Cloud Pricing: The Spot Market Strikes Back*

Working Paper, July 2020

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

Cloud computing providers must constantly hold many idle compute instances available (e.g., for maintenance, or for users with long-term contracts). A natural idea, which should intuitively increase the provider's profit, is to sell these idle instances on a secondary market, for example, via a preemptible spot market. However, this ignores possible "market cannibalization" effects that may occur in equilibrium as well as the additional costs the provider experiences due to preemptions. To study the viability of offering a spot market, we model the provider's profit optimization problem by combining queuing theory and game theory to analyze the equilibria of the resulting queuing system. Our main result is an easy-to-check condition under which a provider can simultaneously achieve a profit increase and create a Pareto improvement for the users by offering a spot market (using idle resources) alongside a fixed-price market. Finally, we illustrate our results numerically to demonstrate the effects the provider's costs and her strategy have on her profit.

1 INTRODUCTION

Providers of cloud services like Amazon EC2 or Microsoft Azure rent out computing capacity to millions of users. These cloud services generate billions of dollars in yearly revenue and the market for these services is still growing exponentially.[1] While some users enter into year-long contracts, many prefer to obtain resources on-demand (i.e., just when they need them and without long-term commitments). Today, on-demand compute resources (i.e., CPU, RAM, bandwidth, etc.) are most commonly combined into compute instances (e.g., virtual machines) and rented out via *fixed-price* markets. In these markets, users pay a fixed price per time unit and the provider aims to keep enough instances available to be able to almost instantly satisfy all requests. This approach is considered to be simple, reliable, and to satisfy the requirements of most users; it is therefore widely used in practice.[2] At the same time, all cloud computing centers always contain many *idle* instances: for example, to guarantee the 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). In effect, a sizable number of instances do nothing at any given time. However, low utilization rates are inherently undesirable, because most of the overall per-instance costs are independent of utilization (Barroso et al. 2018). A lot of research has been done to increase these utilization rates, e.g., by allowing advance reservation of resources (Azar et al. 2015, Babaioff et al. 2017), by predicting future usage (Cohen et al. 2019, Cortez et al. 2017, Dierks et al. 2019, Jyothi et al. 2016) or by incentivizing users to reduce the variance of their demand (Dierks and Seuken 2020). However, any cloud computing center will always have a non-negligible number of idle instances.

Naturally, a cloud provider may also want to sell those idle instances; but unfortunately, they are only "temporarily" idle. Whenever an instance is needed for its primary purpose (e.g., for maintenance or for a user with a long-term contract), then the job that is currently running on the instance must be *preempted*, i.e., the job is shut down. Of course, getting preempted and restarting a job on a different instance is not lossless for a user; for example, because a new instance may not be immediately available, because unsaved progress to completion may be lost, or because it may require some time to reconfigure the new instance. For this reason, idle instances cannot be

*Some of the ideas presented in this paper were also described in a one-page abstract that was published in the conference proceedings of EC'19 (Dierks and Seuken 2019).

[1] https://www.microsoft.com/en-us/Investor/earnings/FY-2018-Q2/press-release-webcast

[2] For example: https://aws.amazon.com/ec2/pricing/on-demand/ or https://azure.microsoft.com/en-us/offers/pay-as-you-go/

sold on the regular fixed-price market where users are guaranteed the continued (non-preemtible) usage of their resources. However, researchers (e.g., Abhishek et al. (2017) and Hoy et al. (2016)) as well as cloud providers (e.g., *Amazon EC2 Spot instances* and *Microsoft Azure Low Priority VMs*)[3] have considered to instead sell the idle instances on a secondary (cheaper) market where the users know that their jobs may be preempted. A user would use this market with the understanding that, when his job is shut down, he can restart the job once another instance becomes available.[4]

Given that the supply of idle resources changes over time, "dynamic pricing" is a natural choice for the secondary market.[5] Since in any secondary market consisting of idle instances the users have to cope with preemptions by the nature of the instances, additional preemptions caused by dynamic pricing do not qualitatively change the user experience. We focus on dynamic pricing that is implemented as a *preemptible spot market*. This means that users bid for resources and are served whenever the current market price is lower than their bid. If the market price rises too far while their job is running, they are preempted until the price drops again. Note that the idea of using preemptible spot markets for unused resources is not new: similar spot markets have also recently been proposed for the power capacity in multi-tenant data centers (Islam et al. 2018).

At first sight, it may seem obvious that the provider should offer idle instances on such a secondary (spot) market — after all, these instances seem "free" for the provider and selling them (at essentially any price) generates revenue. However, this thinking ignores two important points. First, putting otherwise idle instances under load causes additional load-dependent costs for the provider, which can be much larger in a *preemptible* secondary market than in a *non-preemptible* market. Second, it ignores possible "market cannibalization" effects that may occur in equilibrium, i.e., that users may choose to move from the more expensive fixed-price market to the cheaper spot market. Indeed, Abhishek et al. (2017) have shown that certain cannibalization effects can occur, at least in terms of revenue. Intuitively, market cannibalization becomes particularly problematic when the preemption costs on the spot market are large (i.e., when the users get preempted often and need to re-run large parts of their jobs), because then most users are only willing to pay very little for joining the spot market.

In this paper, we ask the following research question: *when can a cloud provider offer a spot market in addition to a fixed-price market to increase her profit?* Note that a provider trying to maximize her profit faces two basic questions: (1) what *price* should she ask for instances in the fixed-price market, and (2) how many (possibly zero) spot instances should she offer? Her profit then depends on the users' actions, given the offered markets. To answer our research question, we combine queuing theory and game theory to analyze the equilibrium behavior of the users. We model the two different markets as distinct *queues* that arriving users can choose from, a modeling framework well studied for classical service systems (Banerjee et al. 2015, Hassin 2016, Hassin and Haviv 2003) and previously applied to cloud computing (Abhishek et al. 2012, 2017, Gao et al. 2019). While such a queueing-theoretic approach is not the only viable approach to model cloud markets (Hoy et al. 2016, Kash and Key 2016, Zhang et al. 2016), it is particularly well suited to model the temporal nature of users continuously arriving and departing.

In contrast to prior work, we present a significantly more realistic model for the cloud domain that also captures all relevant costs of the provider and the users (Section 3), enabling us to perform

[3]See https://aws.amazon.com/ec2/spot/ and https://azure.microsoft.com/en-us/pricing/details/batch/. Note that, Agmon Ben-Yehuda et al. (2013) collected evidence, suggesting that, at some point, the EC2 spot market (despite its name) may not actually have used real spot pricing.

[4]To distinguish the provider from the users, we use "she/her" when referring to the provider and "he/his" for a user.

[5]Note that there are alternative pricing mechanisms for the secondary market that are also plausible. For example, the secondary market could also use fixed prices and this would likely increase the provider's profit as well. Future work should compare the two alternatives; but before we can do so, we first need to analyze the spot market case.

a profit analysis. First, we assume that the provider incurs *fixed costs* for each instance in the fixed-price market and therefore only offers a finite number of fixed-price instances. For the spot market, we assume that the provider has a finite number of instances that she can offer without incurring any fixed costs (since she uses existing idle instances). We assume that the provider takes those spot instances from a pool of idle resources that is distinct from the fixed-price instances (e.g., long-term reserved instances, maintenance capacity, etc.). For most large cloud providers, a sizable pool of such instances exists (Yan et al. 2016), and these instances are more "reliably idle" than most instances on the fixed-price market.[6] We assume that, for any instance (in the fixed-price or spot market), the provider incurs *load-dependent costs* whenever a job is running on it. Furthermore, we assume that a user has costs for waiting until his job is completed. Thus, his payoff consists of his value for job completion minus his incurred waiting cost and payment. Importantly, our model also includes *preemption costs* that a user incurs when his job is preempted. This takes the form of additional time required for the user's computation to again reach the point at which his job was preempted.

On the way towards answering our research question, we first provide a full characterization of the resulting user equilibria depending on the provider's strategy (Section 4). Our main result is a condition (see Definition 5.3 in Section 5.2) under which any provider who only offers a fixed-price market can increase her profit and simultaneously create a Pareto improvement for the users by also offering some of her idle instances on a spot market (Sections 5.1 and 5.2).[7]

In practice, our results can provide managerial guidance for cloud providers, in particular because our condition is easy to check (Section 5.3). Our condition is also mild enough to be satisfied for most cloud providers. Furthermore, we discuss how a cloud provider may increase her profit even if she is unable to compute her optimal strategy.

To illustrate our theoretical results, we numerically calculate equilibria for multiple examples where users arrive according to a Poisson process and require exponentially distributed service times (Section 6). We use these examples to study the effect that different cost structures have on the profitability of offering a spot market, and how the provider's strategy impacts her profit. In particular, we illustrate our main result, i.e., how a profit increase can be combined with a Pareto improvement for the users. Further, we show that even under relatively pessimistic conditions for a spot market, sizeable profit increases can still be attainable.

2 RELATED WORK

Offering a spot market is a form of *product differentiation* (Desai 2001, Maskin and Riley 1984, Shaked and Sutton 1982), where a provider offers *differentiated* products to appeal to different users. This contrasts with standard *price discrimination*, where typically identical or very similar products are sold at varying prices (Varian 1989). In their seminal paper on product differentiation, (Mussa and Rosen 1978) first formalized the relationship between prices and a user's obtained quality level, similar to results (Myerson 1981) later obtained for optimal auctions. For posted price mechanisms, (Mussa and Rosen 1978) further showed that optimal pricing policies can result in the segmentation of users into some segments where all users obtain the same product (as in our fixed-price market), while for all other users (who do not balk), the obtained product quality increases monotonically in

[6]In Appendix D, we provide intuition for why the fixed-price market, even if it is large, might not have a lot of reliably idle instances. In D, we also show how our results translate to providers who nevertheless want to use idle *fixed-price instances* (instead of some other, distinct pool of idle instances) for their spot market.

[7]The ability to combine the profit increase with a Pareto improvement for the users is particularly important in practice, where the competition between providers like Google, Microsoft, or Amazon is fierce. Thus, any strategy for increasing a provider's profits must guarantee that users are not worse off than before to ensure that they do not migrate to competing providers.

their type (as in our spot market). (Moorthy 1984) later extended the model of (Mussa and Rosen 1978) to non-linear user preferences (for discrete user types), showing that limiting the number of offered quality levels is often profit-optimal. These classic product differentiation papers give hope that offering a spot market alongside a fixed-price market may be profit increasing in some cases. However, because their analysis only considers simple posted price mechanisms (where users have no *direct* effect on each other) it does not apply to our cloud computing setting where users' actions have more complicated effects. In our problem, the values of the users for "obtaining a product" (i.e., running a job in a particular market) are not fixed but depend on the user equilibrium. Furthermore, the costs of the provider for offering a product are also not fixed but again depend on the user equilibrium. This is why we need to combine queuing theory and game theory (see, e.g., Hassin and Haviv (2003)) to answer the question when offering a spot market increases profits.

The work most related to ours is (Abhishek et al. 2012, 2017), who like us study cloud computing markets. Interestingly, the authors found that offering a spot market often decreases the provider's revenue. However, this does not contradict our results: their model was tailored towards a revenue analysis and therefore could assume an infinite number of fixed-price instances and did not need to model the provider's costs. Given this, they could not make any statements about profits. Gao et al. (2019) used a similar modeling approach to study the competition between two firms, where one firm only offers a fixed-price market of fixed finite size (i.e., that is not necessarily large enough for demand) while the other firm only offers a spot market.

Recently, a series of papers has studied the problem of selling resources through more complicated auction and pricing mechanisms that take individual requirements of jobs into account, including deadlines (Zhou et al. 2017), multi-dimensional resource requirements (Shi et al. 2014, Zhang et al. 2014), and the provider's opportunity cost for scheduling a given job (Boodaghians et al. 2019, Kash et al. 2017). However, in contrast to our model, these papers do not consider the following important business constraint: in practice, any provider who wants to keep her market share *must* also offer a non-preemptible fixed-price market alongside any other offerings, because many users want access to a fixed-price market. Furthermore, most of the prior work on cloud spot markets (including the papers cited above) do not consider the users' preemption costs. However, preemption costs are an important factor to analyze the profitability of the spot market, and previous authors (e.g., Subramanya et al. (2016)) have even argued that spot markets may become too unattractive once they become congested. However, we show that the costs incurred due to preemption are bounded, such that even congested spot markets remain attractive for the provider.

