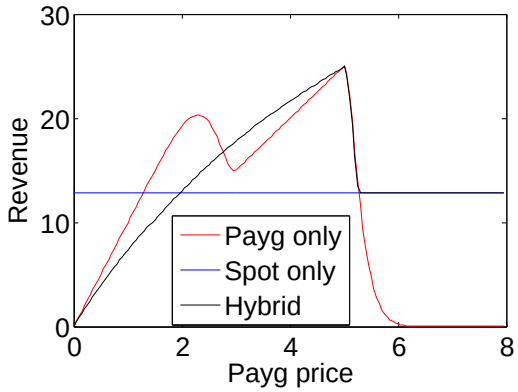

Fig. 4 Expected revenue $R^{PAYG}(p), R^H(p), R^S(p)$ for PAYG, Hybrid and Spot respectively as a function of PAYG price p for fixed $\lambda_1 = \lambda_2 = 5$, $k = 10$, $v_1 = 10$ and $v_2 = 3$. Here PAYG is (just) optimal. Waiting cost pdfs $f_i(c)$ as in Figure 3.

pseudo-efficient, because it would be efficient if the PAYG price represented a real cost.

Theorem 4 *The equilibrium of the hybrid system is pseudo-efficient: given the PAYG price p, agents make the socially optimal decision in determining whether to send their job to the spot market, send it to PAYG, or balk. Equivalently, this mechanism implements the VCG outcome. Thus, the payment of an agent is his (expected) externality*

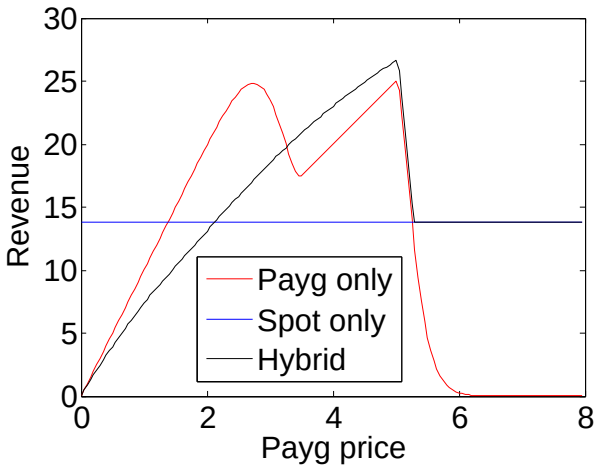

Fig. 5 Expected revenue $R^{PAYG}(p), R^H(p), R^S(p)$ for PAYG, Hybrid and Spot respectively as a function of PAYG price p for fixed $\lambda_1 = \lambda_2 = 5$, $k = 10$, $v_1 = 10$, as in Figure 4, but with $v_2 = 3.5$. Now the Hybrid scheme is optimal.

Proof Suppose we use the VCG mechanism, which selects the pseudo-efficient outcome and charges each agent his externality. In this outcome a given class of job must have all jobs with waiting costs below a cutoff sent to the spot market. Otherwise, jobs with a lower waiting cost that are not sent to it could be swapped with jobs with a higher waiting cost, increasing social welfare. Consider the marginal effect of admitting a job at the cutoff to the spot market for any class. Since the outcome is pseudo-efficient, the externality this causes plus its own waiting cost must be equal either to the cost it would experience under PAYG (if it would otherwise go there) or to its value (if it would otherwise balk). Since VCG is incentive compatible, we know that its payment (equal to the externality) satisfies (4). But then the equations that define the cutoffs of the spot market are exactly the same as those from (11), whose unique solution is the outcome of the hybrid system. □

Given a PAYG price p, the expected social welfare per unit time for the hybrid system, denoted by $W^H(p)$, is:

$$W^H(p) \triangleq \sum_i \lambda_i \left(\int_{c_i(p)}^{\mu v_i - p} \left(v_i - \frac{t}{\mu} \right) f_i(t) dt + \int_0^{c_i(p)} \left(v_i - w(t; \mathbf{c}(p)) \right) f_i(t) dt \right). \quad (17)$$

Figure 6 illustrates the effect of prices on social welfare in the same example from Figure 2. For much of the range of prices, the welfares of PAYG and the hybrid sytem are similar, with the better solution changing several times. Once prices are high enough that no one participates in PAYG in the hybrid system, social welfare in the hybrid system is higher; because PAYG tends to extract more revenue at a given price, Figure 7 shows that, at least in this example, PAYG enables a better tradeoff interesting most operating points. In particular, while there are choices of social welfare for which the hybrid market raises more

