Cloud Pricing: The Spot Market Strikes Back

LUDWIG DIERKS, University of Zurich
SVEN SEUKEN, University of Zurich

Cloud computing providers must constantly run many idle compute instances to ensure compliance with service level agreements towards their users with long-term contracts. A natural idea to increase the cloud provider’s profit is to sell some of these idle instances on a spot market, even though they often already offer a short term fixed-price market. It is therefore not clear if offering such a spot market is in the providers interest or if market cannibalization would occur. In this paper, we show that in equilibrium, offering a market consisting of some regular fixed-price instances as well as some spot instances (using idle resources), is always increasing the provider’s profit over only offering fixed-price instances. At the same time it also constitutes a Pareto improvement for the users. We show that this holds even if preemptions are costly for users. Finally, we illustrate the results numerically for a specific setting that demonstrates the large effect that market price and the provider’s instance cost have on the users’ equilibrium behavior and the provider’s profit.

1 INTRODUCTION

Providers of cloud services like Amazon EC2 or Microsoft Azure rent out computing capacity to company internal users, as well as to external customers. While some of this is done via year-long contracts, a large part consists of instances (virtual machines) that get rented on-demand for computation. As cloud markets already contain multi billion dollars in revenue and are still growing at an astounding rate, even a relatively small percentage of revenue or profit translates to hundredth of millions of dollars. Since these markets may contain many company internal users and are at the same time highly competitive for external users, providers cannot simply optimize theoretical profits without taking customer satisfaction into account. Any possible strategy for increasing profits therefore needs to make sure that no user is notably worse off than currently or take into account that they might migrate to competitors.

In current markets, instances are most commonly rented via fixed-price offerings where users pay a fixed price per time unit and the provider aims to offer enough instances to be able to almost instantly satisfy all requests. Simple and reliable, this satisfies the requirements of most users and seems to be the ideal baseline for any provider. Yet at the same time providers must constantly run many idle instances, for example to guarantee service level agreements of long-term contracts, for maintenance, as fail safe redundancy, or simply as a buffer for future growth ([Yan et al., 2016]). Effectively, a sizeable number of instances at any time does nothing. The resulting low utilization is undesirable, because much of the overall per-instance costs are paid upfront and are independent of usage. As these instances might be required for their dedicated purpose on a moment’s notice, they cannot be sold on the normal fixed-price market. This raises the question whether it would be in the provider’s interest to sell idle instances on a secondary market where any user’s job can simply be preempted. Getting preempted means that the job is shut down in favour of workload with higher priority until a new instance becomes available. As just shutting down a job is usually not without loss, preemption also carries a cost for the user beyond the time he waits until a new instance is ready. Pricing such a secondary market dynamically as a preemptible spot market seems

2For example according to Amazon (Retrieved 2.2.2017: https://s3-eu-west-1.amazonaws.com/donovapublic/TCOOutput.pdf), the cost of a server rack greatly depends on fixed costs (hardware/software/space) which remain even if the instance only runs during peak times.

Manuscript submitted for review to ACM Economics & Computation 2018 (EC ’18).
natural, as users already have to cope with preemption by the nature of the offered instances. In a preemptible spot market, users bid for resources and are served whenever the price is below their bid. If the market price rises above their bid while their job is running, they are preempted until the price drops again.

It is not directly clear if this would increase or decrease overall profits. While the hardware for such a spot market might come for free in the sense that usage independent costs (hardware, space, infrastructure etc.) are already incurred whether the spot market is offered or not, it might cannibalize the existing fixed-price markets. Indeed, Abhishek et al. [2012] show that this risk is quite real: offering a spot market will often decrease revenue over only offering instances at a fixed-price. They modelled the two different markets as distinct queues that arriving users could choose from and analyzed the resulting equilibria, an approach well studied for classical service systems ([Hassin and Haviv, 2003], [Hassin, 2016]). Yet, as their model does not incorporate any costs for the provider and assumed an infinite supply of instances, they did not make any statements about profit. Offering a spot market could therefore still be in the providers interest if it reduces the number of instances required in the data center.

In this paper, we present a related model that takes the cost of instances into account and is suited to also analyze profits. We assume a fixed-price market is offered and an additional spot market could be built from idle capacity. A provider trying to maximize his profit then faces two basic questions: what price should he ask in the fixed-price market and how many (possibly zero) spot instances should he offer. His profit then depends on the user’s actions given the offered market. Our model here is different from Abhishek et al. [2012] in two fundamental ways. First, we assume that the pool of fixed-price instances is not infinite. Instead, the provider offers only enough instances to able to serve demand and incurs a cost for every fixed-price instance. Second, we assume that for the spot market, the provider has a finite number of instances available that he can offer without incurring a cost. One-time costs like setting up a market are ignored, as they get amortized over time.

We analyze the structure of the resulting market equilibria depending on the provider’s strategy. We start with the assumption that being preempted does not carry a cost for the users besides the time waiting until a new instance becomes available. In our full model, we then relax this assumption and include a cost in the form of additional time required to reach the point a job was preempted at. This lost time is in addition to the time the job is not running. It models that many jobs lose some progress whenever preempted and require time to restart. Our main result then shows that for any strategy without spot instances the provider could choose, offering some spot instances would increase his profit while at the same time being a Pareto improvement for his customers. It is therefore in both relevant dimensions optimal to offer idle instances on a spot market. The main driver of this increase is that by moving some users from the fixed-price market to the spot market, the provider has to maintain fewer instances in the fixed-price market. The provider thus saves the costs for those instances. Conversely, users on average do pay enough in the spot market that the saved costs outweigh the revenue loss.

The obtained Pareto improvement for consumers especially shows that offering a spot market is advantageous for providers in competitive settings where multiple providers are competing for users, though we do not give an in-depth analysis of provider competition.

For illustration purpose, we finally give some numerical examples for the case when users arrive according to a Poisson process and require exponentially distributed service times. Here we visualize the effects that the pricing of the fixed price instances has on the equilibrium structure of the hybrid market.
Overall, our results have significant implications for practical market design. We show that it is in every provider’s best interest to sell instances that are idle for longer continuous intervals of time as spot instances, even if preemption is costly for users. We therefore encourage providers to, where technically and contractually possible, offer idle capacities on spot markets to increase efficiency and profit.

1.1 Related work

There is a host of research on the economics of short term cloud computing markets. An overview of the research are can be found in [Kash and Key, 2016]. Gao et al. [2016] recently studied the competition between two providers where one only offers a fixed-price market while the other only offers a spot market. They showed that this setting produces an equilibrium structure closely resembling what we observe whenever preemption carries no or a very low cost. They further showed that the competition between the providers does not seem to greatly influence the revenue per user. Zhang et al. [2013] presented a truthful online auction framework to accommodate different types of user profiles (as opposed to only looking at either users who only want a given job being run or who want certain resources). They analyzed the resulting social welfare, but made no statement about revenue or profit. Hoy et al. [2016] proposed a myopic model where users desire instances for one unit of time. Users are able to either directly buy an instance from a limited instance pool for a fixed price or wait and compete in a spot market for those instances that weren’t sold in the fixed price market. Under these assumptions they showed that offering such a two-step market is better than selling everything on the spot market. Agmon Ben-Yehuda et al. [2013] collected evidence suggesting that the EC2 spot market might not actually use true spot pricing. Instead they seem to use posted prices that vary over time. In the domain of ride-sharing platforms, [Banerjee et al., 2015] showed that a spot market, while not having a higher theoretically attainable revenue than fixed-price markets, can more easily realize its revenue potential if not all parameters are known to the provider or change over time.

Other strains of research studied the possibilities of increasing the systems efficiency and revenue by allowing advanced reservation of resources ([Azar et al., 2015], Babaioff et al., 2017]), predicting resource usage ([Cortez et al., 2017], Jyothi et al., 2016], Rajan et al., 2016]) or doing some form of price discrimination ([Kash et al., 2017], Xu and Li, 2013]).

2 MODEL

To analyze the profit a cloud provider obtains from running a short term cloud market, we need to model the effect its decisions have on the actions of potential customers. To model this, we define a two-step setting that is closely related to a Stackelberg game ([Maharjan et al., 2013]): first, the cloud provider chooses his actions and by doing so defines the game the users play. The users then play the resulting game only with each other, without further input from the provider. In contrast to a normal Stackelberg game, this second game takes the form of a queueing system. As the users are assumed to be rational, the characteristics of the provider’s potential strategies can be fully analyzed based on the equilibria of the queuing system. We denote a strategy of the provider by \( \eta \) and the aggregated strategy of all users by \( \xi \). Whenever a provider strategy \( \eta \) is given, \( \xi^* \) denotes a corresponding equilibrium user strategy.

2.1 Provider model

The provider defines the setting for the users. A strategy for him consists of a tuple \( \eta = (p, l_s) \). \( p \) is the price any a user joining the fixed price market has to pay per unit time his job is running. \( l_s \) is the number of spot instances the provider offers. \( l_s = 0 \) denotes that he decides to only offer a fixed
price market. We assume he only has a limited number of idle instances \( l \) available, the number of spot instances is therefore bounded by \( l_S \leq l \). For simplicity, we assume that the provider sets no reserve price for the spot market. Introducing a reserve price only strengthens our results as it broadens the strategy space available to make the spot market more profitable. The provider cannot directly choose how many fixed-price instances it offers. Instead, we assume that the offered fixed-price market is made sufficiently large for the amount of users joining it. Sufficiently large here means that for any user strategy \( \xi \) it consists of a number of instances \( l_F(\eta, \xi) \), such that any user joining it only has to wait a very short time \( T \) until his job is expected to get started. The expected waiting time \( T \) is assumed to be exogenously fixed and not part of the providers strategy space, but our results hold for every \( T \). As the provider strategy \( \eta \) influences the equilibrium strategy of the users \( \xi^* \), the number of instances \( l_F(\eta, \xi) \) can still be indirectly influenced by the provider. Any instance in the fixed-price market causes the provider a cost \( \kappa \) per unit of time. This is the fixed amount a new fixed-price instance costs, averaged out over the instances lifetime. The cost mostly consists of having to buy new (physical) servers (e.g. hardware, software, space etc., but excluding running costs like power) and does not need to mean that offering the spot instances is actually completely free. Given provider and user strategies \( \eta \) and \( \xi \), we denote by \( \Pi_F(\eta, \xi) \) and \( \Pi_S(\eta, \xi) \) the revenue from fixed-price and spot instances, respectively. The providers expected profit is then defined as

\[
\Pi(\eta, \xi) = \Pi_F(\eta, \xi) + \Pi_S(\eta, \xi) - l_F(\eta, \xi)\kappa \tag{1}
\]

### 2.2 User model

To model the resulting game for the users given provider strategy \( \eta \), we employ a combination of queueing theory combined and Bayes-Nash equilibrium analysis to model and analyze the different markets.

Queueing theory provides tools to analyze the expected queueing time of users (i.e. the time a newly arriving user has to wait until he gets an instance to run his job). By adding the known expected service time of the service process (i.e. the time a user’s job has to run on an instance), it allows us to derive the expected total waiting time of users (i.e. the time a newly arriving user has to wait until his job is run to completion). The basic queueing theoretic approach is then extended by assuming that every user is a rational agent who can choose to join or not to join one of the offered queues. They also have to pay some amount of money for getting their job serviced. We thus obtain a framework to analyze equilibria of the underlying market mechanism.