3 MODEL

In this section, we introduce our model. Before starting with the formal definitions, we provide some brief intuition for our framework. To analyze the profit a cloud provider obtains from running the different markets, we need to consider how her decisions affect the actions of potential users. To this end, we define a two-step model that is reminiscent of a Stackelberg game (Maharjan et al. 2013) with an important difference. As in a Stackelberg game, in a first step, the cloud provider chooses her actions (i.e., what markets to offer). This defines the parameters of the game the users will play in the next step. In the second step, with the provider's strategy fixed, the users then play the resulting game only with each other; i.e., they decide which market to join and potentially what to bid. In contrast to a Stackelberg game, the sub-game in the second step takes the form of a queuing system in steady state and therefore has no fixed set of users. Instead, users with certain parameters continuously arrive and depart. We assume that the users act rationally; thus, the provider's strategies can be fully analyzed by backwards induction from the equilibria of the steady state of the queuing system.

Remark. *As we analyze the queuing system in steady state, our model works directly on the stochastic processes and not on individual realizations. Thus, all outcomes of interest, such as the provider's profit or the users' waiting times and payments (which we will introduce in the next sections), are always "in expectation," even if we do not always explicitly denote them accordingly.*

The remainder of this section is structured as follows: First we introduce the models for the provider (Section 3.1) and for the users (Section 3.2). Then we present the models for how the fixed-price and spot markets work in Sections 3.3 and 3.4, respectively.

3.1 Provider Model

The type of a provider is defined by a tuple $(\kappa_F, \kappa_L, T, l, \psi_E)$. Here, κ_F denotes the *fixed costs* an instance causes per time unit in the provider's cloud computing center, i.e., the total fixed costs the instance causes over its lifetime amortized per time unit. We can think of this as mainly hardware, infrastructure, and maintenance costs that are independent of the actual utilization. Conversely, κ_L denotes all *load-dependent costs* that an instance causes per time unit it is running. This overwhelmingly consists of increased electricity costs. We call the sum of fixed and load-depended costs the *instance costs* $\kappa := \kappa_F + \kappa_L$. T is an internal *SLA* (Service Level Agreement) for the fixed-price market that ensures a satisfactory quality of service. The SLA is said to be satisfied if the expected time until a newly arriving job in the fixed-price market starts running is below T.[8] In practice, the choice of T is influenced by many factors outside our model. We therefore assume T to be exogenously given and not to be part of the provider's strategy space, but our results hold for any T. The number of idle instances the provider has available for the spot market, and thus the maximum number of instances she could sell on the spot market, is denoted by l. We assume that the provider draws these idle instances from any part of the cloud computing center (e.g., maintenance instances, long-term reserved instances), except from those instances offered on the fixed-price market.[9] As these instances are already part of the provider's cloud computing center, they do not incur fixed costs κ_F a second time when offered on the spot market but they still incur load-dependent costs κ_L.

Since the steady state analysis is only concerned with mean service times, we do not specifically model a process by which idle capacity becomes available and unavailable. Instead, we let any $l' \leq l$ denote the l' idle instances with the lowest individual probabilities to become unavailable. $\psi_E(l')$ then denotes the expected number of times an instance randomly selected from among these l' instances becomes unavailable per time unit. Note that this makes $\psi_E(l')$ weakly increasing in l'. We set $\psi_E(0) = 0$ by convention. Obviously, whenever an instance that is currently running becomes unavailable, a user gets preempted. We call $\psi_E(l')$ the number of *external preemptions* because it only encompasses preemptions caused by the unreliable availability of idle instances. In addition to this, there are *internal preemptions* caused by changes in the current market price (which we introduce in Section 3.4).

The choices of the provider define the setting in which the users find themselves. A strategy for the provider consists of a tuple $\rho = (p_F, l_S)$. We let p_F denote the price any user joining the fixed-price market has to pay per time unit his job is running. $l_S < l$ is the number of instances

[8]Note that the SLA can also take different shapes in practice and our model can easily be modified such that, instead of a limit on the expected queuing time T, some threshold on the percentage of rejections has to be met. Another possibility is to assume that users who cannot be served instantly simply do not join at all. Neither of these modifications change our main results.

[9]In Appendix D, we also provide an analysis for the case where the provider instead draws the idle instances from the *fixed-price market.*

the provider decides to offer on the spot market.[10] $l_S = 0$ denotes that she decides to only offer a fixed-price market. For simplicity, we assume that the provider does not set a reserve price for the spot market. Introducing a reserve price would only strengthen our results as it expands the provider's strategy space to make the spot market more profitable.

Note that while the number of fixed-price instances l_F could technically be seen as part of the provider's strategy, the fixed-price market *must* satisfy the SLA T. This bounds the number of fixed-price instances from below. To keep our arguments simple, we assume that the provider will always offer the smallest number of fixed-price instances for a given user strategy profile σ (as defined in Section 3.4) such that the SLA is satisfied, as this minimizes her costs.[11] Consequently, in our model, l_F is not itself part of the provider strategy, but given as a function $l_F(\rho, \sigma)$ of the provider strategy ρ and user strategy profile σ.

Given provider strategy ρ and user strategy profile σ, we let $R_F(\rho, \sigma)$ denote the revenue from all fixed-price instances and $C_F(\rho, \sigma)$ denote the costs from all fixed-price instances. Similarly, $R_S(\rho, \sigma)$ and $C_S(\rho, \sigma)$ denote the revenue and cost from all spot instances. The provider's (expected) *profit per time unit* is then defined as the sum of her revenues minus her costs, i.e.,

$$\Pi(\rho, \sigma) := R_F(\rho, \sigma) + R_S(\rho, \sigma) - C_F(\rho, \sigma) - C_S(\rho, \sigma). \tag{1}$$

3.2 User Model

We model the resulting game for the users, given provider strategy ρ, as a queuing system. We assume this system to be in steady state (i.e., the state probabilities do not change over time). Thus, there is no fixed set of players, because users arrive and depart over time. This allows us to analyze strategies for every parameter set a user's *job* could have instead of having to artificially enumerate each individual user. Queuing theory provides us with tools to analyze (a) the time a newly arriving user has to wait until he gets to run his job and (b) his expected payment.

Formally, let there be *n job classes* with fixed *values* $v = (v_1, \ldots, v_n)$ for completion where $v_i > v_{i+1}$ for all $i \in \{1, \ldots, n-1\}$. New jobs from each class arrive sequentially according to a memoryless arrival process.[12] The *arrival rates* of the different job classes are $\lambda = (\lambda_1, \ldots, \lambda_n)$; i.e., in expectation, λ_i jobs of class i arrive per time unit. Each individual job requires exactly one instance to run and is associated with a distinct *user*. Users are only identified by the parameters of their jobs; the terms "user" and "job" can thus be used interchangeably. The *service time* for each job (i.e., the time it has to run on an instance assuming it is not preempted) is independently drawn according to a distribution with expectation $\frac{1}{\mu}$. To keep our expositions and proofs concise, we assume that all classes of jobs have the same mean service time. Our main results do not require specific service processes, as they only make use of the first moments of the distributions.[13] For every job of class i that arrives, a *waiting cost* c is independently and privately drawn from a distribution $F_i(c)$. This distribution has a strictly positive PDF $f_i(c)$ on $[0, \mu v_i]$.[14] The waiting cost is incurred once per time unit until job completion. Every time a job is preempted, its user further incurs *preemption costs* in the form of an additional expected time loss τ. Concretely, this means that the expected

[10] To keep the exposition simple and avoid special handling of corner cases, our formal model allows fractional instances (in both markets). As the cost of a single instance is negligible in realistic settings, this is a reasonable abstraction.

[11] Note that in practice, cloud providers can approximately follow such a cost minimization strategy because of the high turn-over rate of their hardware.

[12] This assumption is natural, especially for large cloud computing centers, and is supported by empirical studies (Zaharia et al. 2010, Zheng et al. 2016).

[13] The queuing theory literature often denotes this combination of memoryless arrival and general (independent) service processes as *M/GI/"number of instances"*.

[14] Note that jobs with waiting costs $c > \mu v_i$ could only ever expect a negative payoff, even if they run instantly and pay nothing, and thus do not have to be considered.

payoff of a user with waiting cost c decreases by $c\tau$ for every expected preemption. Again, to keep the proofs concise, we assume that τ is independent of the job class. Note that when $\tau\psi_E(l_S) \geq 1$, then the time loss due to preemption per time unit is larger than one time unit (during which the job can again be preempted), such that (in expectation) the job will need an infinite amount of time to run to completion. Since that would trivially make offering a spot market of size l_S meaningless (as no user ever joins), we assume w.l.o.g. $\tau\psi_E(l) < 1$ (recall that l denotes the *maximum* number of spot instances the provider can offer). As is common in queuing 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 and functions $(n, v, \lambda, \mu, F, \tau, \kappa_F, \kappa_L, T, l, \psi_E)$ a *setting*. The setting is assumed to be fixed and known by the provider and all users.

For any single user, a possible strategy consists of the tuple (α, β). $\alpha \in \{\mathcal{F}, \mathcal{S}, \mathcal{B}\}$ represents the decision whether to join the fixed-price market $\mathcal{F}$, the spot market $\mathcal{S}$ or to *balk* $\mathcal{B}$ (i.e., not to join any market and obtain zero payoff). To simplify notation, we assume that the action $\mathcal{S}$ is equivalent to balking when there is no spot market, i.e., when the provider sets $l_S = 0$. Further, any user submits a bid β for the spot market (which, if he joins the spot market, determines how quickly he gets an instance and how much he has to pay). For users who do not join the spot market, this bid has no effect, and thus, w.l.o.g., is set to be equal to their waiting cost c. The current state of the queues (i.e., which other users are currently in the system) is unobservable for users and thus cannot influence their strategies. A strategy profile σ encodes the strategies for any possible user. It consists of functions $\sigma_i : [0, \mu v_i] \rightarrow \{\mathcal{F}, \mathcal{S}, \mathcal{B}\} \times \mathbb{R}$, one for each class of jobs $i \in \{1, \ldots, n\}$, that map waiting costs c to strategies (α, β). Whenever a provider strategy ρ is given, a strategy profile with an asterisk (i.e., σ^*) denotes a corresponding equilibrium strategy profile for the users.

For any user, we now denote by $q(\alpha, \beta, \rho, \sigma)$ the expected *queuing time* (i.e., the time he spends in a queue without running his job on an instance) when he plays strategy (α, β), assuming provider strategy ρ and that all other users play according to the strategy profile σ. By $r(\alpha, \beta, \rho, \sigma)$ we denote the expected *running time* a user requires, i.e., the total time the user's job has to run on an instance until completion. $r(\alpha, \beta, \rho, \sigma)$ is the sum of the user's "normal" service time $\frac{1}{\mu}$ and the additional time his job requires because of preemptions. The expected *total waiting time* until job completion is the sum of queuing time and running time: $w(\alpha, \beta, \rho, \sigma) := q(\alpha, \beta, \rho, \sigma) + r(\alpha, \beta, \rho, \sigma)$. The user has to pay some amount of money for using an instance. We denote this expected *payment* $m(\alpha, \beta, \rho, \sigma)$. Overall, the expected *payoff* for a user of class i with waiting cost c is then given by $\pi_i^c(\alpha, \beta, \rho, \sigma) := v_i - cw(\alpha, \beta, \rho, \sigma) - m(\alpha, \beta, \rho, \sigma)$ for joining a market, and zero for balking.

Remark. *Our motivation for the notational separation of the 6 parameters of the payoff function $\pi_i^c(\alpha, \beta, \rho, \sigma)$ is as follows. The parameters i and c fully identify an individual user, where i denote this user's job class and c denotes this user's realized waiting cost. The remaining parameters $(\alpha, \beta, \rho, \sigma)$ all represent strategies.*

3.3 Fixed-price Market and Queue

The fixed-price market consists of a queuing system where users pay a fixed price p_F for every time unit their job is running. This results in an expected payment of $m(\mathcal{F}, \beta, \rho, \sigma) = \frac{p_F}{\mu}$. Since users do not get preempted in the fixed-price market, their running time is equal to their service time, i.e., $r(\alpha, \beta, \rho, \sigma) = \frac{1}{\mu}$. In contrast to Abhishek et al. (2017), we assume that sometimes, users have to wait until their job finds a free instance and begins running, leading to a short (expected) queuing time $T > 0$.[15] Recall that T is the SLA, which is given exogenously.