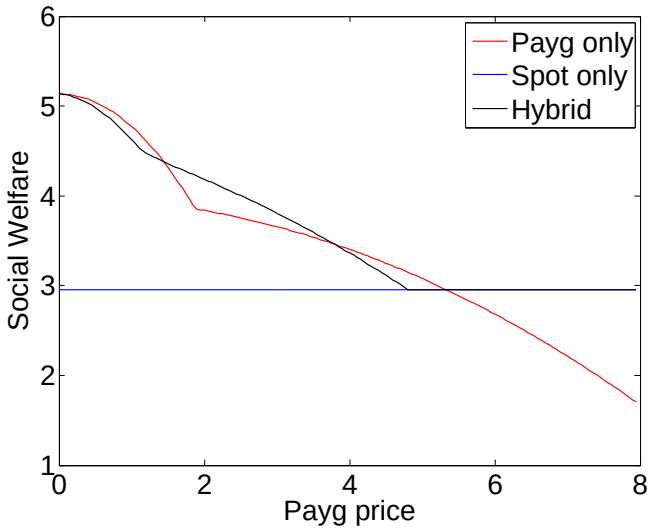

Fig. 6 Social welfare $W^{PAYG}(p), W^{S}(p), W^{H}(p)$ as a function of PAYG price for Figure 2 parameter settings .

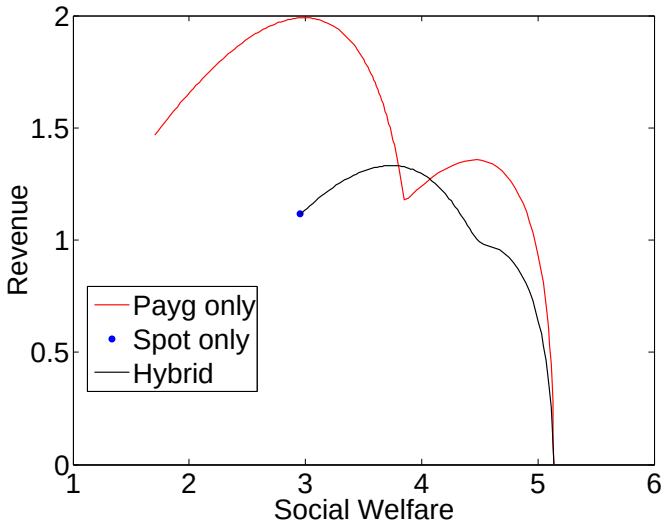

Fig. 7 Tradeoff between revenue $R(p)$ and welfare $W(p)$ for Figure 2 parameter settings.

revenue, for every such choice there is a choice with higher social welfare where the hybrid market simultaneously achieves more revenue (i.e. a Pareto improvement in economic terminology). So in this example, PAYG not only raises more revenue but enables better tradeoffs between revenue and social welfare. Gaining a theoretical understanding of these tradeoffs is an important question for future work.

7 Relaxing Model Assumptions

We now consider extensions to our model, which relax some of our assumptions.

7.1 Finite Capacity PAYG

We have modeled PAYG as infinite capacity system, whereas the spot market is has a finite capacity (k) servers. This dichotomy is intended to represent current cloud-computing markets, where IaaS is primarily sold using a PAYG model with availability guarantees, whereas the spot market is much smaller. However, we believe that with respect to our results, the assumption that PAYG is modeled as a $GI/GI/\infty$ queue, rather than (say) a $GI/GI/n/n$ queue is essentially without loss of generality.

If PAYG has a finite capacity, then in order to give service level guarantees, there is some arrival rate λ^* and associated load ρ^* such that the performance of PAYG meets service level guarantees for all $\lambda \leq \lambda^*$. For example, by choosing λ^* such that $\Pr\{\text{arriving jobs sees all } n \text{ servers busy}|\lambda = \lambda^*\} \leq \epsilon$, for some small $\epsilon \approx 0$. Provided $\lambda \leq \lambda^*$, the behavior of the PAYG queue will approximated by $GI/GI/\infty$ queue. For $\lambda > \lambda^*$, with a finite capacity queue, certain jobs would not be admitted (c.f [4] for some information on Amazon EC2 PAYG policies). Thus, we could incorporate this in our model by assuming that there is a small probability that jobs submitted to PAYG are not served and pay cost 0. This slightly reduces revenue from PAYG and the PAYG part of hybrid and results in classes choosing slightly higher cutoffs in Theorem 2, but does not materially affect our results.