For the formal model, let there be \( n \) classes of jobs with fixed values \( v = (v_1, \ldots, v_n) \) for completion where \( v_i > v_{i+1} \) for all \( i \in \{1, \ldots, n\} \). New jobs arrive sequentially according to a stationary stochastic process with independent interarrival times. The arrival rates of the different job classes are \( \lambda = (\lambda_1, \ldots, \lambda_n) \) (i.e. in expectation, \( \lambda_i \) jobs of class \( i \) arrive per unit of time). Each job can run on exactly one instance and is associated with a distinct user. Thus, we can use the terms “user” and “job” interchangeably. The service time for each job (i.e. how long it needs to run on an instance until completion) is independently drawn according to some distribution with expected service time \( \frac{1}{\mu} \). The queueing theory literature often denotes this combination of general independent arrival and service processes \( GI/GI^\infty/\text{number of instances} \). For every job of class \( i \) that arrives, a

---

\( ^3 \)For ease of notation and in order to not have to specifically treat extreme corner cases where queues only consist of very few instances, we assume \( l_S \in \mathbb{R} \). As the cost of one additional instance is negligible in reasonable settings, this is a reasonable abstraction.

\( ^4 \)Note that our model can easily be modified such that, instead of a limit on the expected queueing time \( T \), some threshold on the percentage of rejections has to be met. Another possible modification is to assume that all or some users that cannot be served instantly simply balk instead. Neither of these modifications changes our central profit result.
waiting cost $c$ for every unit of time until job completion is independently and privately drawn from a distribution $F_i(c)$. This distribution has strictly positive PDF $f_i(c)$ on $[0, \mu v_i]$. As jobs with waiting cost $c > \mu v_i$ get a negative payoff even if they run instantly and pay nothing, jobs with higher waiting costs would never join and do not have to be considered. As is common in queueing theory, we assume that each job is infinitesimally small and does not affect the system dynamics on its own.

We call the tuple of exogenous parameters $(n, \nu, \lambda, F, \kappa, T, l)$ a setting. These parameters are assumed to be fixed and known by the provider.

For any single user, a possible strategy consists of the tuple $(\alpha, \beta)$. $(\alpha, \beta) \in \{F, S, B\}$ stands for the decision whether to join the fixed-price market $F$, the spot market $S$ or to balk $B$, i.e. not to join any market and obtain zero payoff. When the provider only offers a fixed-price market by setting $l_S = 0$, choosing the action $S$ would also mean to balk. Further, any user submits a bid $\beta$, which is used to determine his priority and payment when he joins the spot market. For any user that does not join the spot market this bid has no effect. For readability we w.l.o.g. set it to be equal to his waiting cost $c$. The current state of the markets is unobservable for users and thus cannot influence the strategy. An aggregated strategy $\zeta$ for all users gives a strategy for every possible class and waiting cost combination.

For any user that joins a queue, we now denote by $q(\alpha, \beta, \eta, \zeta)$ the expected queuing time (i.e. the time he spends in the system without running his job on an instance) when he plays strategy $(\alpha, \beta)$ assuming provider strategy $\eta$ and that all other users commit to the strategy encoded by $\zeta$.

The expected total waiting time until job completion is the sum of queuing time and service time $w(\alpha, \beta, \eta, \zeta) = q(\alpha, \beta, \eta, \zeta) + \frac{1}{\mu}$. The user has to pay some amount of money for using an instance. We denote this expected payment $m(\alpha, \beta, \eta, \zeta)$. Overall, the expected payoff a user obtains is thus given by $\pi^c_i(\alpha, \beta, \eta, \zeta) = v_i - c w(\alpha, \beta, \eta, \zeta) - m(\alpha, \beta, \eta, \zeta)$ for joining a market, and zero for balking.

### 2.3 Fixed-price Market and Queue

The fixed-price market consists of a queueing system where users have to pay a fixed price $p$ for every unit of time their job is running. In contrast to Abhishek et al. [2012], we assume that the queue has a finite number of instances $I_F(\zeta)$. This makes it impossible to guarantee that every user that arrives can instantly run his job, as the number of users that potentially arrive in one unit of time is unbounded. We instead assume that the number of instances $I_F(\eta, \zeta)$ the provider offers is set depending on the aggregate user strategy $\zeta$ such that the expected queuing time $q(\zeta)$ of any user that joins the fixed-price market is equal to some small queuing time limit $T$. Users that arrive while no instance is available are assumed to continuously keep resubmitting their request and are served in random order when instances become available.

In further contrast to Abhishek et al. [2012], the expected payoff $\pi^c_i(\alpha, \beta, \eta, \zeta)$ of a user of class $i$ with waiting cost $c$ that joins the fixed-price market now also depends on the queuing time $T$ and is given by

\begin{align*}
\pi^c_i(F, \beta, \eta, \zeta) &= v_i - c w(F, \beta, \eta, \zeta) - m(F, \beta, \eta, \zeta) \\
&= v_i - c(T + \frac{1}{\mu}) - \frac{p}{\mu}. \quad (2)
\end{align*}

Note that this payoff is independent of the action of any other user, it only depends on the providers choice for the price $p$.

---

5For ease of notation and calculation, we generally assume that $I_F \in \mathbb{R}$ and therefore that $q(F, \beta, \eta, \zeta) = T$ holds exactly. This could be seen as either offering a slower instance or an instance that is not always active. For reasonably sized markets (> 1000) this is generally a very weak abstraction as the price of one instance negligible.
2.4 Spot Market and Queue

The spot market consists of a preemptible priority queue where both payments and the order in which jobs get run depend on the users’ bids. It is thus completely equivalent to [Abhishek et al., 2012]. The preemptible priority queue consists of \( l_S \) (where \( l_S \) is set by the providers’ strategy) instances running jobs in a priority-based order. A job’s priority is set by the bid given on arrival. Servicing is preemptible, which means that whenever a user submits a high bid and no instance is free, then the lowest priority job currently running is interrupted. We assume that the interrupted job can later continue from where it left off, without loss, except for the additional queuing time. In Section 4 we analyze the full model that includes costs associated with every preemption. Payment in the spot market are set according to some spot pricing mechanism that is not explicitly modeled. Abhishek et al. [2012] showed that for any Bayes Nash Incentive Compatible (BNIC) mechanism, expected payments have to equal the payments of a mechanism where a user’s bid \( \beta \) only consists of a revelation of its true waiting cost \( c \), i.e. \( \beta = c \). It is therefore enough to only analyze the Bayes-Nash equilibria for such a mechanism. In the following we will use bid and waiting cost interchangeably and assume a user’s priority to be equal to his bid. The payments of any such mechanism are uniquely defined by its allocation.

The expected total waiting time \( w(S, c, \eta, \zeta) \) of a user in the spot market does only depend on the number of instances the provider offers and the users that join the spot market with a higher bid.

From Lemma 5 of Abhishek et al. [2012] we know for any BNIC pricing rule employed in the spot market that the expected payments is:

\[
m(S, c, \eta, \zeta) = \int_0^c w(S, x, \eta, \zeta)dx - cw(S, c, \eta, \zeta)
\]

The total waiting time, and thus the expected payment, depend on the number of users joining the spot market encoded by the strategy \( \zeta \). Formalized for any user of class \( i \) with waiting cost \( c \) that joins the spot queue this means that the expected payoff \( \pi^c_i(S, c, \eta, \zeta) \) is

\[
\pi^c_i(S, c, \eta, \zeta) = \nu_i - cw(S, c, \eta, \zeta) - m(S, c, \eta, \zeta)
\]

\[
= \nu_i - \int_0^c w(S, x, \eta, \zeta)dx.
\]

3 EQUILIBRIA AND PROFIT

We now analyze the resulting Bayes-Nash equilibria (BNE’s) of the user game dependent on the provider strategy \( \eta \). When only fixed-price instances are offered, i.e. \( l_S = 0 \), in equilibrium either some users join the fixed-price market or everyone balks when the price is too high. Whenever at least some spot instances are offered, i.e. \( l_S > 0 \), some users will join the spot market. But equilibria still look vastly different depending on whether anyone joins the fixed-price market or not. In the following, we first characterize all possible equilibria. We describe the three basic types: pure fixed-price, pure spot and proper hybrid equilibria. We then show how to identify in which type any given provider strategy \( \eta \) would result. Lastly we compare the expected profit of different strategies \( \eta \) and show that it is always in the providers best interest to play \( l_S > 0 \), i.e. to offer spot instances.

3.1 Pure Fixed-price Equilibria

We have already seen in section 2.3 that the payoff of a user in the fixed-price market does not depend on the actions of the other users. When no spot market is offered, the strategy of any
user is therefore independent of the other users. The payoff \( \pi_i^F(\mathcal{F}, \beta, \eta, \zeta) = v_i - c(T + \frac{1}{\mu}) - \frac{p}{\mu} \) for submitting a job is clearly decreasing in \( c \) for every class. From this we easily see that there exists a unique waiting cost cutoff vector for the fixed-price market \( \mathcal{c}^F = (c_1^F, \ldots, c_n^F) \) where for every class \( i \) the payoff is \( \pi_i^F(\mathcal{F}, \beta, \eta, \zeta) = 0 \). To maximize payoffs, users with jobs of class \( i \) with waiting cost below \( c_i^F \) join the fixed-price queue. Users with with waiting cost above \( c_i^F \) do not join and obtain a payoff of zero. We denote the users playing such an aggregate strategy is by \( \zeta = \mathcal{c}^F \).

**Proposition 3.1.** For any provider strategy \( \eta = (p, l_S) \), any equilibrium strategy takes the form \( \zeta^* = \mathcal{c}^F \) and is uniquely defined by

\[
c_i^F = \frac{v_i - p\frac{1}{\mu}}{(T + \frac{1}{\mu})} \quad \forall i \in N.
\]

**Proof.** Whenever \( l_S = 0 \) \( \zeta^* = \mathcal{c}^F \) clearly has to hold. \( c_i^F \) follows by simple algebra from setting \( \pi_i^F(\mathcal{F}, \beta, \eta, \zeta) = v_i - c_i^F(T + \frac{1}{\mu}) - \frac{p}{\mu} = 0 \).

The revenue \( \Pi_F \) of the pure fixed-price market with price \( p \) in any BNE with cutoff vector \( \mathcal{c}^F \) then simply arises as the arrival rate of all classes of jobs into the market multiplied with their payment.

\[
\Pi_F(\eta; \mathcal{c}^F) = \frac{p}{\mu} \left( \sum_i \lambda_i F_i(c_i^F) \right)
\]

The profit is consequently

\[
\Pi(\eta; \mathcal{c}^F) = \frac{p}{\mu} \left( \sum_i \lambda_i F_i(c_i^F) \right) - \kappa l_F(c_i^F)
\]

Thus, in order to maximize his profit, the provider has to optimize the price \( p \). As we have seen, the cutoff vector \( \mathcal{c}^F \) in equilibrium, and consequently the number of instances \( l_F(c_i^F) \), are uniquely set by the price and the setting.