[15] Note that it does not influence the equilibrium structure nor our results that jobs are queued and serviced in a first-come, first-served order. The same results are obtained if users instead continuously resubmit their jobs until they get a free instance (and are therefore effectively served in random order), because the expected service time would be the same.

Remark. *An (expected) queuing time of $T = 0$ is not attainable with any finite number of instances. An infinite number of instances would not be realistic and makes any profit analysis meaningless, as the costs would also be infinite.*

The expected payoff of a user of class i with waiting cost c that joins the fixed-price market is thus equal to:

$$\pi_i^c(\mathcal{F}, \beta, \rho, \sigma) = v_i - cw(\mathcal{F}, \beta, \rho, \sigma) - m(\mathcal{F}, \beta, \rho, \sigma) \tag{2}$$

$$= v_i - c(T + \frac{1}{\mu}) - \frac{p_F}{\mu}. \tag{3}$$

Note that, in the fixed-price market, the user's payoff (i.e., Equation (3)) is independent of the actions of other users because it only depends on the provider's choice for the price p_F.

The provider's revenue $R_F(\rho, \sigma)$ in the fixed-price market is straightforwardly given by the arrival rate of users into the market, while the costs $C_F(\rho, \sigma)$ additionally depend on how many instances she has to offer in order to guarantee the SLA T:

$$R_F(\rho, \sigma) = \frac{p_F}{\mu}\left(\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{F}} f_i(x)dx\right) \tag{4}$$

$$C_F(\rho, \sigma) = \frac{\kappa_L}{\mu}\left(\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{F}} f_i(x)dx\right) + \kappa_F l_F(\rho, \sigma) \tag{5}$$

3.4 Spot Market and Queue

Following Abhishek et al. (2017), we model the spot market as a preemptible priority queue where both payments and the order in which jobs are run depend on the users' bids. The preemptible priority queue consists of l_S instances, running jobs in a priority-based order, where l_S is set as part of the provider's strategy. A job's priority is set by the bid given on arrival. A running job may be preempted for two different reasons. First, the job may be outbid by a job with a higher bid (*internal preemption*). Second, the instance the job is running on may become unavailable for some exogenous reason (like the instance being required for its primary purpose), independent of the job's priority (*external preemption*). The expected number of external preemptions per time unit $\psi_E(l_S)$ is independent of bids or other users and only depends on the provider's strategy. In contrast, how often a user's job gets *internally* preempted depends on the arrival rate of users with higher bids. We let $\psi_I(c, \rho, \sigma)$ denote the expected number of times a user with bid c in the spot market gets internally preempted, i.e., outbid by other users during a time unit. Note that ψ_I is fully determined by the queuing model and thus, in contrast to $\psi_E(l_S)$, it is not part of the setting. To summarize, a user's running time also depends on how often he gets preempted and how much time he loses with each preemption. The following proposition formally shows all of these dependencies.

Proposition 3.1. *A user's running time with bid c is*

$$r(\mathcal{S}, c, \rho, \sigma) = \begin{cases} \frac{1}{\mu}\frac{1}{1-\tau(\psi_I(c,\rho,\sigma)+\psi_E(l_S))} & \text{for } \tau(\psi_I(c, \rho, \sigma) + \psi_E(l_S)) < 1 \\ \infty & \text{otherwise} \end{cases} \tag{6}$$

Proof. During any time unit where his job is running, a user is on average preempted $(\psi_I(c, \rho, \sigma) + \psi_E(l_S))$ times, causing him to require an additional running time of $\tau(\psi_I(c, \rho, \sigma) + \psi_E(l_S))$. During this additional running time he is then on average again preempted $\tau(\psi_I(c, \rho, \sigma) + \psi_E(l_S))$ times. Summing over these recursive preemptions and multiplying by his service time $\frac{1}{\mu}$ yields a geometric series, i.e.,

$$r(\mathcal{S}, c, \rho, \sigma) = \sum_{k=0}^{\infty} \frac{1}{\mu}\Big(\tau\Big(\psi_I(c, \rho, \sigma) + \psi_E(l_S)\Big)\Big)^k \tag{7}$$

$$= \begin{cases} \frac{1}{\mu}\frac{1}{1-\tau(\psi_I(c,\rho,\sigma)+\psi_E(l_S))} & \text{for } \tau(\psi_I(c, \rho, \sigma) + \psi_E(l_S)) < 1 \\ \infty & \text{otherwise} \end{cases} \tag{8}$$

□

Note that in any equilibrium, the running time of all jobs is trivially finite. Going forward, we can therefore safely ignore the case $\tau(\psi_I(c, \rho, \sigma) + \psi_E(l_S)) \geq 1$. A user's expected total waiting time when joining the spot market is therefore given by

$$w(\mathcal{S}, c, \rho, \sigma) = q(\mathcal{S}, x, \rho, \sigma) + \frac{1}{\mu}\frac{1}{1 - \tau(\psi_I(c, \rho, \sigma) + \psi_E(l_S))}. \tag{9}$$

Payments in the spot market are set according to some spot market mechanism which we do not explicitly model, as we are only interested in the expected payment in equilibrium. Abhishek et al. (2017) showed that it suffices to analyze a Bayes Nash Incentive Compatible (BNIC) spot market mechanism for which a user's bid β only consists of a revelation of its true waiting cost c, i.e. $\beta = c$. In the following, we therefore use the terms bid, waiting cost and priority interchangeably.

From Lemma 5 of Abhishek et al. (2017), which is an adaptation of Myerson's famous Lemma (Myerson 1981) to spot markets, we know that for any BNIC market mechanism employed in the spot market, the expected payment has to be:

$$m(\mathcal{S}, c, \rho, \sigma) = \int_0^c w(\mathcal{S}, x, \rho, \sigma)dx - cw(\mathcal{S}, c, \rho, \sigma) \tag{10}$$

We always assume that the provider employs a market mechanism whose payment rule for the spot market satisfies Equation (10). With such a payment rule, each user has to pay the difference between the overall cost caused by waiting he would incur at the mean waiting time of lower bids and the cost he incurs with his bid.

The total waiting time and the expected payment thus both depend on the number of users joining the spot market as determined by the strategy σ. For any user of class i with waiting cost c who joins the spot market the expected payoff can now be formulated as:

$$\pi_i^c(\mathcal{S}, c, \rho, \sigma) = v_i - cw(\mathcal{S}, c, \rho, \sigma) - m(\mathcal{S}, c, \rho, \sigma) \tag{11}$$

$$= v_i - \int_0^c w(\mathcal{S}, x, \rho, \sigma)dx. \tag{12}$$

The provider's revenue $R_S(\rho, \sigma)$ from the spot market now consists of the average payments users make. The costs $C_S(\rho, \sigma)$ are more complex, as a user getting preempted is also costly for the provider, since any job that loses time through preemption effectively has a longer running time and therefore causes more load-dependent costs. This means that the cost of the provider $C_S(\rho, \sigma)$ also depends on the number of preemptions:

$$R_S(\rho, \sigma) = \sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} m(\mathcal{S}, x, \rho, \sigma) f_i(x)dx \tag{13}$$

$$C_S(\rho, \sigma) = \kappa_L \sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} \frac{1}{\mu}\frac{1}{1 - \tau(\psi_I(x, \rho, \sigma) + \psi_E(l_S))} f_i(x)dx \tag{14}$$

4 EQUILIBRIUM ANALYSIS

In this section, we analyze the resulting Bayes Nash equilibria (BNEs) of the user game given a provider strategy ρ. A provider strategy can result in three basic types of equilibria: (a) *pure fixed-price equilibria*, where no user joins the spot market, (b) *pure spot equilibria*, where no user joins the fixed-price market and (c) *real hybrid equilibria*, where each market is chosen by some users. In the following, we first individually characterize the three types of equilibria (Sections 4.1, 4.2 and 4.3). In Section 4.4, we then show how to identify the type of equilibrium in which any given provider strategy ρ would result.

4.1 Pure Fixed-price Equilibria

As we have shown in Section 3.3, 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 $\pi_i^c(\mathcal{F}, \beta, \rho, \sigma) = v_i - c(T + \frac{1}{\mu}) - \frac{p_F}{\mu}$ for submitting a job is a monotonically decreasing function of the waiting cost c for every job class. We can therefore easily show that all equilibrium strategy profiles take the form of threshold functions.

Proposition 4.1. *For any provider strategy $\rho = (p_F, 0)$, the users' equilibrium strategy profile in any BNE takes the form $\sigma^* = \overrightarrow{c}^F$. Here, overloading our previous notation, $\sigma = \overrightarrow{c}^F$ denotes that all users of class i with waiting cost $c < c_i^F$ join the fixed-price market, i.e., they play $\alpha = \mathcal{F}$, and all users of class i with waiting cost $c > c_i^F$ balk and obtain zero payoff. The cutoff vector $\overrightarrow{c}^F = (c_1^F, \ldots, c_n^F)$ is the unique solution to the following system of equations:*

$$c_i^F = \frac{v_i - p_F \frac{1}{\mu}}{(T + \frac{1}{\mu})} \quad \forall i \in \{1, \ldots, n\} \tag{15}$$

Proof. Whenever $l_S = 0$, it directly follows that any equilibrium strategy profile is defined by a cutoff vector $\sigma^* = \overrightarrow{c}^F$ by the monotonicity of the payoff $\pi_i^c(\mathcal{F}, \beta, \rho, \sigma)$ in c. The expression for c_i^F follows via simple algebra by setting $\pi_i^{c_i^F}(\mathcal{F}, \beta, \rho, \sigma) = v_i - c_i^F(T + \frac{1}{\mu}) - \frac{p_F}{\mu} = 0$. □

Remark. *With a strategy profile of the form $\sigma = \overrightarrow{c}^F$, users whose jobs have waiting costs $c = c_i^F$ can join the fixed-price market or balk and obtain zero payoff either way. We do not need to use a tie-breaking rule because these users constitute a set with measure zero and do not influence the provider's profit or any user's payoff. The same holds for all future strategy profile characterizations we present using cutoff vectors.*

4.2 Pure Spot Equilibria

Pure spot equilibria arise when at least some spot instances are offered and no user joins the fixed-price market. If the provider chooses the price per time unit for fixed-price instances too high, every user either joins the spot market or cannot obtain a positive payoff in either market and balks. For settings without preemption costs, these equilibria have previously been studied in Abhishek et al. (2017). We now provide an analogous result for settings with preemption costs.

Proposition 4.2. *For any provider strategy $\rho = (p_F, l_S)$ with $l_S > 0$, in any BNE of the user game where no user joins the fixed-price market, the equilibrium strategy profile has the form $\sigma^* = \overrightarrow{c}^S$. Here, $\sigma^* = \overrightarrow{c}^S$ denotes that all users of class i with waiting cost $c < c_i^S$ join the spot market, i.e., they play $\alpha = \mathcal{S}$ and truthfully bid c, and all users of class i with waiting cost $c > c_i^S$ balk and obtain zero* *payoff. The cutoff vector* $\overrightarrow{c}^S = (c_1^S, \ldots, c_n^S)$ *is the unique solution to the following system of equations:*

$$v_i - \int_0^{c_i^S} w(\mathcal{S}, x, \rho, \overrightarrow{c}^S)dx = 0 \quad \forall i \in \{1, \ldots, n\} \tag{16}$$

The proof of Proposition 4.2 can be found in Appendix A.2. Intuitively, the existence of a cutoff vector $\overrightarrow{c}^S$ follows because, for any fixed strategy profile σ, the payoff of a user of class i, i.e., $\pi_i^c(\mathcal{S}, c, \rho, \sigma)$, is monotone decreasing in his waiting cost c.

4.3 Hybrid Equilibria

We now analyze *hybrid equilibria* that arise when the provider plays $l_S > 0$ on the spot market and some users still join the provider's fixed-price market. These equilibria can again be subdivided into two cases. In the first case, jobs in the spot market incur relatively high preemption costs and (independent of their bid) take longer until completion than in the fixed-price market, i.e., $w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*) \leq w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*)$, where $\overline{\beta}$ denotes a bid that is higher than any other bid in σ^*. Plugging in the definitions and using simple algebra, this holds whenever $T \leq \frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right)$. To understand the second case, note that the jobs with the highest bids are started instantly in the spot market, while they always have to wait some small time $T > 0$ in expectation to start running in the fixed-price market. When spot instances are very reliably idle, this can in theory lead to situations where the spot market requires a shorter time until completion than the fixed-price market for jobs with very high bids (i.e. $w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*) > w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*)$). While in practice, this is unlikely to happen, we still need to include this case in the equilibrium analysis.