In more detail, let $B(\rho, n)$ denote the probability that an arriving job sees all servers busy in an $GI/GI/n/n$ queue, which serves PAYG customers. Define $\rho^{PAYG}(p) = \frac{1}{\mu}\sum_i \lambda_i F_i(\mu v_i - p)$, and let $\epsilon = B(\rho^{PAYG}(p), n)$. Then Theorem 2 holds as stated, but with the BNE in part (iv) replaced by ϵ^*-BNE, where ϵ^* is upper bounded by $v_i(1 - p/\mu) \times \epsilon$. Compared to $n = \infty$, $R^{PAYG}(p)$ decreases by $p\rho^{PAYG}(p) \times B(\rho^{PAYG}(p), n)$, while $R^H(p)$ decreases by $p\rho^H(p) \times B(\rho^H(p), n)$, where $\rho^S(p) \triangleq \frac{1}{\mu}\sum_i \lambda_i F_i(c_i(p))$, $\rho^H(p) = \rho^{PAYG}(p) - \rho^S(p)$; both decrements are $O(\epsilon)$ under our assumptions that $\rho^s \ll \rho^{PAYG}$.

Alternatively Theorem 2 can be amended to allow BNE rather than ϵ^*-BNE on replacing (11) with

$$\int_0^{x_j} w(t;\mathbf{x})dt = \begin{cases} \frac{p+x_j}{\mu}(1 - B(\rho^H(p), N) + v_j B(\rho^H(p), N) & \text{if } j \geq i \\ v_j & \text{otherwise,} \end{cases}$$

and using this to define the unique $\mathbf{c}(p)$ and $\bar{c}_{i^*}$ via (ii). The finite capacity results in an $O(\epsilon)$ increase in the cut-offs. Notice that now there is weak $O(\epsilon)$ dependence of the cut-offs on the class, and hence payments also will have this weak class dependence. However, these do not affect our results that in general PAYG raises more revenue.

7.2 Reserve Prices

In our basic model, we assumed that $m(0) = 0$. This means that there is no minimum payment (reserve price), so jobs willing to tolerate a long enough wait can be served for free. Imposing such a reserve price is typically a way to raise revenue, so intuitively adding them to the spot market seems beneficial and potentially a way to make the hybrid market more profitable than PAYG alone. Indeed, it is at least weakly better because a PAYG is the special case of a hybrid market with a reserve price equal to the PAYG price. In this section, we explain why this extra option need not actually help.

First, we explain how our characterizations from Theorems 1 and 2 change in the presence of reserve prices. If we institute a reserve price of $m(0) = r$, then the incentive compatibility constraint given by Lemma 5 becomes

$$m(c) = r + \int_0^c w(t; \mathbf{c})dt - cw(c; \mathbf{c}).$$

With this change, we need to change the systems of equations based on indifference between the spot market and PAYG or balking. Specifically, (8) becomes

$$\int_0^{x_i} w(t; \mathbf{x})dt = v_i - r,$$

while (11) becomes

$$\int_0^{x_j} w(t; \mathbf{x})dt = \begin{cases} \frac{p+x_j}{\mu} - r & \text{if } j \geq i \\ v_j - r & \text{otherwise.} \end{cases}$$

The same proofs go through, *mutatis mutandis*.

This new characterization, shows that adding a reserve price doesn't change anything significant about the behavior or jobs. The spot market is now expensive, so fewer jobs use it with the rest either choosing PAYG or balking, but this structure is still determined by cutoffs as before. Thus it is still the case that, for every class with jobs that participate in both PAYG and the spot market, all those jobs would have been willing to pay the higher PAYG price. This is the main insight behind Theorem 3, so the proof of that theorem goes through as well. Thus, if the optimal combination of PAYG price and reserve price for a hybrid system is such that all classes of jobs participate in PAYG, then the reserve price is equal to the PAYG price and the system is effectively PAYG alone.