### 3.2 Pure Spot Equilibria

Whenever the provider chooses the fixed-price price \( p \) high enough, no user will want to submit his job to the fixed-price market, either because his payoff would be negative (if \( p > v_i \)) or because he would obtain a higher payoff in the spot market. The equilibria in those cases have been fully analyzed by Abhishek et al. \[2012\]. We repeat these results here, adjusted for our notation. Like in the pure fixed-price market case, there again exists a unique waiting cost cutoff vector \( \mathcal{c}^S = (c_1^S, \ldots, c_n^S) \).

**Theorem 3.2.** \[Abhishek et al., 2012\] For any provider strategy \( \eta = (p, l_S) \), in any BNE of the user game where no user joins the fixed-price market it holds that \( \zeta^* = \mathcal{c}^S \) and the cutoff vector \( \mathcal{c}^S = (c_1^S, \ldots, c_n^S) \) is the unique solution to the following system of equations:

\[
v_i - \int_0^{c_i^S} w(S, x, \eta; \mathcal{c}^S) dx = 0 \quad \forall i \in N
\]

The reason for this is that, for any fixed strategy \( \zeta \) of the other users with truthfully reported waiting costs, the payoff \( \pi_i^S(S, c, \eta, \zeta) \) is monotone decreasing in \( c \). Given the cutoff vector, the profit \( \Pi \) only consists of the spot market’s revenue, as instances are required for the fixed-price market and the spot instances come free. The revenue can be calculated as the sum over the average
payments of users of class $i$. These are given as the integrals over the payments at every waiting cost, multiplied by the arrival rate of class $i$:

$$
\Pi(\eta; \bar{c}^S) = \Pi_S(\eta; \bar{c}^S)
$$

(11)

$$
= \sum_i \lambda_i \int_0^{\bar{c}^S_i} \left( \int_0^{y} w(S, x, \eta, \bar{c}^S) \, dx \right) f_i(y) \, dy
$$

(12)

3.3 Hybrid Equilibria

We now analyze those equilibria when the provider plays $l_S > 0$, i.e. when there exists a spot market, and some users still join the fixed-price market.

Users with waiting cost $c = 0$ are willing to wait arbitrarily long and can get serviced for free in the spot market. Conversely, the users with the highest waiting costs that join the spot queue are finished faster than they would be in fixed-price queue. They are thus willing to pay more than $p$ per unit of time. This means that both the users with the lowest and highest waiting costs that join any market join the spot queue. The users that join the spot queue are therefore going to be split into two types: those that pay a lot and get served faster than in the fixed-price queue (and are quite unlikely to ever get preempted for being outbid) and those that wait longer and pay less than the fixed-price price. Between these lies a continuous interval of waiting costs containing the users that join the fixed-price queue.

**Lemma 3.3.** For any provider strategy $\eta = (p, l_S)$, in any BNE of the user game where some users join the fixed-price market, there exists an interval $[c^L, c^U]$, such that almost all users, i.e. all besides possibly a null set that doesn’t influence system dynamics, with waiting costs $c \in [c^L, c^U]$ join the fixed-price queue. For these waiting costs $c$, the total waiting time and the payment would be equal for either queue:

$$
p \frac{1}{\mu} = m(S, c, \eta, \xi^*)
$$

(13)

$$
= m(F, c, \eta, \xi^*)
$$

(14)

$$
T + \frac{1}{\mu} = w(S, c, \eta, \xi^*)
$$

(15)

$$
= w(F, c, \eta, \xi^*)
$$

(16)

For waiting costs $c \notin [c^L, c^U]$ it holds

$$
\pi_i(S, c, \eta, \xi^*) > \pi_i(F, c, \eta, \xi^*)
$$

(17)

and these users join the spot queue or no queue at all.

**Proof.** The proof can be found in the e-companion B

This leads to the existence of two general cutoff points in equilibrium: A waiting cost $c^L$ where a user is still served as fast in the spot queue as in the fixed-price queue but also has to pay the same. Below $c^L$, a user that joins the market will thus join the spot queue, above $c^L$ the fixed-price queue. Then a cost $c^U$ above which users join the spot market again, as they will then be served faster than in the fixed-price market and are thus willing to pay more than $p$.

Since the payoff in both markets is monotone decreasing in the waiting cost, there is again a cutoff vector $\bar{c}^H = (c^H_1, \ldots, c^H_n)$ denoting the waiting cost for each class $i$ above which a user would obtain negative payoff in either market and thus will not join. Whether $c^H_i$ is above or below $c^L$ and/or $c^U$ consequently determines if any users of class $i$ join the fixed-price market and/or the part of the spot market where they pay more than $p$. For notational reasons we introduce...
cutoff vectors $\overrightarrow{c}_L = (c^L_1, \ldots, c^L_n)$ and $\overrightarrow{c}_U = (c^U_1, \ldots, c^U_n)$ that denote for every class of jobs up to which waiting costs users join the spot market paying less than $p$ and the fixed-price market, respectively. They are defined as $c^L_i = \min \{c^L, c^H_i\}$ and $c^U_i = \min \{c^U, c^H_i\}$. For any class of users $i$ where $c^H_i < c^L$, no user joins the fixed-price queue. When $c^L < c^H_i < c^U$, the users of class $i$ with the highest waiting cost that do not balk do join the fixed-price queue. For classes where $c^U < c^H_i$, the top waiting cost users join the spot market again. We denote the users playing according to this aggregate strategy by $\xi = (\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H)$.

We can now use Lemma 3.3 to characterize any such equilibrium strategy as the solution of a system of equations.

**Proposition 3.4.** For any provider strategy $\eta = (p, l_S)$, in any BNE of the user game where any user joins the fixed-price market it holds $\xi = (\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H)$ and the cutoff vectors $\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H$ are the unique solution to the following system of equations:

\begin{align}
0 &= c^L(T + \frac{1}{\mu}) + p - \frac{1}{\mu} \int_0^{c^L} w(S, x, \eta, (\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H)) dx \quad (18) \\
0 &= c^U(T + \frac{1}{\mu}) + p - \frac{1}{\mu} \int_0^{c^U} w(S, x, \eta, (\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H)) dx \quad (19) \\
0 &= v(i) - \int_0^{c^U} w(S, x, \eta, (\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H)) dx \quad \forall i \in N \quad (20)
\end{align}

With aggregate strategy $\xi = (\overrightarrow{c}_L, \overrightarrow{c}_U, \overrightarrow{c}_H)$, a user of class $i$ with waiting cost $c$ will join the hybrid market if and only if $c \leq c^H_i$. Almost all users, i.e. all besides possibly a null set that doesn’t influence system dynamics, will choose the fixed-price queue if and only if $c^L \leq c \leq c^U$ and the spot queue otherwise.

**Proof.** That equations 18, 19 and 20 have to hold immediately follows from Lemma 3.3. The proof that a unique solution always exists can be found in appendix C.

Recall that, for every class $i$, the arrival rate into the market is given by $\lambda_i$ and the distribution function of the waiting cost is $F_i(c)$. To further simplify notation denote by $\lambda_F(\eta, \xi) = \sum_i \lambda_i(F_i(c^L_i) - F_i(c^U_i))$ the arrival rate into the fixed-price market given strategy $\xi$. Given the strategy $\eta = (p, l_S)$ and resulting equilibrium strategy $\xi^* = (\overrightarrow{c}_L^*, \overrightarrow{c}_U^*, \overrightarrow{c}_H^*)$, the profit of the hybrid market can then simply be calculated as the sum of the revenue of its fixed-price queue and spot queue, minus the cost of the offered instances:

\begin{align}
\Pi(\eta, \xi^*) &= \Pi_F(\eta, \xi^*) - \kappa l_F + \Pi_S(\eta, \xi^*) \\
&= \frac{p}{\mu} \lambda_F(\eta, \xi^*) - \kappa l_F \quad (21)
\end{align}

\begin{align}
&+ \sum_i \lambda_i \left[ \int_0^{c^L_i} m(S, y, \eta, \xi^*) f_i(y) dy + \int_{c^H_i}^{c^U_i} m(S, y, \eta, \xi^*) f_i(y) dy \right]
\end{align}

The revenue in each queue is calculated similarly to the pure fixed-price and spot market (see Equations (9) and (12)) and only modified for the cutoff vectors of the hybrid market.

### 3.4 Resulting Equilibria

We now analyze which kind of equilibrium results from a given provider strategy $\eta$.

As users with very low waiting cost always join the spot queue, pure fixed-price equilibria can only result when no spot instances are offered, i.e. when the provider plays $l_S = 0$. It is not directly
visible whether a given strategy with \( l_s > 0 \) degenerates the market to a pure spot market or results in a proper hybrid equilibrium. Fortunately, enforcing a pure spot market is always possible by simply removing the option for users to join the fixed-price market. Because of this, \( c^S \) as the solution to equation 10 exists even if it is not the equilibrium strategy for a given \( \eta \). This allows us to formulate a criterion for whether it would be rational for any player to deviate from the strategy \( \bar{\zeta} = \bar{c}^S \). When this criterion is satisfied, any equilibrium must have some users join the fixed-price market, leading to a proper hybrid outcome.

**Definition 3.5.** For any \( \eta \) with \( l_s > 0 \) we say the hybrid market is proper if a waiting cost \( \bar{c} \) exists, such that the total waiting time at bid \( \bar{c} \) for joining the spot market given \( \bar{\zeta} = \bar{c}^S \) is the same as the fixed-price market waiting time, i.e.

\[
\begin{align*}
\text{w}(S, \bar{c}, \eta, \bar{c}^S) &= T + \frac{1}{\mu}.
\end{align*}
\]

and for a user with waiting cost \( \hat{c} \) deviating from \( \zeta \) and joining the fixed-price market would be a best response, i.e.

\[
\hat{c}(T + \frac{1}{\mu}) + \mu^{-1} < \int_0^{\hat{c}} \text{w}(S, x, \eta, \bar{c}^S)dx
\]

This allows us to formulate the following result for deciding in what equilibrium a given provider strategy \( \eta \) would result.

**Proposition 3.6.** For any provider strategy \( \eta \), the aggregate equilibrium strategy of the users is

- \( \bar{\zeta}^* = \bar{c}^F \) if and only if \( l_s = 0 \).
- \( \bar{\zeta}^* = \bar{c}^L, \bar{c}^U, \bar{c}^H \) if and only if \( l_s > 0 \) and the market is proper.
- \( \bar{\zeta}^* = \bar{c}^S \) otherwise

**Proof.** The formal proof is found in the e-companion D □

### 3.5 Profit

Given the equilibrium strategies we can now finally analyze the profit of different provider strategies \( \eta \) and show the optimality of offering a spot market. Note that for general arrival and service processes, the equilibrium strategies cannot be directly calculated. This is because there exists no general formula for the total waiting times in either queue type. Our results in this section are therefore based on analyzing the structure of the equilibria. For those interested, a concrete formulation for the special case of a Poisson arrival rate and exponential service times, a concrete formulation can be found in the e companion.