In the following, we first analyze the case where the spot market is slower than the fixed-price market. We then present the equilibrium analysis for the more exotic case where the spot market is faster for a small number of user.

4.3.1 Case 1: Spot Market Slower Than the Fixed-price Market.

If the overall time lost through external preemptions with the highest bid (i.e., without ever getting internally preempted) is higher than the expected queuing time in the fixed-price market, then no user is willing to pay more in the spot market than in the fixed-price market. As the overall costs (i.e., the costs for waiting plus the payment) of a user who has arrived into the system do not depend on his class, there exists a waiting cost for which both markets result in the same payoff, independent of a job's class. Below that waiting cost users prefer the spot market and above it they prefer the fixed-price market; though in either case they might still balk if their value is too low (leading to a negative payoff in both markets). This again allows us to state the equilibrium strategy profiles as cutoff vectors, this time with two vectors $\overrightarrow{c}^P$ (P for payoff equivalence) and $\overrightarrow{c}^B$ (B for balking).

Proposition 4.3. *For any provider strategy* $\rho = (p_F, l_S)$ *with* $l_S > 0$ *and* $w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*) \leq w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*)$, *in any BNE of the user game where any user joins the fixed-price market, the equilibrium strategy profile is of the form* $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$. *Here,* $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ *denotes that a user of class i with waiting cost c joins the spot market when* $c < c_i^P \leq c_i^B$ *and the fixed-price market when* $c_i^P < c < c_i^B$*; when* $c > c_i^B$*, he balks and does not join any market. The cutoff point* c_1^P *and the cutoff vector* $\overrightarrow{c}^B$ *are* *the unique solution to the following system of equations:*

$$0 = c_1^P(T + \frac{1}{\mu}) + \frac{p_F}{\mu} - \int_0^{c_1^P} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B))dx \tag{17}$$

$$0 = v_i - \min\left\{c_i^B(\frac{1}{\mu} + T) + \frac{p_F}{\mu}, \int_0^{c_i^B} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B))dx\right\} \quad \forall i \in \{1, \ldots, n\} \tag{18}$$

The rest of the cutoff vector $\overrightarrow{c}^P$ *is given as* $c_i^P = min(c_1^P, c_i^B)$.

The proof of Proposition 4.3 can be found in Appendix A.3. In words, $\overrightarrow{c}^P$ denotes until which waiting cost the payoff in the spot market is higher than either balking or joining the fixed-price market. The second vector (i.e., $\overrightarrow{c}^B$) denotes above which waiting cost neither market allows users to obtain a positive payoff anymore. Note that there are often some classes i for which $c_i^P = c_i^B$, i.e., no user from those classes joins the fixed-price market.

4.3.2 *Case 2: Spot Market Faster Than the Fixed-price Market.*

When the spot market for very high bids is faster than the fixed-price market, i.e., $w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*) < w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*)$, then some users are willing to pay more in the spot market than what they would in the fixed-price market. However, most users in the spot market still have to wait longer, and are therefore not willing to pay as much as in the fixed-price market. As a result, we obtain the following equilibrium strategy profiles, now with three cutoff vectors: $\overrightarrow{c}^L$ (L for lower bound of the fixed-price market), $\overrightarrow{c}^U$ (U for upper bound of the fixed-price market) and $\overrightarrow{c}^H$ (H for hybrid).

Proposition 4.4. *For any provider strategy* $\rho = (p_F, l_S)$ *with* $l_S > 0$, *in any BNE of the user game where any user joins the fixed-price market and where it holds that* $w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*) < w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*)$, *the equilibrium strategy profile is of the form* $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$. *Here,* $\sigma = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$ *denotes that a user of class* i *with waiting cost* c *joins the spot or the fixed-price market if and only if* $c \leq c_i^H$; *out of those users that join any market, almost all (i.e., all besides possibly a set of measure zero that does not influence system dynamics) choose the fixed-price market if and only if* $c_i^L \leq c \leq c_i^U$ *and choose the spot market otherwise. The cutoff points* c_1^L, c_1^U *and the cutoff vector* $\overrightarrow{c}^H$ *are the unique solution to the following system of equations:*

$$c_1^L(T + \frac{1}{\mu}) + p_F\frac{1}{\mu} - \int_0^{c_1^L} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H))dx = 0 \tag{19}$$

$$c_1^U(T + \frac{1}{\mu}) + p_F\frac{1}{\mu} - \int_0^{c_1^U} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H))dx = 0 \tag{20}$$

$$v_i - \int_0^{c_i^H} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H))dx = 0 \quad \forall i \in \{1, \ldots, n\} \tag{21}$$

The rest of the cutoff vectors $\overrightarrow{c}^L$ *and* $\overrightarrow{c}^U$ *are given as* $c_i^L = min(c_1^L, c_i^H)$ *and* $c_i^U = min(c_1^U, c_i^H)$.

The proof of Proposition 4.4 can be found in Appendix A.5. In words, $\overrightarrow{c}^H$ denotes the waiting costs of each class above which users cannot obtain a positive payoff in either market and balk. Of those users that do not balk, only users of class i whose waiting cost c lies between c_i^L and c_i^U join the fixed-price market, while every other user whose waiting cost c lies below c_i^H joins the spot-market.

4.4 Equilibria Resulting From Provider Strategy

So far, we have analyzed different forms of equilibria for the user game. We now analyze what kind of equilibrium a given provider strategy ρ results in. Because users with very low waiting costs always join the spot market, pure fixed-price equilibria can only result when no spot instances are offered, i.e., when the provider plays $l_S = 0$. Going forward, we call such provider strategies *fixed-price strategies*. Conversely, we call all strategies where the provider offers any spot instances (i.e., where she plays $l_S > 0$) *hybrid strategies*.

Given a hybrid strategy ρ, it is easy to differentiate which of the two different hybrid equilibria the strategy ρ potentially results in by simply checking whether a congestion free spot market is faster than the fixed-price market, i.e., by comparing T to $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right)$. But it is not directly apparent whether a hybrid strategy results in a "real" hybrid equilibrium or degenerates the market into a pure spot equilibrium where no user joins the fixed-price market. Fortunately, we can formulate a simple condition that, as we will show, distinguishes these two types of equilibria.

Definition 4.5 (Proper Hybrid Strategy). For any ρ with $l_S > 0$, let $\sigma = \overrightarrow{c}^S$ be a strategy profile satisfying Equation (16). Recall that, with such a strategy profile σ, no user joins the fixed-price market and σ would be an equilibrium strategy profile (of the spot market) if no fixed-price market existed. We say that ρ is a *proper hybrid strategy* (or *proper*) if one of the following holds:

(1) $w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*) \leq w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*)$ and the payoff for users with waiting cost c_1^S is higher in the fixed-price than in the spot market (and thus a beneficial deviation from $\sigma = \overrightarrow{c}^S$ exists), or
(2) $w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*) < w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*)$ and, under strategy profile $\sigma = \overrightarrow{c}^S$, there exists a waiting cost c', such that (a) the total waiting time with bid c' in the spot market is equal to the waiting time in the fixed-price market, and (b) the payoff for a user with waiting cost c' is higher in the fixed-price market than in the spot market (and thus a beneficial deviation from $\sigma = \overrightarrow{c}^S$ exists).

Informally, a *proper hybrid strategy* only requires that users with one specific waiting cost (i.e., c') prefer the fixed-price market over the spot market. The definition is well defined because Eq. (16) always has a solution and thus allows us to calculate the strategy profile $\overrightarrow{c}^S$. We now show that this definition enables us to determine whether or not a given provider strategy results in a BNE where some users join the fixed-price market.

Lemma 4.6. *Let $\rho = (p_F, l_S)$ be a hybrid strategy, i.e., a provider strategy with $l_S > 0$. In any BNE σ^* of the user game, some users join the fixed-price market (i.e., $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ or $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$) if and only if ρ is proper.*

The proof of Lemma 4.6 can be found in Appendix A.6.

With this result, we can now give a full equilibrium characterization result based purely on the provider strategy ρ.

Theorem 4.7. *For any provider strategy ρ, the equilibrium strategy profile of the users is*

(1) $\sigma^* = \overrightarrow{c}^F$ *if and only if ρ is a fixed-price strategy, i.e., $l_S = 0$.*
(2) $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ *if and only if ρ is a proper hybrid strategy and it holds that $w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*) \leq w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*)$.*
(3) $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$ *if and only if ρ is a proper hybrid strategy and it holds that $w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*) < w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*)$.*
(4) $\sigma^* = \overrightarrow{c}^S$ *otherwise.*

Proof. We show the "if" direction of each case in turn. Once the "if" direction for all cases have been shown, the "only if" direction immediately follows for each of them.

(1) Assume that $l_S = 0$, then $\sigma^* = \overrightarrow{c}^F$ follows from Proposition 4.1. For $l_S > 0$ at least users with waiting costs in some small neighborhood around zero will trivially prefer the spot market.

(2) Assume that $l_S > 0$, $w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*) \leq w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*)$ and the strategy is proper. By Lemma 4.6 some users join the fixed-price market. By Proposition 4.3 it follows that $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$.
(3) Assume that $l_S > 0$, $w(\mathcal{S}, \overline{\beta}, \rho, \sigma^*) < w(\mathcal{F}, \overline{\beta}, \rho, \sigma^*)$ and the strategy is proper. By Lemma 4.6 some users join the fixed-price market. By Proposition 4.4 it follows that $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$.
(4) Assume that $l_S > 0$ and the strategy is not proper. By Lemma 4.6 no users join the fixed-price market. By Proposition 4.2 it follows that $\sigma^* = \overrightarrow{c}^S$.

□

5 PUTTING IT ALL TOGETHER: PROVIDER PROFIT AND USER WELFARE

In the previous section, we have derived the user equilibrium strategy profiles σ^*, given different provider strategies ρ. As these uniquely define the provider's costs and revenue, we can now bound how a provider's costs change when she offers a spot market. Using these bounds, we then assemble a condition that only depends on the setting and show that, when this condition holds, a provider can always simultaneously increase her profit and achieve a Pareto improvement for the users by also offering a spot market alongside her fixed-price market.

5.1 Bounding the Provider's Costs

Towards deriving our results, we now bound the costs of the spot and fixed-price markets. First, we state a condition to formally bound the fixed-cost savings a provider can obtain by offering spot instances. Second, we derive a lemma to bound the average running time in the spot market. We then use these to bound the provider's costs. We begin with bounding the reduction in fixed-price instances. Going forward, we denote as σ_0^* the user equilibrium resulting from a provider strategy with $l_S = 0$ spot instances to distinguish it from the user equilibrium σ^* resulting from another strategy.

LEMMA 5.1. *For any fixed-price strategy $\rho_0 = (p_F^0, 0)$ and any hybrid strategy $\rho = (p_F^0, l_S)$ with the same price p_F^0, denote by $\Delta l_F(l_S) := l_F(\rho_0, \sigma_0^*) - l_F(\rho, \sigma^*)$ the reduction in required fixed-price instances and by $\lambda_S(l_S)$ the arrival rate of all users who join the fixed-price market under ρ_0 but move to the spot market if the provider offers l_S spot instances. Then it holds that*

$$\Delta l_F(l_S) \geq \frac{1}{\mu}\lambda_S(l_S). \tag{22}$$

PROOF. Given that the arrival process into the system (i.e., before users take an action) is memoryless (i.e., the number of arrivals in any time interval is distributed as a Poisson random variable), the arrival process into the fixed-price market for any strategy profile σ is also memoryless. Furthermore, the number of users who choose the fixed-price market in equilibrium is strictly decreasing in the number of offered spot instances l_S. Thus, in equilibrium, the fixed-price markets under any ρ_0 and ρ can be seen as two queues Q_1, Q_2 with the same queuing time, Poisson arrival rates $\lambda_1 > \lambda_2$, the same service process, and l_1, l_2 instances, respectively. To keep the notation simple, w.l.o.g., we assume that the service time is normalized to $\frac{1}{\mu} = 1$. The statement of the lemma is therefore equivalent to the difference between the number of instances of any such Q_1, Q_2 being larger or equal to the difference between the arrival rates, i.e., $l_1 - l_2 \geq \lambda_1 - \lambda_2$.