Furthermore, even in ranges of parameters where a hybrid market is superior, this analysis suggests there is an option that is better than either: offer only PAYG, but with a menu of prices where lower prices receive a larger (artificial) delay before (an approach explored by Afèche [1]). By way of intuition if the menu of prices and delays are chosen to be the payments and expected waiting time at the various cutoff values for different classes in the spot market, all jobs of a class that were making less than their cutoff payment would still be willing to pay the higher amount rather than balk. Indeed, in the example of Figure 5 where a hybrid market can raise revenue from 25.0 to 26.6, such a scheme can generate revenues upwards of 33. Such artificial delays are not without precedent: among other restrictions the Amazon Glacier cloud storage system may delay responding

to requests for up to 3 to 5 hours. Further, subsequent to our work Google launched "Preemptible VMs" which take exactly this approach of offering a fixed discount for the risk of preemption.

This analysis is in stark contrast to that of Doroudi et al.[8], who found that a spot market with a reserve price is optimal in their setting (and in particular superior to PAYG). However, their result relies on PAYG having the same resource constraints as the spot market does, a feature we have argued is not present in current cloud systems.

These results are under our Incentive Compatible allocation assumption, which implies that jobs are able to discover enough information from the system for it to be optimal to bid truthfully. This implies that any reserve price must be public knowledge or discoverable. Ben-Yehuda et al. [4] argue that Amazon EC2 spot prices appears to use a dynamic or random reserve price. Equivalently, this result could be due to having a limited supply of spot instances that is variable over time. Such opacity can cause agents (jobs) to be unsure what is their best report, thus making the mechanism (and pricing) not incentive compatible.

7.3 Spot Market Costs

Another variant of our model would be to assume that participation in the spot market is costly. In practice, designing systems to backup and resume work from checkpoints may require additional effort, so being preempted may not actually be costless the way our basic model assumes. If all classes of jobs must pay the same cost to participate in the spot market, then this essentially serves as a reserve price, except that it simply represents a loss of efficiency rather than an increase in revenue. If the cost is not the same for all classes, the situation is more interesting. In particular, if the cost is higher for classes with higher values, this opens up an opportunity to discriminate between classes that can make a hybrid market clearly profitable. Such "damaged goods" approaches to market segmentation are common across a variety of markets [17].

Suppose there are only two classes with $v_1 > v_2$ and class 1 is required to pay a cost of s_1 to participate in the spot market. Our results are easily adapted to the general case of indirect spot-market participation costs: for example, in (7) and in Theorem 1 v_i is simply replaced by $v_i - s_i$. However, now it is not necessarily the case that $c_1(p) > c_2(p)$. To see this, take the extreme case of $s_1 = v_1$, in which case jobs of class 1 will not participate in the spot market while jobs of class 2 will still do so. Thus, a high PAYG price could be set to optimally extract revenue from jobs of class 1, and even if this price is higher than μv_2 some revenue would still be extracted from the jobs of class 2 in the spot market with no jobs of class 1 defecting to it.

As related differentiation occurs when the support of the waiting cost is $[s_i, \mu v_i]$ rather than $[0, \mu v_i]$. Our analysis goes through using obvious alterations to the lower limits of integrals (equivalently, putting $f_i(t) = 0$ for $t \in [s_i, 0]$). The quantitative behavior is similar to when s_i is an indirect cost. For example, in the two-class example, if $s_1 = v_i$ and $s_2 = 0$ then class 1 jobs will not participate in the spot market.

8 Discussion and Future Work

Our analysis 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. In particular, we have proved that this always holds true when all classes participate in the PAYG market. It will also hold true in the degenerate case, where the PAYG is chosen suboptimally high, so that all cases prefer to enter the spot market. Simulations suggest that this may be true in general, regardless of whether individual classes of enter the PAYG market.

Our results contrast with previous work, and may also appear counterintuitive to those expecting price discrimination to automatically yield higher revenues. One significant difference between our work and previous analyses is the combination of two different markets operating simultaneously and under different mechanisms, one an incentive-compatible mechanism and the other a fixed price design requiring no information from the users. We specifically chose such a system as an abstraction of current pricing systems. If there is but a single mechanism for all users, that is a single scheduling and pricing design, then optimal dynamic or auction-based pricing can raise move revenue than fixed pricing, a result that has been proved under a variety of assumptions. The introduction of a secondary market changes the picture by changing user incentives. With the combined system, it is very difficult to extract extra revenue by using a spot market, because of the difficulty of avoiding cannibalizing the primary market, where low waiting cost, high value customers choose to drop to the (cheaper) spot market, thus decreasing revenue. Our results frame this more precisely.