Our main result is that choosing a strategy that results in a proper hybrid market will always lead to a higher profit and a Pareto improvement for the user compared to only offering a fixed-price market. This means that a provider that has access to idle instances should sell at least some of them on a spot market. This holds no matter whether he only wants to optimize for profits or if he also cares about user welfare.

**Theorem 3.7.** For every provider strategy that doesn’t offer a spot market \( \eta_0 = (p_0, 0) \), there exists a strategy \( \eta = (p, l_s) \) with \( 0 < l_s \leq l \) that yields a higher profit for the provider, i.e.

\[
\Pi((p, l_s), \bar{\zeta}^*) > \Pi((p_0, 0), \bar{\zeta}^*)
\]

and a Pareto improvement for the users, i.e.

\[
\forall c \; \pi^c_\epsilon(\alpha, \beta, \eta, \bar{\zeta}^*) \geq \pi^c_\epsilon(\alpha, \beta, \eta_0, \bar{\zeta}^*)
\]

\[
\exists c \; \text{s.t.} \; \pi^c_\epsilon(\alpha, \beta, \eta, \bar{\zeta}^*) > \pi^c_\epsilon(\alpha, \beta, \eta_0, \bar{\zeta}^*)
\]

\[ 25 \]

\[ 26 \]

\[ 27 \]
This theorem shows that it will always be advantageous for a cloud provider to offer at least some spot instances in addition to the fixed-price market if he has idle capacities that can be used without acquiring additional instances. This is the case even in settings with competition or when user satisfaction is more important than profit. Note however that this does not mean that the profit maximizing strategy and the welfare maximizing strategy for the provider are identical. It only means only that both offer a spot market and that there is a strategy that where profit and welfare are better than for any strategy that does not offer a spot market.

One problem that Theorem 3.7 does not solve is to find optimal strategies. This inability to easily calculate equilibrium strategies ζ for a given provider strategy η might make it seem that offering a spot market could still be risky. Luckily, simply using the price that a provider would use when only offering a fixed-price market can simply be used in the hybrid market without risk of losing profit, as long as the spot market is small enough.

Proposition 3.8. For every provider strategy that doesn’t offer a spot market \( \eta_0 = (p_0, 0) \), there exists a number of spot instances \( l^*_S > 0 \) such that any strategy \( \eta = (p_0, l_S) \) with \( l_S \leq l^*_S \) yields a higher profit for the provider, i.e.

\[
\Pi((p_0, l_S), \zeta^*) > \Pi((p_0, 0), \zeta^*)
\]

(28)

and a Pareto improvement for the users, i.e.

\[
\forall c \pi^c_i(\alpha, \beta, \eta, \zeta^*) \geq \pi^c_i(\alpha, \beta, \eta_0, \zeta^*)
\]

(29)

\[
\exists c \text{ s.t. } \pi^c_i(\alpha, \beta, \eta, \zeta^*) > \pi^c_i(\alpha, \beta, \eta_0, \zeta^*)
\]

(30)

Proof. Follows directly from the proof of Theorem 3.7. □

Calculating \( l^*_S \) would again require to calculate the equilibrium strategies \( \zeta \). A way around this in practice would be to start with a very rather spot market size \( l_S \) and slowly scaling it up over time.

4 COSTS OF PREEMPTION

The model given so far assumed that preemption does not incur any actual cost for the user besides the increase in queueing time while waiting to be serviced again. In reality, this is clearly not the case: few jobs can continue without any loss from the point they left off, so that congestion in the spot queue becomes more expensive. In this section, we add costs for getting preempted to our model. A job can either get preempted by a job of higher priority in the market (internal preemption) or because the instance it is running on becomes unavailable for some exogenous reason independent of the job’s priority (external preemption); like the instance being required for its dedicated purpose. Assume that a preemption carries a cost in form of an additional expected time loss \( \tau \), meaning that the expected payoff of a user with waiting cost \( c \) decreases by \( c \tau \) for every expected preemption. We assume that \( \tau \) is independent of a user’s class. Analyzing preemption costs that varies between classes, while interesting, requires extensive treatment of different cases and is beyond the scope of this paper. It needs to be reserved for future work.

How often a user’s job gets internally preempted, i.e. outbid while in service, depends on the arrival rate of users with a higher bid.

Define \( \Pi(c, \eta, \zeta) \) as the expected number of times a user with bid \( c \) in the spot market gets internally preempted by other users over the runtime of his job.

The number of external preemptions, i.e. how often the instance his job runs on is removed from the market, does not depend on his priority. External preemptions in expectation happens
to all user in the spot queue on average \( t_E(l_S) \) times. It can depends on the providers strategy \( \eta \), as offering more spot instances \( l_S \) means he has to utilize some that are not as reliably idle. We assume \( t_E(l_S) \) is monotonically increasing and set \( t_E(0) = 0 \) as convention.

A user’s expected total waiting time for when joining the spot market therefore is

\[
w(S, c, \eta, \zeta) = q(S, x, \eta, \zeta) + \frac{1}{\mu} + \tau(t_I(c, \eta, \zeta) + t_E(l_S)).
\] (31)

Internal preemption does not affect the equilibrium structure at all besides reducing the overall number of users that join the spot market. This is because it only changes how fast waiting times grow for lower bids. Instead, most changes to the equilibrium structure come from external preemption. This is because when external preemption becomes too common and costly, users with high waiting cost start to prefer the fixed-price market. This can lead to an equilibrium structure that more closely resembles [Abhishek et al., 2012]. When the overall time loss of external preemption \( t_I(c, \eta, \zeta) \) is higher than the expected queuing time in the fixed-price market \( T \), no rational user will ever be willing to pay more in the spot market than in fixed-price market. This means that even if \( \eta \) that leads to a proper hybrid markets, there will only be one general cutoff point \( c^p \). Below it users that do not balk join the spot market and above which they join the fixed-price market. A cutoff vector \( \bar{c}^B = (c^B_1, \ldots, c^B_n) \) above which users would obtain negative payoff and balk exists as normal.

**PROPOSITION 4.1.** For any provider strategy \( \eta = (p, l_S) \) with \( l_S > 0 \) and \( T < t_I(c, \eta, \zeta^*) \), in any BNE of the user game where any user joins the fixed-price market it holds \( \zeta^* = (c^p, \bar{c}^B) \). With aggregate strategy \( \zeta = (c^p, \bar{c}^B) \), a user of class \( i \) with waiting cost \( c \) will join the spot market when \( c < c^p \leq c^B_i \) and the fixed-price market when \( c^p < c < c^B_i \). He will balk and not join any market when \( c > c^B_i \).

\( c^p, \bar{c}^B \) are the unique solution to the following system of equations:

\[
0 = c^p(T + \frac{1}{\mu}) + \frac{1}{\mu} \int_0^{c^p} w(S, x, \eta, (c^p, \bar{c}^B))dx
\] (32)

\[
0 = v(i) - \min \left\{ \frac{p}{\mu} + c^B_i(1/\mu + T), \int_0^{c^B_i} w(S, x, \eta, (c^p, \bar{c}^B))dx \right\} \quad \forall i \in N
\] (33)

**PROOF.** Users with waiting cost close to zero will always prefer the spot queue, no matter how much time they loose. Since some users join the fixed-price market, there has to be a smallest point \( c^p \) such that

\[
c^p(T + \frac{1}{\mu}) + \frac{1}{\mu} \int_0^{c^p} w(S, x, \eta, (c^p, \bar{c}^B))dx
\] (34)

Since \( T < t_I(c, \eta, \zeta) \) it has to hold \( m(S, c^p, \eta, \zeta) < \frac{p}{\mu} \) and there cannot exist any \( c > c^p \) with

\[
c^p(T + \frac{1}{\mu}) + \frac{1}{\mu} \int_0^{c^p} w(S, x, \eta, \zeta)dx > \int_0^{c^p} w(S, x, \eta, \zeta)dx
\] (35)

No user with waiting cost greater \( c^p \) thus joins the spot market. This means the spot market can be fully defined by the actions of players with \( c < c^p \). For this denote by \( \bar{c}^S \) the vector of cutoff points at which a job becomes indifferent between the spot market and either the fixed-price market or balking. It holds
This is because for users with very low waiting even very high preemption costs do not influence $\zeta$ a strategy preemption, though the proof becomes a bit more involved, directly equivalent to Theorem 3.7 and Proposition 3.8 thus hold for hybrid markets with external their payoff much. They willing to wait a long time for very small reductions in payment. Results □ any equilibrium.

Otherwise, no one prefers the fixed-price market over the spot market and

\[
\text{if } c_i^S < c_i^B, \quad \text{else }
\]

\[
c_i^B = \left\{ \begin{array}{ll}
 c_i^S, & \text{if } c_i^S < c_i^S \\
 \frac{\mu_i - c_i^S}{T + 1}, & \text{else } (37)
\end{array} \right.
\]

With this we can now extend Proposition 3.6 to markets with external preemption.

**Corollary 4.2.** For any provider strategy $\eta$, the aggregate strategy equilibrium strategy of the users is

- $\zeta^* = \overline{c}^F$ if and only if $l_S = 0$.
- $\zeta^* = (\overline{c}^L, \overline{c}^U, \overline{c}^H)$ if and only if $l_S > 0$, the market is proper and it holds $t_E(l_S) \tau < T$
- $\zeta^* = (c^P, \overline{c}^B)$ if and only if $l_S > 0$, it holds $T \leq t_E(l_S) \tau$ and $\nu_1 - \frac{\mu}{\mu + 1} > 0$
- $\zeta^* = \overline{c}^S$ otherwise

**Proof.** If $l_S = 0$, trivially no user can join the spot market. Otherwise, a user with waiting cost close enough to zero will prefer any waiting time to paying $\frac{p}{\mu} > 0$.

If $t_E(l_S) \tau < T$ the proof is equivalent to Proposition 3.6.

If $T < t_E(l_S) \tau < \frac{p}{\mu} + T$, no one is willing to pay a higher price in the spot market than in the fixed-price market. The equilibrium strategy therefore has to be $\zeta^* = (c^P, \overline{c}^B)$ or $\zeta^* = \overline{c}^S$. If $\nu_1 - \frac{\mu}{\mu + 1} c^S > 0$, then some users that would have zero payoff when playing $\zeta = \overline{c}^S$ could deviate by joining the fixed-price market in order to increase their payoff. The strategy therefore has to be $\zeta^* = (c^P, \overline{c}^B)$.