Since $\lambda_1 > \lambda_2$, we can write $\lambda_1 = \lambda_A + \lambda_2$ for some $\lambda_A > 0$. Since the arrival process into Q_1 is memoryless, we can further see it as a mix of two independent arrival processes A and B with rates λ_A and $\lambda_B = \lambda_2$. Now, for the sake of contradiction, assume that $l_1 - l_2 < \lambda_1 - \lambda_2 = \lambda_A$. Since the number of instances l_1 is finite, some users will sometimes have to wait to be served. Thus, at any randomly chosen point in time, there are, in expectation, strictly more than λ_A users that arrived from arrival process A in Q_1. Given that $l_1 - \lambda_A < l_2$, there are, at any random point in time (in expectation) less than l_2 instances available to serve the users from arrival process B in Q_1. Given that the arrival processes are memoryless, the distribution of the states of the queue at a random point in time is the same as whenever a random user arrives.[16] Therefore, there are also, in expectation, less than l_2 instances available for users from process B whenever a random user from process B arrives. This implies that when a random user arrives from process B into Q_1, in expectation, there are strictly fewer idle instances available in Q_1 than when a random user arrives into Q_2. It follows that users from arrival process B (in expectation) have a longer queuing time in Q_1 than users have in Q_2. Since Q_1 is a FIFO queue with memoryless arrivals, every user in Q_1 has the same expected queuing time independent of whether he arrives by process A or B. Consequently, the queuing time of any user in Q_1 is longer than the queuing time of any user in Q_2, a contradiction to our definition of Q_1 and Q_2. Thus, it must hold that $l_1 - l_2 \geq \lambda_1 - \lambda_2$. □

Note that Equation (22) from Lemma 5.1 immediately implies a lower bound on the fixed costs the provider saves by offering a spot market. This is the case because the provider's fixed costs only depend on κ_F and the number of fixed-price instances required in equilibrium.

Next, we bound the average running time in the spot market.

LEMMA 5.2. *The average running time in the spot market (i.e., the left-hand side of the following inequality) is bounded above as follows:*

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} < (\frac{1}{\mu} + \tau)\frac{1}{1-\tau\psi_E(l_S)} \tag{23}$$

PROOF. First, note that in any queue that is stable, i.e., where every user has a finite waiting time, (on average) the exact same number of users arrive and depart per time unit. Now assume that any arriving user preempts one other job through its arrival. Obviously, this is an upper bound, as no job can preempt more than one job due to its arrival. Since the average number of arrivals is the same as the average number of departures, after a job arrives and causes a preemption, another job has to depart (in expectation) before the next preemption. Thus, (on average) a job cannot be internally preempted more than once during its whole running time. Denoting the number of internal preemptions a job suffers in expectation by $\psi_I(c,\rho,\sigma)r(\mathcal{S},c,\rho,\sigma)$, it follows that:

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} \psi_I(x,\rho,\sigma) r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} < 1 \tag{24}$$

Now note that the full running time of a job can be split into the running time it would require without internal preemption and any additional running time $r_I(\mathcal{S},c,\rho,\sigma)$ needed because of internal preemption (caused either directly, or indirectly via additional external preemptions that occur during the additional running time), i.e.,

$$r(\mathcal{S},c,\rho,\sigma) = (\frac{1}{\mu}\frac{1}{1-\tau\psi_E(l_S)}) + r_I(\mathcal{S},c,\rho,\sigma). \tag{25}$$

While the time lost directly through internal preemptions is given by $\psi_I(c,\rho,\sigma)r(\mathcal{S},c,\rho,\sigma)\tau$, additional external preemptions can occur during this time. By the same argument as for the

[16]This is called the PASTA (Poisson Arrivals See Time Averages) property (see (Wolff 1982)).

running time in Proposition 3.1 (i.e., viewing it as a geometric series) it follows that

$$r_I(\mathcal{S},c,\rho,\sigma) = \sum_{k=0}^{\infty} \psi_I(c,\rho,\sigma) r(\mathcal{S},c,\rho,\sigma)\tau(\tau\psi_E(l_S))^k \tag{26}$$

$$= \psi_I(c,\rho,\sigma) r(\mathcal{S},c,\rho,\sigma)\tau\frac{1}{1-\tau\psi_E(l_S)}. \tag{27}$$

We can now upper bound the average additional running time caused by internal preemptions:

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r_I(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} \tag{28}$$

$$= \frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} \psi_I(x,\rho,\sigma) r(\mathcal{S},x,\rho,\sigma)\tau\frac{1}{1-\tau\psi_E(l_S)} f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} \tag{29}$$

$$= \frac{\tau}{1-\tau\psi_E(l_S)} \frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} \psi_I(x,\rho,\sigma) r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} \tag{30}$$

$$< \frac{\tau}{1-\tau\psi_E(l_S)}. \tag{31}$$

Here, the inequality in (31) follows by plugging (30) into Inequality (24). It now directly follows for the average running time that

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} \tag{32}$$

$$= \frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} (\frac{1}{\mu}\frac{1}{1-\tau\psi_E(l_S)}) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} + \frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r_I(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} \tag{33}$$

$$< \frac{1}{\mu}\frac{1}{1-\tau\psi_E(l_S)} + \frac{\tau}{1-\tau\psi_E(l_S)} \tag{34}$$

$$= (\frac{1}{\mu}+\tau)\frac{1}{1-\tau\psi_E(l_S)}. \tag{35}$$

□

Lemma 5.2 immediately implies a bound on the load-dependent costs incurred from offering a spot market by combining Equation (23) with the definition of the costs $C_S(\rho,\sigma)$ in Equation (14).

5.2 Well-behaved Settings: Increasing Provider Profit and User Welfare

In this subsection, we use Lemmas 5.1 and 5.2 to derive a very mild condition on the setting under which we then show our main result. First, recall that the average running time in the fixed-price market is equal to the service time $\frac{1}{\mu}$. As we have shown in Lemma 5.2, the average running time in the spot market is bounded by $(\frac{1}{\mu}+\tau)\frac{1}{1-\tau\psi_E(l_S)}$. This immediately implies an upper bound on the *difference* between the average running time in the spot and the fixed-price market: $\frac{1}{\mu}\left(\frac{1+\tau\mu}{1-\tau\psi_E(l_S)}-1\right)$. The following condition puts this bound (after normalizing by the service time $\frac{1}{\mu}$) in relation to the ratio between the provider's fixed and load-dependent costs.

Definition 5.3. We call a setting $(n, v, \lambda, \mu, F, \tau, \kappa_F, \kappa_L, T, l, \psi_E)$ *well-behaved* if there exists a number of spot instances l^w with $0 < l^w \leq l$ such that the following holds:

$$\frac{1+\tau\mu}{1-\tau\psi_E(l^w)} - 1 < \frac{\kappa_F}{\kappa_L} \tag{36}$$

In the following theorem, we show that, in a well-behaved setting, a provider can increase her profit as well as achieve a Pareto improvement for the users by offering a spot market, compared to only offering a fixed-price market.

THEOREM 5.4. *Given a well-behaved setting, for every fixed-price strategy $\rho_0 = (p_F^0, 0)$ that results in a positive profit, there exists a hybrid strategy $\rho = (p_F^0, l_S)$ with the same price p_F^0 and with $0 < l_S \leq l$ that yields a higher profit for the provider, i.e.,*

$$\Pi((p_F^0, l_S), \sigma^*) > \Pi((p_F^0, 0), \sigma_0^*), \tag{37}$$

and the same strategy also yields a Pareto improvement for the users, i.e.,

$$\forall i \in \{1, \dots, n\}\, \forall c \in [0, \mu v_i] : \quad \pi_i^c(\alpha, \beta, \rho, \sigma^*) \geq \pi_i^c(\alpha, \beta, \rho_0, \sigma_0^*), \text{ and} \tag{38}$$

$$\exists i \in \{1, \dots, n\}\, \exists c \in [0, \mu v_i] : \quad \pi_i^c(\alpha, \beta, \rho, \sigma^*) > \pi_i^c(\alpha, \beta, \rho_0, \sigma_0^*). \tag{39}$$

PROOF. First, we collect all required auxiliary results. Denote by $\lambda_S(l_S)$ the arrival rate of all users who join the fixed-price market under σ_0^* but move to the spot market under σ^*. Denote by $\lambda_N(l_S)$ the arrival rate of all users who balk under σ_0^* but newly join the spot market under σ^*. Thus, the arrival rate of all users into the spot market under σ^* is $\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx = \lambda_N(l_S) + \lambda_S(l_S)$.

Recall that Lemma 5.1 directly translates to a lower bound on the difference in the fixed costs incurred under ρ and ρ_0, i.e., $\kappa_F \Delta l_F(l_S) \geq \kappa_F \frac{1}{\mu}\lambda_S(l_S)$. Denote by $b(l_S)$ the bound on the normalized average running time as derived in Lemma 5.2, i.e., $b(l_S) := \frac{1+\tau\mu}{1-\tau\psi_E(l_S)}$. By combining Equation (23) with the definition of $C_S(\rho, \sigma)$ in Equation (14), it then follows that the increase in load-dependent costs caused by $\lambda_S(l_S)$ users switching from the fixed-price to the spot market and $\lambda_N(l_S)$ users switching from balking to the spot market is bounded by $\frac{\kappa_L}{\mu}(\lambda_S(l_S) - (\lambda_N(l_S) + \lambda_S(l_S))b(l_S))$. We can now use both of these results to bound the cost difference between ρ_0 and ρ:

$$\begin{aligned} C_F(\rho_0, \sigma_0^*) - \left(C_F(\rho, \sigma^*) + C_S(\rho, \sigma^*)\right) \quad &\geq \quad \frac{\kappa_F}{\mu}\lambda_S(l_S) + \frac{\kappa_L}{\mu}(\lambda_S(l_S) - (\lambda_N(l_S) + \lambda_S(l_S))b(l_S)) && (40) \\ &= \quad \lambda_S(l_S)(\frac{\kappa_F}{\mu} + \frac{\kappa_L}{\mu}(1 - b(l_S))) - \lambda_N(l_S)\frac{\kappa_L}{\mu}b(l_S) && (41) \end{aligned}$$

Lemma A.3 in the appendix shows that for every $\varepsilon > 0$, there exists an l_S such that the *average* payment of the users that join the spot market $\bar{m}(\mathcal{S}, (p_F^0, l_S), \sigma^*)$ is at least the expected payment in the fixed-price market minus ε, i.e.,

$$\bar{m}(\mathcal{S}, (p_F^0, l_S), \sigma^*) \geq \frac{p_F^0}{\mu} - \varepsilon. \tag{42}$$

For any $l_S \leq l^w$ that satisfies Equation (42) for a given ε, the revenue difference between ρ and ρ_0 can therefore be bounded as follows:

$$R_S((p_F^0, l_S), \sigma^*) + R_F((p_F^0, l_S), \sigma^*) - R_F((p_F^0, 0), \sigma_0^*) \tag{43}$$

$$= (\lambda_S(l_S) + \lambda_N(l_S))\left[\bar{m}(\mathcal{S}, (p_F^0, l_s), \sigma^*)\right] - \lambda_S(l_S)\left[\frac{p_F^0}{\mu}\right] \tag{44}$$

$$\geq \; -\lambda_S(l_S)\varepsilon + \lambda_N(l_S))\left[\frac{p_F^0}{\mu} - \varepsilon\right] \tag{45}$$

Combining the bounds on revenue and costs for any $l_S \leq l^w$ that satisfies Equation (42) for a given ε, the profit difference between the hybrid strategy $\rho = (p_F^0, l_S)$ and the fixed-price strategy $\rho_0 = (p_F^0, 0)$ can now be bounded as follows:

$$\Pi((p_F^0, l_S), \sigma^*) - \Pi((p_F^0, 0), \sigma_0^*) \tag{46}$$

$$\geq \lambda_S(l_S)\left[\frac{\kappa_F}{\mu} + \frac{\kappa_L}{\mu}(1 - b(l_S)) - \varepsilon\right] + \lambda_N(l_S)\left[\frac{p_F^0}{\mu} - \varepsilon - \frac{\kappa_L}{\mu}b(l_S)\right] \tag{47}$$

To see that this expression is positive for correctly chosen ε and l_S, note that for well-behaved settings, it directly follows that

$$\frac{\kappa_F}{\mu} + \frac{\kappa_L}{\mu}\left(1 - b(l_S)\right) > 0. \tag{48}$$

Since offering more spot instances causes a higher external preemption rate, the left-hand side of Equation (48) decreases in l_S. Further, since the fixed-price strategy obtains a positive profit, it also holds that

$$\frac{p_F^0}{\mu} > \frac{\kappa_F}{\mu} > \frac{\kappa_L}{\mu}b(l_S). \tag{49}$$

Thus, for ε small enough, any strategy $\rho = (p_F^0, l_S)$ with l_S satisfying Equation (42) increases the profit compared to the fixed-price strategy $\rho_0 = (p_F^0, 0)$. Since the price p_F^0 did not change, the users still have access to the same fixed-price market as before, but additionally now also have access to a spot market (which some users prefer), which leads to a Pareto improvement for the users. □

Theorem 5.4 shows that in any well-behaved setting, a provider can increase her profit and at the same time achieve a Pareto improvement for the users by playing a hybrid strategy compared to playing a fixed-price strategy. The following corollary follows immediately.