Our analysis is based on a number of assumptions. However, as we have shown, many of them are not critical for our conclusions to hold. For example, we model the PAYG system as having infinite capacity, which we believe is reasonable given that capacity is endogenous and PAYG jobs are more profitable than spot market jobs. However, this can be relaxed as long as the capacity is "large." Similarly, our analysis is robust to the ability to set reserve prices in the spot market.

One assumption that does affect our findings is the assumption that the only indirect costs to jobs are waiting time, and hence that other indirect costs are zero, costs such as those associated with preemption or rewriting applications to enable them to cope with possible preemption. If these additional costs, s_i, differ among classes and are non-zero, it need no longer be the case that the cut-offs c_i decrease with (increasing) i. Indeed, s_i may be sufficiently large to offset the adverse selection problem, causing instead high value jobs to stay with PAYG so that the hybrid market extracts more revenue. For example, having both v_i and s_i decreasing with i but differences $v_i - s_i$ increasing implies high value jobs favour PAYG while low value favour a spot market.

Lastly, there are three assumptions which merit further study: first, the current analysis assumes 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. An example of this would be if more valuable jobs tend to arrive at certain times of day. Then it is possible that there are equilibria where jobs of different classes but the same cost have different outcomes. Though as both classes have the same set of optimal outcomes, this requires an amount of coordination

on tiebreaking that may be unreasonable in practice. Secondly, our framework is for a monopolistic provider. The effect of competitive pressures needs to be investigated. Third, we have assumed that waiting costs are linear. This does make the equilibrium analysis dramatically easier because only the expected waiting time matters rather than the full distribution, but does mean that the waiting cost can exceed the value of the job, resulting in negative utility.

We conclude by discussing the important point that Amazon does in fact operate a spot market, despite our results suggesting that it may not be optimal from a revenue perspective to do so. One possibility is that their arrival distribution happens to be in one of the ranges where this is in fact optimal or some assumption our model makes is not applicable in their setting. However, another possibility is that it is being used for reasons other than revenue optimality. For example, even if Amazon is making less money, they may be gaining useful information about what jobs can easily be interrupted if the system experiences an unexpected spike in demand or large-scale failure. Alternatively, the lack of revenue optimality may be exactly the point if the spot market is viewed as a way to gain new customers by offering them lower prices while they are still operating at a smaller scaler. This is consistent with the observation that, anecdotally, Amazon makes it difficult to operate in the spot market at a large scale. Finally, based on the analysis of Ben-Yehuda et al. [4] which found Amazon controls reserve prices and causes them to spike, Amazon may actually be using something closer to the menu pricing approach we discuss in Section 7.2. Perhaps tellingly in this regard, when Amazon introduced Glacier as a less expensive storage service, they adopted artificial delays rather than a sport market for data access. Thus, we do not view Amazon's operation of a spot market as necessarily contradicting our model or results.

A Proof of Theorem 1

We begin with a technical lemma that will be used several times in the appendix.

Lemma 7 *Let $(x_1, \ldots, x_k$ and $g_i(x_i), \ldots, g_n(x_n)$ be given such that $g_j(x_j)$ is weakly increasing and semidifferentiable[5] with left derivative at most $1/\mu$. Then there exists unique $x_i, \ldots, x_n$ such that*

$$\mathbf{x} = (x_1, \ldots, x_k, x_i, \ldots, x_i, x_{i+1}, \ldots, x_n)$$

satisfies

$$\int_0^{x_i} w(t; \mathbf{x}) dt = g_i(x_i) \tag{18}$$

for all $j \geq i$.