Otherwise, no one prefers the fixed-price market over the spot market and $\zeta^* = \overline{c}^S$ is played in any equilibrium. □

Perhaps surprisingly, the profit and welfare optimality of offering a spot market is retained. This is because for users with very low waiting even very high preemption costs do not influence their payoff much. They willing to wait a long time for very small reductions in payment. Results directly equivalent to Theorem 3.7 and Proposition 3.8 thus hold for hybrid markets with external preemption, though the proof becomes a bit more involved,

**Theorem 4.3.** For every provider strategy that doesn’t offer a spot market $\eta_0 = (p_0, 0)$, there exists a strategy $\eta = (p, l_S)$ with $l_S \leq 1$ that yields a higher profit for the provider, i.e.

\[
\Pi((p, l_S), \zeta^*) > \Pi((p_0, 0), \zeta^*)
\]

and a Pareto improvement for the users, i.e.

\[
\forall c \quad \pi^c_i(\alpha, \beta, \eta, \zeta^*) \geq \pi^c_i(\alpha, \beta, \eta_0, \zeta^*)
\]

\[
\exists c \text{ s.t. } \pi^c_i(\alpha, \beta, \eta, \zeta^*) > \pi^c_i(\alpha, \beta, \eta_0, \zeta^*)
\]
Proof. For settings where there exists \( l_S > 0 \) with \( l_E(l_S) \tau < T \), this theorem can be proofed equivalently to Theorem 3.7. Otherwise the proof is more involved. It is given in Appendix F. □

As in queues without preemption costs, the optimal price does not need to be known in order to take advantage of a spot market. Simply using the price used when only offering a fixed-price market does again increase profits and welfare.

**Proposition 4.4.** For every provider strategy that doesn’t offer a spot market \( \eta_0 = (p_0, 0) \), there exists a number of spot instances \( l_S^* > 0 \) such that any strategy \( \eta = (p_0, l_S) \) with \( l_S \leq l_S^* \) yields a higher profit for the provider, i.e.

\[
\Pi((p_0, l_S), \zeta^*) > \Pi((p_0, 0), \zeta^*)
\]

and a Pareto improvement for the users, i.e.

\[
\forall c \pi^c_i(\alpha, \beta, \eta, \zeta^*) \geq \pi^c_i(\alpha, \beta, \eta_0, \zeta^*)
\]

\[
\exists c \text{ s.t. } \pi^c_i(\alpha, \beta, \eta, \zeta^*) > \pi^c_i(\alpha, \beta, \eta_0, \zeta^*)
\]

Proof. Follows directly from the proof of Theorem 4.3. □

### Numerical Example

In the following we will give some illustrating numerical examples of the equilibrium structure. While no exact formulas exist for calculating waiting times of general queue of either market, the well known Erlang C formula allows us to calculate waiting times for queues with memoryless processes. This allows us to in turn calculate the cutoff vectors and profits for the following illustrative example with two classes of jobs. All required calculations can be found in the e-companion. The parameters of the example are chosen as follows: the values for completion are \( \nu = (1, 0.75) \), the arrival rates are \( \lambda = (100, 50) \), and the expected service time is \( \frac{1}{\mu} = 1 \). The waiting costs are uniformly distributed on \([0, 1]\) and \([0, 0.75]\), respectively. The bound on the expected total waiting time for the fixed-price market is set to \( T = 0.001 \) and there are \( l = 20 \) spot instances available. To keep the example concise, we only give results for \( l_S = 0 \) and \( l_S = l = 20 \). The preemption cost is \( \tau = 0 \).
For $l_S = 20$, Figures 1 and 2 illustrate the user strategy $\eta^*$ for class 1 and 2 respectively. The price $p$ is varied over 400 uniformly spaced gridpoints between 0 and 1. Remember that jobs of class 1 have a higher value for completion that those of class 2 and are thus willing to pay at the same waiting cost. Depending on the price of the fixed-price market and their waiting cost, users with waiting costs in the horizontally shaded area join the fixed-price queue while those in the vertically shaded areas join the spot queue. For prices above $p \approx 0.796$, where there is no horizontally shaded area, the hybrid market is no longer proper and no one joins the fixed-price market. We can also see that for this setting, no user of class 2 is joining the spot queue with a bid high enough to queue less than $T$, independent of what value $p$ takes. This is simply because there are always enough class one users that are willing to pay more for the privilege of getting served quickly. The number of users from both classes in the spot market stays roughly constant for low to medium $p$, as can be seen by the roughly constant height of the vertically shaded area. But once the price is high enough, the class one users start to push out almost all class two users and $e_2^H$ goes almost to zero. Note that while the figure makes it seem as no user from class two joining the spot queue for $p > 0.796$, this is not the case. For prices where no users join the fixed-price market, still about 0.3 of the 50 class 2 users that arrive per unit of time do join. All others are displaced by users with jobs of the first class willing to pay 0.25 more for the same total waiting time.

In Figure 3, we show the profit of the provider as a function of the price $p$ for $\kappa = 0.25$ for $l_S = 0$ and $l_S = 20$. The profit for any $p$ depends on two factors: how many users join and how much they pay. As the price goes up, average payments go up, but overall fewer users join. This means that the profit goes up until the rise in payments is overwhelmed by the loss of users. For $l_S = 20$ it then falls until the market degenerates to the spot market, from which point on higher prices have no effect. Shortly before no user joins the fixed-price market any more, the profit falls below what would be gained if all users would choose the spot market. This is because for such few users, disproportionately many fixed-price instances have to be kept to ensure the queueing time limit $T$. For these prices the fixed-price market would operate at a loss. Overall, in Figure 3 it can be clearly seen that the profit is far higher when offering spot instances than when only offering fixed-price market for any price.

Finally, Figure 4 shows the maximal attainable profit for all values of $\kappa$. It again contains the plots for $l_S = 0$ and $l_S = 20$, as well as for comparison a market where no fixed-price instances are offered. Even with $l_S = 20$, offering spot instances yields a higher profit to the provider than not doing so for instance costs as small as $\kappa = 0.007$. While Theorem 3.7 only shows that offering a hybrid market increases the profit for small $\kappa$ if we set $l_s$ small enough, this example shows that for realistic values of $\kappa$, the spot queue of the hybrid market can be quite large and still lead to a profit increase over the fixed-price market. For even smaller $\kappa$, a smaller spot market would be required.

6 CONCLUSION
In this paper, we have shown that selling some idle instances on a spot market will increase a cloud provider’s profits even if the provider already runs a fixed-price market and preemption is costly for the users. We have also found that the provider can continue to ask the fixed-price price $p$ that he asks in his existing fixed-price market in the hybrid market. Doing so increases his profit and the social welfare of the hybrid market over the fixed-price market. Overall, while our analysis necessarily takes place in an idealized setting that makes a number of assumptions, most do not influence the results. For example introducing a reserve price for the spot market can only increase the attainable profit. Users that cannot immediately be serviced in the fixed-price market balk instead of waiting would necessitates some changes in the model, notably replacing the expected waiting time $T$ with a limit on the expected number of rejections, but the same general
While competition between providers is not directly modeled, offering spot instances at the same price is clearly still optimal as it leads to a Pareto improvement for the users. Our model allows fractional instances to be offered in order to keep the exposition simple and avoid requiring special handling of extreme corner cases, but for any reasonably sized market the price of one more instance is negligible. As an alternative to our proposed use of idle instances, providers could also offer their idle instances on a preemptible fixed-price market. While a spot market with reserved price should theoretically produce more profit, the simplicity of fixed-prices, both for the technical set-up and for users to understand the market, should not be underestimated. Further research comparing both approaches is required to fully characterize the trade-off. Also note again that exogenous factors that are not part of the model, like technical or legal problems setting up a functional spot market out of idle capacity, can still mean that introducing spot instances is not viable for some providers. Baring that, our results imply that it is in any cloud provider’s interest to offer idle instances on a secondary spot market.

REFERENCES


Darrell Hoy, Nicole Immoorlica, and Brendan Lucier. 2016. On-demand or Spot? Selling the cloud to risk-averse customers. In International Conference on Web and Internet Economics. Springer Berlin Heidelberg, 73--86.


Ying Yan, Yanjie Gao, Yang Chen, Zhongxin Guo, Bole Chen, and Thomas Moscibroda. 2016. TR-Spark: Transient Computing for Big Data Analytics. In SoCC.

A BASIS OF THE NUMERICAL EXAMPLE

To calculate the numerical example we need to calculate the expected waiting time of jobs given \( \eta \) and \( \zeta \).

**Proposition A.1.** For any preemptive priority queue with \( l_s \) instances, the expected total waiting time of a user with bid \( c \) is given by

\[
\omega(S, c, \eta, \zeta) = \frac{1}{\mu \varphi(c, l_s, \zeta)},
\]

where \( \varphi(c, l_s, \zeta) \) is the probability that fewer than \( l_s \) jobs with a higher bid than \( c \) are currently in the queue at any point in time given aggregate user strategy \( y \).

**Proof.** Note that jobs with a lower bid do not influence the total waiting time of a user at all and can thus be ignored. As the probability mass for any other user to have a waiting cost of exactly \( c \) is zero, we can assume that every other job has a strictly higher bid and thus a strictly higher priority. This means that a job is only run when there are fewer than \( l_s \) jobs of users with higher waiting cost in the system. To be run for one full unit of time the job with priority \( c \) thus on average needs to be in the system for \( \frac{1}{\mu \varphi(c, l_s, \zeta)} \) units of time. The statement of the Proposition now follows by noting that a job needs to run \( \frac{1}{\mu} \) units of time in expectation to be completed. \( \Box \)

Note that the probability \( \varphi(c, l_s, \zeta) \) depends on the arrival and service processes in the queue, which means we cannot give a closed form expression for \( \varphi \) for general settings. One notable exception are queues with Poisson arrival and exponential service processes, which are the basis for the numerical example.

The resulting queue is usually denoted \( \text{M/M/"number of instance"} \) in the literature. This queuing model is the most common and best understood and still quite powerful. For it exact formulas using only first moments are known, the most important of which is the Erlang C formula. For a given \( \text{M/M/1} \) queue with arrival rate \( \lambda \) and service time \( \frac{1}{\mu} \), the probability \( \varphi(l, \frac{\lambda}{\mu}) \) of finding an empty instance at any time is given by the well known Erlang C formula (see for example [Cooper, 1981]). A proof of the Erlang C formula can be found in [Takagi, 2008]):

\[
\varphi(l, \frac{\lambda}{\mu}) = 1 - \left(1 + \left(1 - \frac{\lambda}{l\mu}\right) \frac{l!}{\frac{\lambda}{\mu}} \sum_{k=0}^{l-1} \frac{\lambda^k}{k!}\right)^{-1}.
\]

This formula is independent of the service discipline, as long as it is work conserving, i.e. doesn’t leave an instance idle if jobs are waiting.

### A.1 Fixed-price Queue

For a fixed-price queue, the total waiting time \( \omega(F, c, \eta, \zeta) \) and payment \( m(F, c, \eta, \zeta) \) are already directly determined by the parameters. The Erlang C formula can now be used to calculate the number of instances \( l_F(\eta, \zeta) \) required to meet the queueing time limit \( T \). The expected queueing time \( q(l) \) of a user joining a randomly served queue with \( l \) instances is given by

\[
q = \frac{1 - \varphi(l, \frac{\lambda}{\mu})}{l\mu - \lambda}
\]

See [Cooper, 1981] for a proof. Plugging in the arrival rate into the fixed price market for given strategies \( \eta, \zeta \) and solving \( q(l_F(\eta, \zeta)) = T \) allows us to find \( l_F(\eta, \zeta) \).
A.2 Spot Queue

For the spot queue, the total waiting time \( w(S, c, \eta, \zeta) \) and payment \( m(S, c, \eta, \zeta) \) depend on the dynamics of the queue. The probability of finding an “empty” instance (in the sense that fewer than \( l_S \) users with a higher bid than \( c \) are currently in the queue) can be found by using the Erlang C formula and plugging in the arrival rate of jobs with higher bid. This follows from Buzen and Bondi [1983] showing that the probability of finding an empty instance does not depend on the service discipline and because users of bid \( c \) only have to wait for jobs of users with a higher priority.