Corollary 5.5. *In any well-behaved setting, the provider's profit-optimal strategy is a hybrid strategy.*

Remark. *Note that the strategies characterized by Theorem 5.4 and Corollary 5.5 are typically not the same. In particular, while the strategy described by Theorem 5.4 simultaneously increases the provider's profit and leads to a Pareto improvement for the users, the profit-optimal strategy described by Corollary 5.5 may increase or* decrease *user welfare.*

5.3 Discussion

Our results show that, in a well-behaved setting, a cloud provider can increase her profit by offering a spot market consisting of idle capacity in addition to offering her existing fixed-price market. In practice, a cloud provider can easily check whether a setting is *well-behaved.* To see this, note that the condition is independent of internal preemptions and only depends on setting parameters: the fixed and load-dependent costs κ_F and κ_L, the preemption costs τ, the expected service time $\frac{1}{\mu}$, and the rate of external preemptions ψ_E (i.e., how reliably idle the provider's capacity is). Whether a setting is well-behaved can therefore directly be evaluated without having to use any queuing-theoretic formulas or equilibrium calculations.

Our well-behavedness condition is quite mild and most cloud providers should find it satisfied in practice. To see this, note that fixed costs κ_F are usually about 5 to 20 times higher than load-dependent costs (Barroso et al. 2018). Even if each preemption in the spot market resulted in an additional run time of 25% of a job's expected service time (which is an unreasonably high number, considering that mostly users with low preemption costs τ would use the spot market), the condition would still be satisfied if a job on average gets externally preempted 3 times per time unit before it finishes. Of course, some cloud providers may see the condition *not* satisfied: for example, if they have very low fixed costs (e.g., by using old instances whose acquisition costs are already amortized) or if they do not have reliably idle capacity. This demonstrates how our condition can inform the provider's managerial decision making process.

A provider whose setting is well-behaved might also want to compute her optimal strategy, i.e., the optimal price and number of spot instances. Unfortunately, directly calculating the optimal strategy is only feasible for very few arrival/service processes (see Section 6 for examples). For general service processes, formulas for queuing times are open problems of queueing theory. However, this does not mean that a provider cannot find a profit-increasing hybrid strategy. In practice, the provider can keep the same price she used when only offering fixed-price instances and start with a relatively small spot market. According to Theorem 5.4, this already leads to a (small) profit increase. Over time, the provider can then successively increase the size of the spot market until she observes no further profit increase. Alternatively, she can employ a reserve price for the spot market, starting with a relatively high price and successively decreasing it.

Note that in this work, we have analyzed the profit per time unit, and therefore our model does not include *one-time costs*. However, depending on how their cloud computing centers are structured, providers might face varying degrees of one-time costs (e.g., to set up a new marketplace and enable offering preemptible instances). While these costs are only incurred once and therefore become negligible over time, a provider with a shorter planning horizon might still want to take these costs into account. Providers may also face additional costs whenever a user gets preempted (e.g., for re-booting a machine after a preemption). As these costs can take different shapes for different providers and do not influence the user sub-game, we did not include them in our model, but it would be straightforward to add them. Such costs simply add another term to the well-behavedness condition, but do not change our results in any meaningful way.

While in this paper, we only model a single provider, some insights from our theoretical results also extend to *competitive* multi-provider settings. For such a setting, we would have to generalize our well-behavedness condition to multiple providers, which is straightforward. Following a similar argument as in the proof of Theorem 5.4, one could then show that, if none of the providers currently offers a spot market, then any provider for whom the well-behavedness condition is satisfied can increase her profit while simultaneously achieving a Pareto improvement for the users by offering a secondary spot market. Since this means that offering a spot market would be a profitable single-provider deviation from any strategy profile where none of the providers offer a spot market, there cannot be an equilibrium where no provider offers a spot market. We leave the formalization of multi-provider settings and a detailed study of the resulting equilibria to future work.

6 NUMERICAL EXAMPLES FOR MEMORYLESS QUEUES

In this section, we provide some numerical examples to illustrate the main results we have derived in the previous sections. So far, we have derived all theoretical results for *general* service processes. Now, we focus on fully *memoryless* queues, for which the well-known Erlang C formula (e.g., (Cooper 1981)) allows us to calculate queuing times and the expected number of preemptions (given some additional technical assumptions). The formal model for the numerical examples and the corresponding formulas for the waiting time are provided in Appendix B.

6.1 Set-up

In our examples, we consider a setting with two classes of jobs. The parameters of the examples are as follows: the values for completion are $v = (1, 0.75)$, the arrival rates are $\lambda = (100, 50)$, and the expected service time is $\frac{1}{\mu} = 1$. The waiting costs c are uniformly distributed on $[0, 1]$ and $[0, 0.75]$, respectively. The SLA on the expected total waiting time for the fixed-price market is set to $T = 0.001$. For any job that joins the spot market, the expected number of external preemptions per time unit is $\psi_E(l_S) = l_S/100$. We assume that the provider can at most offer $l = 100$ spot instances; however, in our examples, this limit is never reached by the provider's optimal strategy. We set the preemption costs to $\tau = 0.25$ (i.e., whenever a job gets preempted it incurs additional running time equal to 25% of its "normal" service time $\frac{1}{\mu}$). We choose these relatively high costs to demonstrate that sizable profit increases are possible even with relatively costly preemptions.

6.2 Example 1: Varying the Instance Costs $\kappa = \kappa_F + \kappa_L$

For the first example, we vary the provider's instances costs $\kappa = \kappa_F + \kappa_L$ between $\kappa = 0$ and $\kappa = 0.2$. We assume that 90% of the costs are fixed-costs κ_F, while the remainder are load-dependent costs κ_L, i.e., $\kappa_F = 9\kappa_L$.

In Figure 1, we show the profit for different strategies, varying the instance costs κ on the x-axis. We plot the following four strategies. First, the red solid line shows the provider's profit-optimal strategy (denoted "hybrid"). Second, the dotted black line shows the profit-optimal strategy for when the provider is restricted to use the price p_F^{*0} that is optimal when only offering a fixed-price market (denoted "hybrid with $p_F = p_F^{*0}$"). Note that, by construction, this strategy guarantees a Pareto improvement for the users compared to the optimal fixed-price strategy. This is true because the users have access to the same fixed-price market, but additionally now also have access to a spot market (which some users prefer). Third, the dashed blue line shows the profit-optimal strategy for when the provider is restricted to only offering a fixed-price market (denoted "fixed-price"). Fourth, and as a reference, the dash-dotted green line shows the profit-optimal strategy for when no fixed-price market exists (denoted "spot-only").

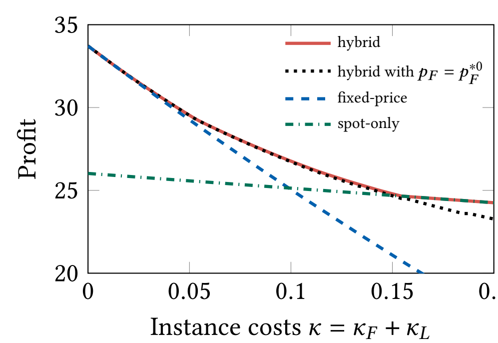

Fig. 1 Profit under different strategies while varying the instance cost κ

Looking at Figure 1, as the instance costs κ increase, we see the expected monotonic decrease of the profit for each of the four strategies. Note that, although the ratio between fixed and load-dependent costs stays constant, the profit of the fixed-price strategy decreases faster than for the other three strategies; the intuition for this is that the ratio between costs and revenue is highest in the fixed-price market. Importantly, we see that the hybrid strategy always achieves the highest profit among all four strategies which illustrates the main point of our paper. For very low instance costs, its profit almost coincides with the profit of the fixed-price strategy. But already for moderate instance costs, the hybrid strategy leads to a significant profit increase. We also see that, when costs are very high, then the provider's optimal strategy is to set the price p_F high enough such that no user joins the fixed-price market (i.e., the equilibrium degenerates into a pure spot equilibrium).[17] Finally, we see that the hybrid strategy with $p_F = p_F^{*0}$ obtains almost the same profit as the (non-restricted) hybrid strategy (until the point where the hybrid strategy stops offering a fixed-price market). This illustrates that the provider does not need to compute a new price for her fixed-price market to achieve a sizeable profit increase via a hybrid strategy.

Overall, Example 1 illustrates our main result (Theorem 5.4): there exists a hybrid strategy (e.g., the hybrid strategy with $p_F = p_F^{*0}$)which simultaneously increases the provider's profit and leads to a Pareto improvement for the users.[18]

6.3 Example 2: Varying the Cost Ratio $\frac{\kappa_F}{\kappa_L}$

For the second example, we vary the ratio between fixed costs κ_F and load-dependent costs κ_L while keeping the sum constant at $\kappa = 0.1$. We again compare the same four strategies as in Example 1.

Figure 2 shows the profit for each strategy, varying the ratio $\frac{\kappa_F}{\kappa_L}$ between 0 and 20 on the x-axis. Thus, at 0, all costs are load-dependent, while at 20, the fixed costs are 20 times as high as the load-dependent costs. As we can see in Figure 2, at $\frac{\kappa_F}{\kappa_L} = 0$, offering no spot instances is optimal. This is expected because at 0, all costs are load-dependent, and therefore the main benefit of using idle instances (i.e., reducing fixed costs) is gone. As the fixed costs increase, the profits of the top three strategies (which at this point offer mostly fixed-price instances) first sharply decrease, even though the instance costs κ stay constant. This happens because the fixed-price instances now also incur a fraction of the instance costs while standing idle, whereas at 0, they only incurred costs while running. Conversely, spot instances get more attractive as $\frac{\kappa_F}{\kappa_L}$ increases. The hybrid strategies therefore start using spot instances to counteract the increased costs for offering fixed-price instances, which can be seen by the flattening of the solid red and dotted black lines.

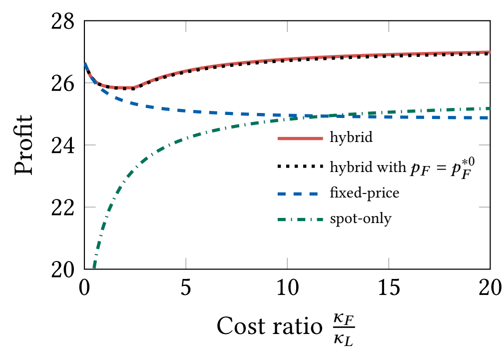

Fig. 2 Profit under different strategies while varying the cost ratio $\frac{\kappa_F}{\kappa_L}$

At the point where the fixed costs are about 3 times as large as the load-dependent costs, increasing $\frac{\kappa_F}{\kappa_L}$ further leads to a profit increase for both hybrid strategies. This happens because, beyond this point, both strategies offer enough spot instances such that the profit increase of the spot market dominates the profit decrease of the fixed-price market. We also again see that the hybrid strategy with $p_F = p_F^{*0}$ achieves close to optimal profits.

Recall that, in practice, fixed costs are usually about 10-20 times as large as load-dependent costs (i.e., $10 \leq \frac{\kappa_F}{\kappa_L} \leq 20$). Thus, this example suggests that, even when the provider incurs relatively large costs for spot instances (e.g., $\frac{\kappa_F}{\kappa_L} = 3$), a cloud provider can expect to achieve a sizable profit increase from offering a spot market in addition to her existing fixed-price market.

[17]Note that this result does not imply that in practice, the provider would not offer a fixed-price market, as there are always some users who would never join the spot market for a variety of reasons (e.g., because their jobs should never be preempted). Instead, this result shows that the provider's optimal strategy incentivizes all users who are *willing to consider* joining a spot market to do so (by setting a relatively high price).

[18]Note that a Pareto improvement for the users obviously implies a welfare increase. We demonstrate this welfare increase in Appendix C.