Proof Suppose this holds for $i+1$ to prove that it holds for i. By our induction hypothesis, given any $z \in [0, \mu v_i]$, there is a unique

$$\mathbf{x}(z) = (x_1, \ldots, x_k, z, \ldots, z, x_{i+1}, \ldots, x_n)$$

satisfying $\int_0^{x_j} w(t; \mathbf{x}(z)) dt = g_j(x_j)$ for all $j \geq i+1$. Next we show that $\phi(z) \triangleq \int_0^z w(t; \mathbf{x}(z)) dt$ is strictly increasing in z for $z \in [0, \mu v_i]$. Since $\int_0^{x_{i+2}} w(t; \mathbf{x}(z)) dt$ is increasing in each x_j,

[5] A function is semidifferentiable if it has left and right derivatives but they need not be equal. Note that this implies continuity. An example relevant for Theorem 2 is a continuous piecewise linear function.

$\mathbf{x}_{i+2}(z)$ is decreasing in z. Consider $\widehat{z} > z$. Then, $[\mathbf{x}_{i+2}(z), z] \subset [\mathbf{x}_{i+2}(\widehat{z}), \widehat{z}]$, and

$$\begin{aligned}
&\int_0^{\widehat{z}} w(t; \mathbf{x}(\widehat{z}))dt - \int_0^{z} w(t; \mathbf{x}(z))dt \\
&= \int_{\mathbf{x}_{i+2}(\widehat{z})}^{\widehat{z}} w(t; \mathbf{x}(\widehat{z}))dt - \int_{\mathbf{x}_{i+2}(z)}^{z} w(t; \mathbf{x}(z))dt \\
&= \int_{t \in [\mathbf{x}_{i+2}(\widehat{z}), \widehat{z}] \setminus [\mathbf{x}_{i+2}(z), z]} w(t; \mathbf{x}(\widehat{z}))dt \\
&\quad + \int_{\mathbf{x}_{i+2}(z)}^{z} \left(w(t; \mathbf{x}(\widehat{z})) - w(t; \mathbf{x}(z))\right) dt \\
&> \int_{\mathbf{x}_{i+2}(z)}^{z} \left(w(t; \mathbf{x}(\widehat{z})) - w(t; \mathbf{x}(z))\right) dt > 0,
\end{aligned}$$

where the last inequality follows from Lemma 4. Hence, $\phi(z)$ is increasing.

Since $\phi(0) = 0 \leq g_i(0)$, $\phi(\mu v_i) \geq v_i \geq g_i(\mu v_i)$, $\phi(z)$ is differentiable with $\phi'(z) > 1/\mu$, and g_i is weakly increasing and semidifferentiable with left derivative at most $1/mu$, there is a unique z solving $\phi(z) = v_{i+1}$. This establishes the claim.

Applying the lemma immediately gives the first part of the theorem. We complete the proof by establish the second part of the theorem.

Given $\mathbf{c}^S$, the expected payoff of a job from class i with waiting cost $c \leq c_i^S$ from participating in the spot market is $v_i - \int_0^c w(t; \mathbf{c}^S)dt$. Since $\mathbf{c}^S$ satisfies (7), the expected payoff is nonnegative. The pricing rule (4) ensures incentive compatibility for $c \leq c_i^S$. We only need to show that if $c > c_i^S$, the job does not participates in the spot market. Since $w(t; \mathbf{c}^S) = 1/\mu$ for $t \geq c_1^S$, reporting a waiting cost larger than c_1^S does not improve the waiting time of the job, and the expected payment is at least $m(c_1^S)$. Hence, if a job with waiting cost $c > c_i^S$ decides to participate in the spot market, it will (mis)report a waiting cost $\widehat{c} \in [0, c_1^S]$. The expected payoff of the job is $v_i - cw(\widehat{c}; \mathbf{c}^S) - m(\widehat{c})$. Then,

$$\begin{aligned}
&v_i - cw(\widehat{c}; \mathbf{c}^S) - m(\widehat{c}) \\
&= v_i - c_i^S w(\widehat{c}; \mathbf{c}^S) - m(\widehat{c}) - (c - c_i^S) w(\widehat{c}; \mathbf{c}^S), \\
&\leq v_i - c_i^S w(c_i^S; \mathbf{c}^S) - m(c_i^S) - (c - c_i^S) w(\widehat{c}; \mathbf{c}^S), \\
&= -(c - c_i^S) w(\widehat{c}; \mathbf{c}^S < 0.
\end{aligned}$$

The first inequality is from the IC constraint (3) and then last equality is because the cutoffs $\mathbf{c}^S$ are the solutions of (18). Hence the expected payoff of a job with waiting cost $c > c_i$ from participating in the spot market is negative and it will not participate.