The expected total waiting time \( w(S, c, \eta, \zeta) \) of a user joining the spot queue in any market is now given by Proposition A.1 as

\[
w(S, c, \eta, \zeta) = \frac{1}{\mu \varphi(c, l_S, \zeta)}.
\] (47)

Using the iterative approach described in the proof of Proposition 3.6 we can now calculate the cutoff vectors of the equilibrium strategies by solving a number of non-linear root searches. This allows us to in turn calculate payments and profits and search for the optimal price \( p \).

B PROOF OF LEMMA 3.3

First note again that for the job at the highest bid that joins the hybrid market, the spot queue is faster than the fixed-price queue as he will be run instantly instead of having to queue on average \( T > 0 \). Those jobs will therefore be willing to pay more than the fixed-price price and be willing to pay more than any job with a lower waiting cost. This means that they strictly prefer and joins the spot queue.

Let \( c^L \) be the lowest waiting cost for which a job prefers the fixed-price queue over the spot queue or is indifferent between the two and \( c^U \) the highest such waiting cost. \( c^L \) and \( c^U \) have to exist when the hybrid market is proper. Now we will show the statement of the Lemma by first showing that a contradiction follows from assuming that any job with waiting cost \( \bar{c} \in (c^L, c^U) \) strictly prefers the spot queue. Then we proof that for any such \( \bar{c} \) for which a job is indifferent between both queues, the spot queue has to be at least as expensive and at least as slow as the fixed-price queue. As this means the spot queue has the same price and queueing time as the fixed-price queue for bids between \( c^L \) and \( c^U \), no more than a null set can join it. It follows that the payment and time in the system of the spot queue are the same at \( c^L \) and \( c^U \), which yields the rest of the Lemma.

Consider a job with waiting cost \( \bar{c} \in (c^L, c^U) \) and assume that jobs with waiting cost \( \bar{c} \) strictly prefer the spot queue to the fixed-price queue. For a job at \( \bar{c} \) to prefer the spot queue, one of two things need to hold: either the spot queue for bid \( \bar{c} \) is cheaper or it is faster than the fixed-price queue. If both hold the contradiction would be trivial. If the spot is cheaper than the fixed-price
queue, i.e. if \( m(S, \bar{c}, \eta, \zeta) < m(\bar{F}, \bar{c}, \eta, \zeta) \), then the following holds:
\[
c^L w(S, c^L, \eta, \zeta) + m(S, c^L, \eta, \zeta) \leq c^L w(S, \bar{c}, \eta, \zeta) + w(S, \bar{c}, \eta, \zeta) \leq c^L \left( \bar{c} w(S, \bar{c}, \eta, \zeta) + \frac{\bar{c}}{c^L} m(S, \bar{c}, \eta, \zeta) \right) \]
\[
= \frac{c^L}{\bar{c}} \left( \bar{c} w(S, \bar{c}, \eta, \zeta) + m(S, \bar{c}, \eta, \zeta) \right) \]
\[
< \frac{c^L}{\bar{c}} \left( \bar{c} w(F, \bar{c}, \eta, \zeta) + m(F, \bar{c}, \eta, \zeta) \right) + (\frac{\bar{c}}{c^L} - 1) m(S, \bar{c}, \eta, \zeta) \]
\[
= \frac{c^L}{\bar{c}} \left( \bar{c} w(F, \bar{c}, \eta, \zeta) + m(F, \bar{c}, \eta, \zeta) \right) \]
\[
= c^L w(F, \bar{c}, \eta, \zeta) + m(F, \bar{c}, \eta, \zeta) \]
\]
\(48\) holds because the pricing rule is BNIC, \(50\) because the job with \( \bar{c} \) prefers the spot market and \(52\) because the spot queue is cheaper. The job with \( c^L \) would therefore also strictly prefer the spot queue, a contradiction.

If the spot queue is faster than the fixed-price queue, i.e. if \( w(S, \bar{c}, \eta, \zeta) < w(F, \bar{c}, \eta, \zeta) \), then the following holds
\[
c^U w(S, c^U, \eta, \zeta) + m(S, c^U, \eta, \zeta) \leq c^U w(S, \bar{c}, \eta, \zeta) + m(S, \bar{c}, \eta, \zeta) \leq c^U \left( \bar{c} w(S, \bar{c}, \eta, \zeta) + m(S, \bar{c}, \eta, \zeta) \right) \]
\[
= \bar{c} w(S, \bar{c}, \eta, \zeta) + m(S, \bar{c}, \eta, \zeta) \]
\[
< \bar{c} w(F, \bar{c}, \eta, \zeta) + m(F, \bar{c}, \eta, \zeta) \]
\[
= \bar{c} w(F, \bar{c}, \eta, \zeta) + m(F, \bar{c}, \eta, \zeta) \]
\[
= c^U w(F, \bar{c}, \eta, \zeta) + m(F, \bar{c}, \eta, \zeta) \]
\]
\(54\) holds again because the pricing rule is BNIC, \(56\) because the job with \( \bar{c} \) prefers the spot market and \(57\) because the spot queue is faster for bid \( \bar{c} \). The job with \( c^U \) would therefore also strictly prefer the spot queue, a contradiction. \( \bar{c} \) therefore either prefers the fixed-price queue or is indifferent between fixed-price and spot queue.

With similar calculation we can also show:

**Lemma B.1.**

1. \( m(S, c^L, \eta, \zeta) \geq m(F, c^L, \eta, \zeta) \), i.e. at the lowest waiting cost for which a job prefers the fixed-price queue over the spot queue or is indifferent between the two, the spot queue can not be more expensive than the fixed-price queue, as there exist jobs with higher waiting cost who strictly prefer the spot queue.

2. \( w(S, c^U, \eta, \zeta) \geq w(F, c^U, \eta, \zeta) \) i.e. at the highest waiting cost where a job prefers the fixed-price queue over the spot queue or is indifferent between the two, the spot queue can not be faster than the fixed-price queue as a job with waiting cost zero always strictly prefers the spot market for \( p > 0 \).

This means that any \( \bar{c} \in (c^L, c^U) \) that joins the spot queue is indifferent between it and the fixed-price queue.
For such \( \tilde{c} \) it thus holds
\[
c^L w(S, c^L, \eta, \zeta) + m(S, c^L, \eta, \zeta) \geq c^L w(F, \tilde{c}, \eta, \zeta) + m(F, \tilde{c}, \eta, \zeta)
\]
(58)
\[
= \frac{c^L}{\tilde{c}} \left[ \tilde{c} w(F, \tilde{c}, \eta, \zeta) + m(F, \tilde{c}, \eta, \zeta) \right]
+ (1 - \frac{c^L}{\tilde{c}}) m(F, \tilde{c}, \eta, \zeta)
\]
(59)
\[
= \frac{c^L}{\tilde{c}} \left[ \tilde{c} w(S, \tilde{c}, \eta, \zeta) + m(S, \tilde{c}, \eta, \zeta) \right]
+ (1 - \frac{c^L}{\tilde{c}}) m(F, \tilde{c}, \eta, \zeta)
\]
(60)
\[
= \left[ c^L w(S, \tilde{c}, \eta, \zeta) + m(S, \tilde{c}, \eta, \zeta) \right]
+ (1 - \frac{c^L}{\tilde{c}}) (m(F, \tilde{c}, \eta, \zeta) - m(S, \tilde{c}, \eta, \zeta))
\]
(61)
and because the payment rule is BNIC it has to hold
\[
(1 - \frac{c^L}{\tilde{c}}) (m(F, \tilde{c}, \eta, \zeta) - m(S, \tilde{c}, \eta, \zeta)) \leq 0
\]
(62)
and therefore
\[
m(F, \tilde{c}, \eta, \zeta) \leq m(S, \tilde{c}, \eta, \zeta)
\]
(63)
This means that for any job that joins the spot queue between \( c^L \) and \( c^U \) the fixed-price price has to be lower or equal than the price they would have to pay in the spot queue.

On the other hand:
\[
w(S, c^U, \eta, \zeta) + m(S, c^U, \eta, \zeta) \geq c^U w(F, \tilde{c}, \eta, \zeta) + m(F, \tilde{c}, \eta, \zeta)
\]
(64)
\[
= \left[ \tilde{c} w(F, \tilde{c}, \eta, \zeta) + m(F, \tilde{c}, \eta, \zeta) \right]
+ (c^U - \tilde{c}) w(F, \tilde{c}, \eta, \zeta)
\]
(65)
\[
= \left[ \tilde{c} w(S, \tilde{c}, \eta, \zeta) + m(S, \tilde{c}, \eta, \zeta) \right]
+ (c^U - \tilde{c}) w(F, \tilde{c}, \eta, \zeta)
\]
(66)
\[
= \left[ c^U w(S, \tilde{c}, \eta, \zeta) + m(S, \tilde{c}, \eta, \zeta) \right]
+ (c^U - \tilde{c}) (w(F, \tilde{c}, \eta, \zeta) - w(S, \tilde{c}, \eta, \zeta))
\]
(67)

Because the payment rule is BNIC it has to hold
\[
(c^U - \tilde{c}) (w(F, \tilde{c}, \eta, \zeta) - w(S, \tilde{c}, \eta, \zeta)) \leq 0
\]
(68)
and thus
\[
w(F, \tilde{c}, \eta, \zeta) \leq w(S, \tilde{c}, \eta, \zeta)
\]
(69)
This means that for any job that joins the spot queue between \( c^L \) and \( c^U \) the fixed-price queue has to be at least as fast as the spot queue.

With (63) and (69), in order to be indifferent between the queues it has to hold
\[
m(F, \tilde{c}, \eta, \zeta) = m(S, \tilde{c}, \eta, \zeta)
\]
(70)
\[
w(F, \tilde{c}, \eta, \zeta) = w(S, \tilde{c}, \eta, \zeta)
\]
(71)

This means that for any \( \tilde{c} \in (c^L, c^U) \) for which a job could potentially join the spot queue in the BNE, the total time in the system and the total payment in the spot queue are independent of \( \tilde{c} \) and
equal to joining the fixed-price queue. Therefore, at most a null set of these jobs can join the spot queue, which does not affect the system dynamics. For all calculations it can thus be assumed no job between \(c^L\) and \(c^U\) joins the spot queue.

This also means that the time in the system and the payment between \(c^L\) and \(c^U\) don’t change and with Lemma B.1 we get:

\[
m(S, c^L, \eta, \zeta) = m(S, c^U, \eta, \zeta) = \frac{p}{\mu} \quad (72)
\]

\[
w(S, c^L, \eta, \zeta)) = w(S, c^U, \eta, \zeta) = T + \frac{1}{\mu} \quad (73)
\]

C PROOF OF PROPOSITION 3.4

That equations 18, 19 and 20 have to hold immediately follows from Lemma 3.3:

1. Equation (18): The payoff at \(c^L\) has to be the same for joining the fixed-price and spot market
2. Equation (19): The payoff at \(c^U\) also has to be the same in fixed-price and spot market.
3. Equations (20): Users join the market as long as their value for doing so is greater than 0.