6.4 Varying the Provider Strategy

For the next two examples, we fix the fixed costs at $\kappa_F = 0.09$ and the load-dependent costs at $\kappa_L = 0.01$. We show how the profit changes, depending on different provider strategies $\rho = (p_F, l_S)$.

Example 3: Varying the Price p_F. Figure 3 shows the optimal profit under two strategies that are restricted to the price p_F shown on the x-axis. The hybrid strategy (solid red line) offers the optimal number of spot instances l_S given p_F. The fixed-price strategy (dashed blue line) only offers a fixed-price market with price p_F.

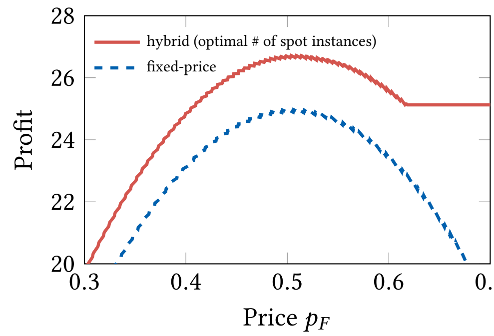

Fig. 3 Profit under different strategies when restricted to different prices p_F

Recall that the provider's profit depends on three factors: how many users join each market, how much they pay, and the provider's costs. As we can see in Figure 3, when the price p_F increases, the profit for both strategies at first also increases. This happens because the users' average payments under both strategies go up; this is also true in the spot market because increasing p_F pushes users from the fixed-price market into the spot market, which increases payments. However, since more users balk when the price increases, the profit only goes up until the rise in payments is dominated by the loss of users, whereafter the profit starts to fall again. Once the fixed-price market becomes too expensive for the users, the user equilibrium under the hybrid strategy degenerates into a pure spot equilibrium, as can be seen by the solid red line becoming constant beyond $p_F = 0.62$.

Example 4: Varying the Number of Spot Instances l_S. Figure 4 shows the optimal profit under three strategies that are all restricted to offering the number of spot instances l_S shown on the x-axis. The hybrid strategy (solid red line) uses the optimal price p_F given l_S. The hybrid strategy with $p_F = p_F^{*0}$ (dotted black line) also offers l_S spot instances but uses the price that would be optimal when only offering a fixed-price market. As a reference, we also include the spot-only strategy (dash-dotted green line) which only offers a spot market with l_S spot instances.

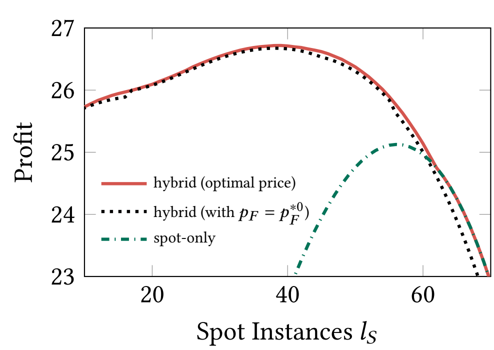

Fig. 4 Profit under different strategies when restricted to different spot market sizes l_S

As we see in Figure 4, as the number of spot instances l_S increases, the profit for the two hybrid strategies at first also increases. The key intuition for this is that, even though the average payments in the spot market may decrease, now more users join the spot market, where (on average) they generate a higher profit than they previously did (either in the fixed-price market, or by balking). For both strategies, the optimal profit is achieved at $l_S = 39$, whereafter the profit starts to decrease.[19] Note that at $l_S = 62$, we again observe a point where the user equilibrium under the hybrid strategy degenerates into a pure spot equilibrium.

Finally, Figure 4 shows that for all numbers of spot instances, the profit achieved by the hybrid strategy with $p_F = p_F^{*0}$ is relatively close to the profit achieved with the optimal price. This suggests that, even when the provider cannot fully optimize her strategy, keeping the price at $p_F = p_F^{*0}$ is a viable approach for a cloud provider when offering a spot market.

7 CONCLUSION

In this paper, we have studied whether a cloud provider can benefit from selling idle instances on a spot market. Our main result is an easy-to-check condition under which a cloud provider can simultaneously increase her profit and achieve a Pareto increase for the users by offering a spot market in addition to a fixed-price market.

In contrast to prior work, we have modeled the provider's fixed and load-dependent costs as well as the users' costs for preemption. As these costs are an important factor for the profitability of any cloud market, modeling them was essential to make valid statements about the provider's profit optimization problem.

Our results have significant implications for practical market design. They suggest that, when our condition is satisfied, offering a spot market alongside her fixed-price market is advantageous for a cloud provider. Our condition is relatively mild and should be satisfied for most providers. Furthermore, even when a provider cannot compute her profit-optimal strategy, there are viable approaches to still achieve a profit increase by offering a spot market. Considering that the preemption costs are one of the main factors determining the profitability of a spot market, we encourage providers to continue to evolve their technology such that the losses incurred from re-starting a job are further reduced.

An interesting direction for future work would be to study how selling idle instances on a spot market compares to alternative market designs, such as the provider selling her idle instances on a preemptible *fixed-price* market. It is not immediately clear whether such a preemptible fixed-price market would be able to generate more or less profit than a spot market (with a reserve price). One would have to account for the differences in average payments, the market cannibalization towards the fixed-price market, and the costs produced by preemptions. Further, possible competitive advantages of one market over the other (i.e., differences in user satisfaction) would have to be taken into account. A complete analysis of this trade-off would be very valuable.

REFERENCES

Abhishek V, Kash IA, Key P (2012) Fixed and market pricing for cloud services. *2012 Proceedings IEEE INFOCOM Workshops*, 157–162.

Abhishek V, Kash IA, Key P (2017) Fixed and market pricing for cloud services, CoRR abs/1201.5621. Extended version of Abhishek et al. (2012).

Agmon Ben-Yehuda O, Ben-Yehuda M, Schuster A, Tsafrir D (2013) Deconstructing amazon ec2 spot instance pricing. *ACM Transactions on Economics and Computation* 1(3):16:1–16:20.

[19]This profit decrease is caused by three factors becoming dominant: the reduction in average payments, the need for a relatively larger buffer in the fixed-price market, and the unreliability of additional spot instances.

Azar Y, Kalp-Shaltiel I, Lucier B, Menache I, Naor J, Yaniv J (2015) Truthful online scheduling with commitments. *Proceedings of the 16th ACM Conference on Economics and Computation*, 715–732.

Babaioff M, Mansour Y, Nisan N, Noti G, Curino C, Ganapathy N, Menache I, Reingold O, Tennenholtz M, Timnat E (2017) Era: a framework for economic resource allocation for the cloud. *Proceedings of the 26th International Conference on World Wide Web Companion*, 635–642.

Banerjee S, Riquelme C, Johari R (2015) Pricing in ride-share platforms: A queueing-theoretic approach. *Proceedings of the 16th ACM Conference on Economics and Computation* 639.

Barroso LA, Hölzle U, Ranganathan P (2018) The datacenter as a computer: Designing warehouse-scale machines, third edition. *Synthesis Lectures on Computer Architecture* 13(3):i–189.

Boodaghians S, Fusco F, Leonardi S, Mansour Y, Mehta R (2019) Online revenue maximization for server pricing. *CoRR* abs/1906.09880.

Buzen JP, Bondi AB (1983) The response times of priority classes under preemptive resume in m/m/m queues. *Operations Research* 31(3):456–465.

Cohen MC, Keller PW, Mirrokni V, Zadimoghaddam M (2019) Overcommitment in cloud services: Bin packing with chance constraints. *Management Science* 65(7):3255–3271.

Cooper RB (1981) *Introduction to queueing theory* (Amsterdam, NL: North Holland Publishing Co.).

Cortez E, Bonde A, Muzio A, Russinovich M, Fontoura M, Bianchini R (2017) Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. *Proceedings of the 26th Symposium on Operating Systems Principles*, 153–167.

Desai PS (2001) Quality segmentation in spatial markets: When does cannibalization affect product line design? *Marketing Science* 20(3):265–283.

Dierks L, Kash IA, Seuken S (2019) On the cluster admission problem for cloud computing. *Proceedings of the 14th Workshop on the Economics of Networks, Systems and Computation.*

Dierks L, Seuken S (2019) Cloud pricing: The spot market strikes back (extended abstract). *Proceedings of the 20th ACM Conference on Economics and Computation*, 593.

Dierks L, Seuken S (2020) The competitive effects of variance-based pricing. *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, Forthcoming.

Gao J, Iyer K, Topaloglu H (2019) When fixed price meets priority auctions: Competing firms with different pricing and service rules. *Stochastic Systems* 9(1):47–80.

Hassin R (2016) *Rational Queueing* (Boca Raton, FL: CRC Press).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems* (Norwell, MA: Kluwer Academic Publishers).

Hoy D, Immorlica N, Lucier B (2016) On-demand or spot? selling the cloud to risk-averse customers. *International Conference on Web and Internet Economics*, 73–86.

Islam M, Ren X, Ren S, Wierman A (2018) A spot capacity market to increase power infrastructure utilization in multi-tenant data centers. *2018 IEEE International Symposium on High Performance Computer Architecture*, 776–788.

Jyothi SA, Curino C, Menache I, Narayanamurthy SM, Tumanov A, Yaniv J, Mavlyutov R, Goiri I, Krishnan S, Kulkarni J, et al. (2016) Morpheus: Towards automated slos for enterprise clusters. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 117–134.

Kash IA, Key P (2016) Pricing the cloud. *IEEE Internet Computing* 20(1):36–43.

Kash IA, Key P, Suksompong W (2017) Simple pricing schemes for the cloud. *International Conference on Web and Internet Economics*, 311–324.

Maharjan S, Zhu Q, Zhang Y, Gjessing S, Basar T (2013) Dependable demand response management in the smart grid: A stackelberg game approach. *IEEE Transactions on Smart Grid* 4(1):120–132.

Maskin E, Riley J (1984) Monopoly with incomplete information. *The RAND Journal of Economics* 15(2):171–196.

Moorthy KS (1984) Market segmentation, self-selection, and product line design. *Marketing Science* 3(4):288–307.

Mussa M, Rosen S (1978) Monopoly and product quality. *Journal of Economic Theory* 18(2):301 – 317.

Myerson RB (1981) Optimal auction design. *Mathematics of operations research* 6(1):58–73.

Shaked A, Sutton J (1982) Relaxing price competition through product differentiation. *The review of economic studies* 3–13.

Shi W, Zhang L, Wu C, Li Z, Lau FC (2014) An online auction framework for dynamic resource provisioning in cloud computing. *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, 71–83.

Subramanya S, Rizk A, Irwin D (2016) Cloud spot markets are not sustainable: The case for transient guarantees. *8th USENIX Workshop on Hot Topics in Cloud Computing*.

Takagi H (2008) *Spectrum Requirement Planning in Wireless Communications*, chapter Appendix A: Derivation of Formulas by Queueing Theory, 199–218 (John Wiley & Sons, Ltd).

Varian HR (1989) Price discrimination. *Handbook of industrial organization* 1:597–654.

Wolff RW (1982) Poisson arrivals see time averages. *Operations Research* 30(2):223–231.

Yan Y, Gao Y, Chen Y, Guo Z, Chen B, Moscibroda T (2016) Tr-spark: Transient computing for big data analytics. *Proceedings of the 7th ACM Symposium on Cloud Computing*, 484–496.

Zaharia M, Borthakur D, Sen Sarma J, Elmeleegy K, Shenker S, Stoica I (2010) Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. *Proceedings of the 5th European conference on Computer systems*, 265–278.

Zhang H, Jiang H, Li B, Liu F, Vasilakos AV, Liu J (2016) A framework for truthful online auctions in cloud computing with heterogeneous user demands. *IEEE Transactions on Computers* 65(3):805–818.

Zhang L, Li Z, Wu C (2014) Dynamic resource provisioning in cloud computing: A randomized auction approach. *2014 IEEE Conference on Computer Communications*, 433–441.

Zheng L, Joe-Wong C, Brinton CG, Tan CW, Ha S, Chiang M (2016) On the viability of a cloud virtual service provider. *ACM SIGMETRICS Performance Evaluation Review* 44(1):235–248.

Zhou R, Li Z, Wu C, Huang Z (2017) An efficient cloud market mechanism for computing jobs with soft deadlines. *IEEE/ACM Transactions on Networking* 25(2):793–805.

A ADDITIONAL STATEMENTS AND PROOFS

A.1 Lemma A.1

A number of our results make use of the following technical Lemma which extends Lemma 7 from Abhishek et al. (2017) to more general g_i and $w()$, which includes our model.