This completes the proof of Theorem 1. ■

B Proof of Theorem 2

Step 1: Existence of the unique solution $\overline{c}$

The existence and uniqueness of $\overline{c}_i$ follows from Lemma 7.

Step 2: Existence of unique i

In equilibrium, jobs of class j whose waiting cost is $c_j(p)$ are indifferent between participating in the spot market and some outside option, either PAYG or balking. That is,

$$v_j - \int_0^{c_j(p)} w(t; \mathbf{c}(p))dt = \max\left(0, v_j - \frac{p + c_j(p)}{\mu}\right),$$

or

$$\int_0^{c_j(p)} w(t; \mathbf{c}(p))dt = \min\left(v_j, \frac{p + c_j(p)}{\mu}\right).$$

The existence and uniqueness of such a $\mathbf{c}(p)$ follows from Lemma 7. Given this solution, there is either a unique c^* such that

$$\int_0^{c^*} w(t;\mathbf{c}(p))dt = \frac{p+c^*}{\mu},$$

or

$$\int_0^{\mu v_1} w(t;\mathbf{c}(p))dt < \frac{p+\mu v_1}{\mu}.$$

In the latter case, no job participates in PAYG so we are done. In the former, there is a unique i such that

$$v_{i+1} \leq \frac{p+c^*}{\mu} < v_i,$$

or

$$p \in [\mu v_{i+1} - c^*, \mu v_i - c^*).$$

By construction, $c^* \geq \overline{c}_j$ for $j > i$ and $c^* \leq \overline{c}_j$ for $j \leq i$. Thus there is a unique such i as desired.

Step 3: Existence of the solution to the equations governing the choice of $\mathbf{c}(p)$.

Our argument in the previous step shows that the $\mathbf{c}(p)$ we construct is a solution to (11). By Lemma 7 it is the unique solution.

Step 4: Characterizing equilibrium.

To be an equilibrium, the cutoff vector $\mathbf{c}(p)$ must satisfy the following constraints for all $c < c_j(p)$:

$$\begin{array}{l} v_j - \int_0^c w(t;\mathbf{c}(p))dt \geq 0 \text{ and} \\ v_j - \int_0^c w(t;\mathbf{c}(p))dt > v_j - \frac{p+c}{\mu}. \end{array} \tag{19}$$

$$\begin{array}{l} \text{Hence either, } \begin{cases} v_j - \int_0^{c_j(p)} w(t;\mathbf{c}(p))dt = 0, \\ v_j - \frac{p+c}{\mu} < 0 \text{ for } c \in [c_j(p), \mu v_j - p], \end{cases} \\ \text{or, } \begin{cases} v_j - \int_0^{c_j(p)} w(t;\mathbf{c}(p))dt = v_j - \frac{p+c_j(p)}{\mu}, \\ v_j - \frac{p+c}{\mu} \geq v_j - \int_0^{c_i(p)} w(t;\mathbf{c}(p))dt \text{ and} \\ v_j - \frac{p+c}{\mu} \geq 0 \text{ for } c \in [c_j(p), \mu v_j - p]. \end{cases} \end{array} \tag{20}$$

Constraint (19) says that the jobs with waiting cost below the cutoff get nonnegative expected payoff from participating in the spot market. Moreover, this expected payoff is strictly higher than that from participating in PAYG. Constraint (20) says that either no jobs from a class i participate in PAYG, or jobs split between the spot market and PAYG with those above the cutoff weakly preferring PAYG.

First, consider classes $j \geq i$. For these classes, $\int_0^{c_i(p)} w(t;\mathbf{c}(p))dt = v_i$, so by construction the first inequality of (19) is satisfied. For the second, $\frac{p+c}{\mu} - \int_0^c w(t;\mathbf{c}(p))dt$ is decreasing in c for $c < c_1(p)$, where it reaches 0. Therefore the second inequality is satisfied. For (20), these classes satisfy the first system of constraints. By construction the equality is satisfied, while the inequality is satisfied because $\frac{p+c}{\mu} - \int_0^c w(t;\mathbf{c}(p))dt > 0$.