We will now show that this system of equations always has a unique solution using a constructive approach. For this we first need some more notation:

Given provider strategy \(\eta\), we know that in order to satisfy Lemma 3.3 jobs that pay more in the spot market than in the fixed-price marked need to arrive at a rate that jobs with waiting cost \(c\) have to wait exactly \(T\). Denote this arrival rate by \(\lambda(T, \eta)\). With some slight abuse of notation, let \(\zeta = (\hat{c}, \lambda(T, \eta))\) for any vector \(\hat{c}\) denote the aggregate strategy where every user of class \(i\) with waiting cost \(c < \hat{c}_i\) joins the spot market, while everyone else doesn’t join at all, but instead dummy jobs of maximal priority arrive with rate \(\lambda(T, \eta)\). This means that \(w(S, c, \eta, (\hat{c}, \lambda(T, \eta))) = T\) by definition.

Combined with Lemma 3.3, this notational trick allows us to simulate the arrival rate of users with waiting costs between \(c^U\) and \(c^H\) without yet knowing all cutoff vectors.

Now note that Equation 18 only depends on \(c^L\) and \(c^H\) for the classes of jobs that do not join the fixed-price queue. We now need to find out which classes join the fixed-price market and which do not. Once that is known, we can split the system of equations into two part that can be solved consecutively.

To check if the \(k\)’th class joins the fixed-price queue, i.e. if \(c^L_k \leq c^H_k\), denote by \(\hat{c} \in [0, v(1)]^n\) the cutoff vector solving

\[
0 = v(i) - \int_0^{\hat{c}_i} w(S, x, \eta, (\hat{c}, \lambda(T, \eta)))dx \quad \forall i \geq k \quad (74)
\]

\[
\hat{c}_i = \hat{c}_k \quad \forall i < k \quad (75)
\]

This has an unique solution according to Lemma 6 of Abhishek et al. [2012]. When

\[
\frac{p}{\mu} > m(S, \hat{c}_k, \eta, (\hat{c}, \lambda(T, \eta))) \quad (76)
\]

\[
= \int_0^{\hat{c}_k} w(S, x, \eta, (\hat{c}, \lambda(T, \eta)))dx - w(S, \hat{c}_k, \eta, (\hat{c}, \lambda(T, \eta))) \quad (77)
\]

holds, then a \(c^L \leq c^H_k\) would mean not enough people join the spot market to reach the fixed-price price at the cutoff point \(c^L\) and equation (18) can’t hold if class \(k\) does not join the fixed-price market. Thus class \(k\) does join the fixed-price queue in every equilibrium strategy.
If it does not hold, then setting $c^L \leq c^H_{k+1}$ would mean that (more than) enough people join the spot market to reach an expected payment equal to the fixed-price price at the cutoff point $c^L$. Thus class $k$ does not join the fixed-price queue in any equilibrium strategy. As $m(S, \hat{c}_k, \eta, (\hat{c}, \lambda(T, \eta)))$ is monotone in $k$, there exists a smallest class $k^*$ that does not join the fixed-price market.

Splitting the system of equations for the equilibrium strategy at this $k^*$, Lemma 6 of Abhishek et al. [2012] yields that

$$0 = c^L(T + \frac{1}{\lambda}) + p\frac{1}{\mu} - \int_0^{c^L} w(S, x, \eta, (\hat{c}^L, \lambda(T, \eta)))dt$$  \hspace{1cm} (78)

$$0 = v(i) - \int_0^{\hat{c}^H_i c^H} w(t, \hat{c}^L, \lambda(T, \eta))dt \hspace{1cm} \forall i \geq k^*$$  \hspace{1cm} (79)

does have a unique solution since $\hat{c}^H_i \leq c^L$.

Given the solution to 78 and 79, we can now equivalently find the smallest class $k^{**}$ that joins the upper area of the spot market (i.e. for which $c^U_{k^{**}} < c^H_{k^{**}}$) by setting $\hat{c}^H_i = c^L$ for all $k^* < i < k$ and solving

$$0 = v(k) - c^U(T + \frac{1}{\mu}) + p\frac{1}{\mu}$$  \hspace{1cm} (80)

$$0 = v(i) - \int_0^{c^H} w(S, x, \eta, (\hat{c}^L, \hat{c}^U, \hat{c}^H))dx \hspace{1cm} \forall i < k^{**}$$  \hspace{1cm} (81)

which has a unique solution according to Lemma 6 of Abhishek et al. [2012] for every $k$.

Equivalently to before, there is a highest class $k^{**} < k^*$ for which

$$T + \frac{1}{\mu} > w(S, \hat{c}_k, \eta, (\hat{c}^L, \hat{c}^U, \hat{c}^H))$$  \hspace{1cm} (82)

$$\hat{c}^H_{k^{**} + 1} \leq c^L < \hat{c}^H_{k^{**}}, \text{ and therefore } v(k^{**} + 1) \leq c^U(T + \frac{1}{\mu}) + p\frac{1}{\mu} < v(k^{**}), \text{ Lemma 6 of Abhishek et al. [2012] yields that}$$

$$0 = c^U(T + \frac{1}{\mu}) + p\frac{1}{\mu} - \int_0^{c^U} w(S, x, \eta, (\hat{c}^L, \hat{c}^U, \hat{c}^H))dx$$  \hspace{1cm} (84)

$$0 = v(i) - \int_0^{\hat{c}^H_i} w(S, x, \eta, (\hat{c}^L, \hat{c}^U, \hat{c}^H))dx \hspace{1cm} \forall i < k^* \hat{c}^H_i = c^L$$  \hspace{1cm} (85)

has a unique solution.

Equation (85) for the classes $i \in (k^*, k^{**})$ then trivially has a unique solution. As each subsystem of equations was, at the time it was solved, independent of any as of then unsolved part, $c^L$, $c^U$ and $\hat{c}^H$ solve the whole system of equations.
D PROOF OF PROPOSITION 3.6

If $l_S = 0$ $\zeta^* = \bar{c}^F$ follows from proposition 3.1. For $l_S > 0$ at least users with waiting times in some neighbourhood around zero will prefer the spot market.

Let $l_S > 0$. If no $\bar{c}$ for which the expected waiting time in both queues is equal, the spot market is trivially faster, and consequently cheaper, for every user. No user would choose to join the fixed-price queue in the hybrid market and by Theorem 3.2 $\zeta^* = \bar{c}^S$.

Let such $\bar{c}$ exists and condition 24 hold, but assume we have an equilibrium where no one joins the fixed-price queue of the hybrid market, i.e. where the hybrid market degenerates to the spot market. A user of class 1 (i.e. the class with maximal value for completion) with waiting cost $\bar{c}$ could then obtain a better payoff by switching to the fixed-price queue, leading to a contradiction.

Any BNE therefore has some users joining the fixed-price market and by Lemma 3.3 it follows $\zeta^* = (\bar{c}^L, \bar{c}^U, \bar{c}^H)$.

Conversely let condition 24 not hold:

$$\bar{c}(T + \frac{1}{\mu}) + \frac{1}{\mu} \geq \int_0^{\bar{c}} w(S, x, \eta, \zeta)dx$$

for $\bar{c}$ such that $T + \frac{1}{\mu} = w(S, \bar{c}, \eta, \zeta)$. It follows that

$$\frac{1}{\mu} \geq \int_0^{\bar{c}} w(S, x, \eta, \zeta)dx - \bar{c}w(S, \bar{c}, \eta, \zeta)$$

This means that even at the bid $\bar{c}$ where the fixed-price market has the same total waiting time as the spot market, joining the spot market is still cheaper. We now show that there exist only BNEs where no user joins the fixed-price queue. Since the payoff in the spot queue for every user is monotone decreasing in the number of users that join, it is enough to show that when playing $\zeta = \bar{c}^S$ no user has any incentive to switch to the fixed-price queue.

A user with waiting cost $\bar{c}$ clearly has no reason to switch. Assume a user of type $i$ with waiting cost $c \neq \bar{c}$ would prefer to change to the fixed-price market. Misreporting his type as $\bar{c}$ and joining the spot queue would then lead to a payoff of

$$\pi^c_i(S, \bar{c}, \eta, \bar{c}^S) = v_i - \int_0^{\bar{c}} w(S, x, \eta, \zeta)dx + \bar{c}w(S, \bar{c}, \eta, \zeta) - cw(S, c, \eta, \zeta)$$

$$= v_i - \int_0^{\bar{c}} w(S, x, \eta, \zeta)dx + \bar{c}w(S, \bar{c}, \eta, \zeta) - c(T + \frac{1}{\mu})$$

$$\geq v_i - \frac{1}{\mu} - c(T + \frac{1}{\mu})$$

$$= \pi^c_i(F, c, \eta, \bar{c}^S)$$

$$> \pi^c_i(S, c, \eta, \bar{c}^S)$$

Misreporting in the spot queue would therefore be beneficial over reporting truthfully, in contradiction to the pricing rule being Bayes-Nash incentive compatible. Consequently, no user prefers the spot market and by Theorem 3.2 $\zeta^* = \bar{c}^S$.

E PROOF OF THEOREM 3.7

This proof requires the introduction of an additional technical Lemma and some more notation.
Ludwig Dierks and Sven Seuken

Lemma E.1. For every setting, \( p < \mu v_1 \) and \( \epsilon > 0 \), there exists a (possibly fractional) number of spot instances \( l_S \) such that for \( \eta = (p, l_S) \) the resulting equilibrium strategy is \( \zeta^* = (\bar{c}_L, \bar{c}_U, \bar{c}_H) \) and it holds that users which join the spot market on average have a expected payment

\[
\bar{m}(S, \eta, \zeta^*) = \frac{\sum_i \lambda_i \left[ \int_0^{\bar{c}_L^i} m(S, y, \eta, \zeta^*) f_i(y) dy + \int_{\bar{c}_L^i}^{\bar{c}_U^i} m(S, y, \eta, \zeta^*) f_i(y) dy \right]}{\sum_i \lambda_i \left[ \int_0^{\bar{c}_L^i} f_i(y) dy + \int_{\bar{c}_L^i}^{\bar{c}_U^i} f_i(y) dy \right]}
\]

(93)

that is greater than the payment at \( \bar{c}_L^i \) highest bidding job pays minus \( \epsilon \) and thus, for \( \epsilon \) small enough, also greater than the average payment in the fixed-price market, i.e.

\[
\bar{m}(S, \eta, \zeta^*) \geq m(S, \bar{c}_L^i, \eta, \zeta^*) - \epsilon
\]

(95)

\[
> \frac{p}{p}
\]

(96)

Proof. To keep the proof readable, we need to introduce notation in order to draw all waiting costs in the spot market from a single distribution, instead of first drawing a jobs class and then its waiting cost. Note that this does not in anyway change the number of jobs or bids in the market. For provider strategy \( \eta \) and equilibrium strategy \( \zeta^* = (\bar{c}_L, \bar{c}_U, \bar{c}_H) \) define the distribution

\[
F_S(c) = \frac{\sum_i \lambda_i \left[ \int_0^{\min\{c, \bar{c}_L^i\}} f_i(y) dy + \int_{\min\{c, \bar{c}_L^i\}}^{\bar{c}_U^i} f_i(y) dy \right]}{\sum_i \lambda_i \left[ \int_0^{\bar{c}_L^i} f_i(y) dy + \int_{\bar{c}_L^i}^{\bar{c}_U^i} f_i(y) dy \right]}
\]

(97)

and the arrival rate

\[
\lambda_S = \sum_i \lambda_i \left[ \int_0^{\bar{c}_L^i} f_i(y) dy + \int_{\bar{c}_L^i}^{\bar{c}_U^i} f_i(y) dy \right]
\]

(98)

(99)

with similarly constructed PDF \( f_S(c) \). Now consider an artificial spot market with arrival rate \( \lambda_S \), where every arriving jobs waiting cost is drawn from \( F_S \) and everyone joins. This market, from the providers view, is the same as the normal spot market that would result from him playing \( \eta \), including users on average having the same expected payments. In order to analyze the providers profit from the spot market when playing \( \eta \) we therefore can instead analyze this artificial market.

The per-user-average profit \( \bar{m}(S, \eta, \zeta^*) \) of the artificial spot market is given by the expected value the expected payment \( m(S, \bar{c}, \eta, \zeta^*) \) takes when \( \bar{c} \) is drawn at random from the PDF \( f_S(c) \):

\[
\bar{m}(S, \eta, \zeta^*) = \frac{\lambda_S \left[ \int_0^{\bar{c}_L^i} m(S, y, \eta, \zeta^*) f_S(y) dy \right]}{\lambda_S}
\]

(100)

\[
= \int_0^{\bar{c}_L^i} m(S, y, \eta, \zeta^*) f_S(y) dy
\]

(101)

\[
= \int_{\bar{c}_L^i}^{\bar{c}_U^i} m(S, y, \eta, \zeta^*) f_S(y) dy
\]

(102)

\[
= \int_{\infty}^{\bar{c}_U^i} m(S, y, \eta, \zeta^*) f_S(y) dy
\]

(103)

\[
= E_{c, f_S}(m(S, \bar{c}, \eta, \zeta^*))
\]

(104)
Now, for any $l_S$ and any $0 < \xi < 1$ define $c^{l_S}_\xi$ as the waiting cost with $F_S(c^{l_S}_\xi) = \xi$. It then follows by Markov’s inequality

\begin{align}
\bar{m}(S, \eta, \zeta^*) &\geq (1 - F_S(c^{l_S}_\xi))m(S, c^{l_S}_\xi, \eta, \zeta^*) \\
&= (1 - \xi)m(S, c^{l_S}_\xi, \eta, \zeta^*)
\end{align}

Further it holds

\begin{align}
\frac{p}{\mu} &= \int_0^{c_L} w(S, x, \eta, \zeta^*)dx - c_L T \\
&\leq c_L \max_{c \in [0, c_L]} w(S, c, \eta, \zeta^*) - c_L T \\
&= c_L w(S, 0, \eta, \zeta^*) - c_L T \\
&\leq c_L w(S, 0, \eta, \zeta^*)
\end{align}

by the integral upper bound as the total waiting time is monotone increasing. As

\begin{equation}
c_L \xrightarrow{l_S \rightarrow 0} 0
\end{equation}

it holds

\begin{equation}
w(S, 0, \eta, \zeta^*) \xrightarrow{l_S \rightarrow 0} \infty
\end{equation}

As a job with bid 0 is served exactly when there is an empty instance in the queue consisting of all other jobs that join, the instance utilization of the spot queue has to go to full, i.e.

\begin{equation}
\frac{\lambda_S}{l_S \mu} \xrightarrow{l \rightarrow 0} 1
\end{equation}

Now fix some $\xi > 0$. It holds

\begin{equation}
\frac{(1 - F_S(c^{l_S}_\xi))\lambda_S}{l_S \mu} \xrightarrow{l_S \rightarrow 0} (1 - \xi)1
\end{equation}

i.e. going towards zero, the utilization by jobs in the queue with priority over $c^{l_S}_\xi$ will always at most be an average of $(1 - \xi)$. This limits the total waiting time at $c^{l_S}_\xi$ to some arbitrary high, but
finite value \( \bar{\omega}_\xi \) depending on \( \xi \) but independent of \( l_S \). Therefore

\[
m(S, c_1^H, \eta, \xi^*) = \int_0^{c_1^I} w(S, x, \eta, \xi^*) dx + \int_{c_1^I}^{c_1^H} w(S, x, \eta, \xi^*) dx - c_1^H w(S, c_1^H, \eta, \xi^*) \leq \int_0^{c_1^I} w(S, x, \eta, \xi^*) dx + (c_1^H - c_1^I) w(S, c_1^I, \eta, \xi^*) - c_1^H w(S, c_1^H, \eta, \xi^*) \leq \int_0^{c_1^I} w(S, x, \eta, \xi^*) dx - c_1^I w(S, c_1^I, \eta, \xi^*) + c_1^H (\bar{\omega}_\xi - w(S, c_1^H, \eta, \xi^*)) = m(S, c_1^I, \eta, \xi^*) + c_1^H (\bar{\omega}_\xi - w(S, c_1^H, \eta, \xi^*))
\]  

(115)  

(116)  

(117)  

(118)  

(119)

As \( c_1^H \to 0 \) and \( m(S, c_1^H, \eta, \xi^*) \geq m(S, c_1^I, \eta, \xi^*) \), it follows

\[
|m(S, c_1^H, \eta, \xi^*) - m(S, c_1^I, \eta, \xi^*)| \to 0
\]  

(120)  

For all \( 0 < \xi < 1, \delta > 0 \) therefore exists a \( l_S \) with \( m(S, c_1^I, \eta, \xi^*) \geq (1 - \delta) m(S, c_1^H, \eta, \xi^*) \) and therefore

\[
\bar{m}(S, \eta, \xi^*) \geq (1 - \xi) m(S, c_1^I, \eta, \xi^*) \geq (1 - \xi)(m(S, c_1^H, \eta, \xi^*) - \delta)
\]  

(121)  

(122)

Choosing \( \xi \) and \( \delta \) such that \( \epsilon > \xi m(S, c_1^I, \eta, \xi^*) - (1 - \xi) \delta \) then gives us the statement of the lemma. \( \square \)

With this technical Lemma, proofing Theorem 3.7 is now rather easy. For any price \( p_0 \) set \( l_S \) such that \( \bar{m}(S, \eta, \xi^*) \geq \frac{p_0}{\mu} \). Such \( l_S \) always exists by Lemma E.1. For any provider strategy \( \eta \) and equilibrium strategy \( \bar{\zeta}^* = (\bar{c}^L, \bar{c}^U, \bar{c}^H) \) let

\[
\lambda_S(\eta, \xi^*) = \sum_i \lambda_i \left( \int_0^{c_1^I} f_i(y) dy + \int_{c_1^I}^{c_1^H} f_i(y) dy \right)
\]  

(123)  

(124)

denote the arrival rate into the spot market.

Then the overall change in profit between playing \( \eta_0 = (p, 0) \) (offering no spot instances) and \( \eta = (p, l_S) \) (offering \( l_S \) spot instances) is positive:
We show Theorem 4.3 whenever there exists no η

As users with waiting cost small enough (what is small enough depends on l

provider therefore has to spend at least κδ

l

δ

than large spot market. It follows that instances offered.

λ

the arrival rate of jobs joining the spot market instead of the fixed-price market

strategy and obtain an increase in payoff,

η

fixed-price queue.

offered. 129 follows because the number of instances can only go down when less people join the fixed-price queue.

Since the price if the fixed-price market didn’t change, every user can at least get the same payoff for provider strategy η as he could for η₀:

$$\pi^c_i(F, c, η, \xi^*) = \pi^c_i(F, c, η_0, \xi^*)$$

As users with waiting cost small enough (what is small enough depends on l₅) change to the spot market and obtain an increase in payoff, η results in a Pareto improvement for the users over η₀.

F CONTINUATION OF THE PROOF OF 4.3

We show Theorem 4.3 whenever there exists no l₅ with tₑ(l₅)τ < T.

Denote by ∆l₅(l₅) = l₅((p₀, 0), ξ*) - l₅((p₀, l₅), ξ*) the reduction in fixed-price instances between strategy η₀ = (p₀, 0) and any strategy η = (p₀, l₅) that sets the same price.

Further denote δ(l₅) = ∆l₅(l₅)/λₛ(l₅) as the ratio between the reduction in fixed-price instances and the arrival rate of jobs joining the spot market instead of the fixed-price market λₛ(l₅) with l₅ spot instances offered.

By the law of large numbers, the variance of the number of arrivals for any period of time is decreasing in the the arrival rate. This means that the more users join the fixed-price market, the fewer instances per expected arrival are required to ensure the queueing time limit T. This means that per spot instance offered, small spot markets lead to a larger reduction in fixed-price instances than large spot market. It follows that δ(l₅) is decreasing in l₅.

Now fix an arbitrary number of spot instances l₅ ≤ l such that the hybrid market is proper. It holds for all l₅ ≤ l₅ that δ(l₅) ≥ δ(l₅) and τₑ(l₅) ≤ tₑ(l₅). When playing η = p₀, l₅ with l₅ ≤ l₅ the provider therefore has to spend at least κδ(l₅)λₛ(l₅) less on fixed-price instances.

Conversely, users of class one at the cutoff waiting cost (i.e. with waiting cost c₁) will be paying p₀ - c₁(tₑ(l₅)τ - T). Equivalently to Lemma E.1 we can show that for every ε > 0, there are l₅
such that the average payment of the users that join the spot queue is
\[
\bar{m}(S, (p_0, l_s), \xi^*) \geq p_0 - c^1_1 (\nu_E(l_s) \tau - T) - \epsilon. \quad (132)
\]

For such \(l_s\), the profit difference between the hybrid and the fixed-price market can therefore be written as
\[
\Pi((p_0, l_s), \xi^*) - \Pi((p_0, 0), \xi^*) = \kappa \delta(l_s) \lambda_S(l_s) - \lambda_S(l_s)(\bar{m}(S, (p_0, l_s), \xi^*) - p_0) \quad (133)
\]
\[
\geq \kappa \delta(l_s) \lambda_S(l_s) - \lambda_S(l_s)(c^1_1 (\nu_E(l_s) \tau - T) - \epsilon) \quad (134)
\]
\[
= \lambda_S(l_s)(\kappa \delta(l_s) - (c^1_1 (\nu_E(l_s) \tau - T) - \epsilon)) \quad (135)
\]

Since \(\kappa \Delta l_F(l_s)\) and \(\nu_E(l_s) \tau - T\) are constants and \(c^1_1 \xrightarrow{l_s \to 0} 0\), choosing \(\epsilon\) and \(l_s\) small enough yields a profit increase for strategy \(\eta = (p_0, l_s)\) over \(\eta_0 = (p_0, 0)\). Since the price didn’t change, users also again obtain an Pareto improvement.