Now consider classes $j < i$. These classes share the same cutoff c^* satisfying $\int_0^{c^*} w(t;\mathbf{c}(p))dt = \frac{p+c^*}{\mu}$. For these classes, $v_j > \frac{p+c_j(p)}{\mu} > 0$, so (19) is satisfied. For (20), these classes satisfy the second system of constraints. By construction the equality is satisfied, while the first inequality is satisfied because $\frac{p+c}{\mu} - \int_0^c w(t;\mathbf{c}(p))dt$ is decreasing in c for $c < c_j(p)$, where it reaches 0. The second inequality trivially holds for all $c \leq \mu v_i - p$. (This point is part of the intuition for our results about revenue).

This gives the desired equilibrium characterization. A job from class $j < i$ with waiting cost c participates in the spot market if and only if $c \leq c_j(p)$; it participates in PAYG if $c_j(p) \leq c \leq \mu v_j - p$. A job from class $j \geq i$ with waiting cost c participates in the spot market if and only if $c \leq c_j(p)$; it never participates in PAYG.

References

1. Afèche, P.: Incentive-compatible revenue management in queueing systems: optimal strategic delay. Manufacturing & Service Operations Management **15**(3), 423–443 (2013)
2. Afèche, P., Mendelson, H.: Pricing and priority auctions in queueing systems with a generalized delay cost structure. Management Science **50**(7), pp. 869–882 (2004).
3. Afèche, P., Pavlin, M.: Optimal price-lead time menus for queues with customer choice: Priorities, pooling & strategic delay. Rotman School of Management Working Paper (2318157) (2013)
4. Agmon Ben-Yehuda, O., Ben-Yehuda, M., Schuster, A., Tsafrir, D.: Deconstructing amazon EC2 spot instance pricing. ACM Trans. Econ. Comput. **1**(3), 16:1–16:20 (2013).
5. Anselmi, J., Ardagna, D., Lui, J., Wierman, A., Xu, Y., Yang, Z.: The economics of the cloud: price competition and congestion. ACM SIGecom Exchanges **13**(1), 58–63 (2014)
6. Balachandran, K.R.: Purchasing priorities in queues. Management Science **18**(5), 319–326 (1972).
7. Cui, T., Chen, Y.J., Shen, Z.J.M.: Optimal pricing, scheduling, and admission control for queueing systems under information asymmetry (2009). Working Paper
8. Doroudi, S., Akan, M., Harchol-Balter, M., Karp, J., Borgs, C., Chayes, J.T.: Priority pricing in queues with a continuous distribution of customer valuations. Tech. Rep. CMU-CS-13-109, CMU (2013)
9. Dube, P., Jain, R.: Queueing game models for differentiated services. In: Game Theory for Networks, 2009. GameNets '09. International Conference on, pp. 523 –532 (2009).
10. Harchol-Balter, M., Osogami, T., Scheller-Wolf, A., Wierman, A.: Multi-server queueing systems with multiple priority classes. Queueing Systems **51**, 331–360 (2005).
11. Hassin, R.: Decentralized regulation of a queue. Management Science **41**(1), 163–173 (1995).
12. Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, 1 edn. Springer (2002)
13. van der Heijden, M., van Harten, A., Sleptchenko, A.: Approximations for markovian multi-class queues with preemptive priorities. Operations Research Letters **32**(3), 273 – 282 (2004).
14. Katta, A.K., Sethuraman, J.: Pricing strategies and service differentiation in queues a profit maximization perspective (2005). Working paper.
15. Krishna, V.: Auction Theory. Elsevier (2010)
16. Lui, F.T.: An equilibrium queuing model of bribery. Journal of Political Economy **93**(4), 760–781 (1985).
17. McAfee, P.: Pricing damaged goods. Economics: The Open-Access, Open-Assessment E-Journal **1**(1), 409–429 (2007)
18. Myerson, R.: Optimal auction design. Mathematics of Operations Research **6**(1), 58–73 (1981).
19. Myerson, R.B.: Incentive compatibility and the bargaining problem. Econometrica **47**(1), 61–73 (1979)
20. Vickrey, W.: Counterspeculation, auctions, and competitive sealed tenders. The Journal of Finance **16**(1), 8–37 (1961).
21. Xu, H., Li, B.: A study of pricing for cloud resources. ACM SIGMETRICS Performance Evaluation Review (2013). To appear in special issue on cloud computing
22. Yahalom, T., Harrison, J.M., Kumar, S.: Designing and pricing incentive compatible grades of service in queueing systems (2006). Working paper