Lemma A.1. *Fix a provider strategy ρ with $l_s > 0$. Let $(x_1, \ldots, x_k)$ be a weakly decreasing sequence. For $i > k$ let $g_i(x_i), \ldots, g_n(x_n)$ be given such that $g_j(x_j)$ is a weakly increasing and semi-differentiable scalar function with left derivative at most $w(\mathcal{S}, x_i, \rho, (x_1, \ldots, x_k, x_i \ldots, x_i, 0_{i+1}, \ldots, 0_n))$. (Here, the notation $(x_1, \ldots, x_k, x_i \ldots, x_i, 0_{i+1}, \ldots, 0_n)$ means that all entries k' with $k < k' \leq i$ are set to x_i.) Further, assume that for all $j \geq i$ it holds that $g_j(\mu v_j) \leq v_j$, as well as $g_j(x) \leq g_i(x)$ for all $x \in \mathbb{R}$. Then there exist unique $x_i, \ldots, x_n$ such that for any strategy profile*

$$\sigma' = \overrightarrow{x} \quad = \quad (x_1, \ldots, x_k, x_i, \ldots, x_i, x_{i+1}, \ldots, x_n) \tag{50}$$

where any user of class j joins the spot market if and only if his waiting cost c is below x_j, it holds that

$$\int_0^{x_j} w(\mathcal{S}, c, \rho, \overrightarrow{x})dc \quad = \quad g_j(x) \quad \textit{for all } j \geq i. \tag{51}$$

Proof. To see this, assume that the claim holds for $i+1$. Then for any $z \in [0, x_k]$ there exists

$$\sigma(z) = \overrightarrow{x}(z) \quad = \quad (x_1, \ldots, x_k, z, \ldots, z, x_{i+1}(z), \ldots, x_n(z)) \tag{52}$$

satisfying Equation (51) for any $j \geq i+1$. We now show that there exists a unique z^* such that $w(z^*) = \int_0^{z^*} w(\mathcal{S}, c, \rho, \sigma(z^*))dz$. As a first step, we show that $w(z) = \int_0^{z} w(\mathcal{S}, c, \rho, \sigma(z))dz$ is increasing in z. Since for any fixed x, $\int_0^{x} w(\mathcal{S}, c, \rho, \overrightarrow{x})dc$ is increasing in each x_j, it follows that $x_{i+1}(z)$ is decreasing in z. Then for any $\hat{z} > z$ it holds by the induction assumption that

$$\int_0^{x_{i+1}(z)} w(\mathcal{S}, c, \rho, \overrightarrow{x}(\hat{z}))dc \quad \geq \quad g_i(x_{i+1}(z)) \tag{53}$$

and therefore

$$\begin{aligned}
& \frac{\int_0^{\hat{z}} w(\mathcal{S}, c, \rho, \overrightarrow{x}(\hat{z}))dc - \int_0^{z} w(\mathcal{S}, c, \rho, \overrightarrow{x}(z))dc}{\hat{z} - z} && (54)\\
\geq \quad & \frac{g_{i+1}(x_{i+1}(z)) - g_{i+1}(x_{i+1}(z))}{\hat{z} - z} + \frac{\int_{x_{i+1}(z)}^{\hat{z}} w(\mathcal{S}, c, \rho, \overrightarrow{x}(\hat{z}))dc - \int_{x_{i+1}(z)}^{z} w(\mathcal{S}, c, \rho, \overrightarrow{x}(z))dc}{\hat{z} - z} && (55)\\
= \quad & \frac{\int_z^{\hat{z}} w(\mathcal{S}, c, \rho, \overrightarrow{x}(\hat{z}))dc + \int_{x_{i+1}(z)}^{z} w(\mathcal{S}, c, \rho, \overrightarrow{x}(\hat{z})) - w(\mathcal{S}, c, \rho, \overrightarrow{x}(z))dc}{\hat{z} - z} && (56)\\
> \quad & \frac{\int_z^{\hat{z}} w(\mathcal{S}, c, \rho, \overrightarrow{x}(\hat{z}))dc}{\hat{z} - z} && (57)\\
> \quad & w(\mathcal{S}, z, \rho, \overrightarrow{x}(\hat{z})). && (58)
\end{aligned}$$

Equation (57) is a direct result of the fact that waiting times of higher priority jobs do not depend on the number of lower priority jobs in the queue. By taking the limit $\hat{z} \longrightarrow z$ it follows that $w'(z) > w(\mathcal{S}, z, \rho, \overrightarrow{x}(\hat{z})) \geq w(\mathcal{S}, z, \rho, (x_1, \ldots, x_k, z \ldots, z, 0_{i+1}, \ldots, 0_n)$. As $w(0) = 0$ and $w(\mu v_i) \geq v_i \geq g_i(\mu v_i)$, the claim for i follows.

To show the induction base case $i = n$, we introduce a dummy variable x_{n+1} with $g_{n+1} = 0$. This means that for any z it trivially holds that $x_{n+1}(z) = 0$ and the statement for n follows. □

A.2 Proof of Proposition 4.2

Any equilibrium strategy for users of class i is trivially a threshold strategy because for any fixed strategy profile σ, the payoff of a user of class i, i.e., $\pi_i^c(\mathcal{S}, c, \rho, \sigma)$, is monotone decreasing in the waiting cost c. By setting $g_i(c) = v_i$, the existence of the unique cutoff vector $\overrightarrow{c}^S$ then follows directly from Lemma A.1. By the incentive compatibility of the payment rule, no user would deviate from $\sigma^* = \overrightarrow{c}^S$.

A.3 Proof of Proposition 4.3

Users with waiting close enough to zero always prefer the spot market, no matter how much time they lose compared to the fixed-price market. Since some users join the fixed-price market and both waiting time and payment are continuous in the bid c, for any potential equilibrium strategy profile σ^*, there has to be a lowest point c_1^P such that

$$c_1^P(T + \frac{1}{\mu}) + p_F\frac{1}{\mu} = \int_0^{c_1^P} w(\mathcal{S}, x, \rho, \sigma)dx. \tag{59}$$

Since it holds $\frac{d}{dc}\pi_i^c(\mathcal{S}, c, \rho, \sigma) = w(\mathcal{S}, c, \rho, \sigma) \geq \frac{1}{\mu}\frac{1}{1-\tau\psi_E(l_S)} > T + \frac{1}{\mu} = \frac{d}{dc}\pi_i^c(\mathcal{F}, c, \rho, \sigma)$, the higher a users waiting cost, the worse the spot market compared to the fixed-price market and there cannot exist any $c > c_1^P$ for which users prefer the spot market, i.e. with

$$c(T + \frac{1}{\mu}) + p_F\frac{1}{\mu} > \int_0^{c} w(\mathcal{S}, x, \rho, \sigma)dx. \tag{60}$$

Thus, no user with waiting cost greater than c_1^P joins the spot market. This means that the spot market can be fully defined by the actions of players with $c \leq c_1^P$. Recall that $\overrightarrow{c}^P$ denotes 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 that

$$\min\left\{\frac{p_F}{\mu} + c_i^P(\frac{1}{\mu} + T), v_i\right\} \quad = \quad \int_0^{c_i^P} w(\mathcal{S}, x, \rho, \overrightarrow{c}^P)dx \quad \forall i \in \{1, \ldots, n\}, \tag{61}$$

which has a unique solution by Lemma A.1. Every job of class i with $c < c_i^P$ joins the spot market, and every job with $c_i^P < c < \frac{\mu v_i - p_F}{\mu T + 1}$ joins the fixed-price market and those with $c > \frac{\mu v_i - p_F}{\mu T+1}$ balk. Setting $c_i^B = max(c_i^P, \frac{\mu v_i - p_F}{\mu T+1})$, it is clear that every solution of (17) and (18) solves (61) and vice-versa.

A.4 Lemma A.2

The following Lemma establishes the broad equilibrium structure when the spot market is faster at the highest bids and shows the existence of two cutoff points in equilibrium between which almost all users join the fixed-price market. It is used in the proof of Proposition 4.4.

Lemma A.2. *For any provider strategy $\rho = (p_F, l_S)$, in any BNE of the user game where some users join the fixed-price market and where the spot market is faster for the highest bids, i.e., $T > \frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right)$, there exists an interval $[c^L, c^U]$, such that almost all users (i.e., all besides possibly a set of measure zero that does not influence system dynamics) with waiting costs $c \in [c^L, c^U]$ join the fixed-price market or balk. For bids $c \in [c^L, c^U]$, the total waiting time and the payment in equilibrium are* *equal in each market, i.e.:*

$$m(\mathcal{S}, c, \rho, \sigma^*) = m(\mathcal{F}, c, \rho, \sigma^*) = p_F \frac{1}{\mu} \tag{62}$$

$$w(\mathcal{S}, c, \rho, \sigma^*) = w(\mathcal{F}, c, \rho, \sigma^*) = T + \frac{1}{\mu} \tag{63}$$

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

$$\pi_i(\mathcal{S}, c, \rho, \sigma^*) > \pi_i(\mathcal{F}, c, \rho, \sigma^*) \;\; \forall i \in \{1, \ldots, n\}, \tag{64}$$

and these users join the spot market or balk.

Proof. For a job with the highest bid that does not balk, the spot market is faster than the fixed-price market because $T > \frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right)$. Any user with such a waiting cost is therefore willing to pay more in the spot market than in the fixed-price market. This means that he strictly prefers the spot market in equilibrium.

Let c^L be the lowest waiting cost for which a job prefers the fixed-price market over the spot market or is indifferent between the two, and let c^U be the highest such waiting cost. c^L and c^U have to exist for any equilibrium where users join both markets. We now show by contradiction that the user's payment in the spot market has to be weakly larger than in the fixed-price market for bids above c^L and that the spot market is weakly slower than the fixed-price market for bids below c^U.

Assume there exists a waiting cost $\bar{c} > c^L$ at which a user would prefer the spot market or be indifferent between spot and fixed-price market, and for which the payment in the spot market is less in expectation than in the fixed-price market, i.e., for which $m(\mathcal{S}, \bar{c}, \rho, \sigma) < m(\mathcal{F}, \bar{c}, \rho, \sigma)$. Then

$$c^L w(\mathcal{S}, c^L, \rho, \sigma) + m(\mathcal{S}, c^L, \rho, \sigma) \tag{65}$$

$$\leq c^L w(\mathcal{S}, \bar{c}, \rho, \sigma) + m(\mathcal{S}, \bar{c}, \rho, \sigma) \tag{66}$$

$$= \frac{c^L}{\bar{c}} \left(\bar{c} w(\mathcal{S}, \bar{c}, \rho, \sigma) + \frac{\bar{c}}{c^L} m(\mathcal{S}, \bar{c}, \rho, \sigma)\right) \tag{67}$$

$$\leq \frac{c^L}{\bar{c}} \left(\bar{c} w(\mathcal{F}, \bar{c}, \rho, \sigma) + m(\mathcal{F}, \bar{c}, \rho, \sigma) + (\frac{\bar{c}}{c^L} - 1) m(\mathcal{S}, \bar{c}, \rho, \sigma)\right) \tag{68}$$

$$< \frac{c^L}{\bar{c}} \left(\bar{c} w(\mathcal{F}, \bar{c}, \rho, \sigma) + m(\mathcal{F}, \bar{c}, \rho, \sigma) + (\frac{\bar{c}}{c^L} - 1) m(\mathcal{F}, \bar{c}, \rho, \sigma)\right) \tag{69}$$

$$= \frac{c^L}{\bar{c}} \left(\bar{c} w(\mathcal{F}, \bar{c}, \rho, \sigma) + \frac{\bar{c}}{c^L} m(\mathcal{F}, \bar{c}, \rho, \sigma)\right) \tag{70}$$

$$= c^L w(\mathcal{F}, \bar{c}, \rho, \sigma) + m(\mathcal{F}, \bar{c}, \rho, \sigma) \tag{71}$$

$$= c^L w(\mathcal{F}, c^L, \rho, \sigma) + m(\mathcal{F}, c^L, \rho, \sigma) \tag{72}$$

(66) holds because the pricing rule is BNIC; (68) holds because at waiting cost $\bar{c}$ the spot market's overall cost has to be lower than the fixed-price market in order for the user to join it. Finally, (69) holds because we assumed the spot market to be cheaper with bid $\bar{c} > c^L$. A job with waiting cost c^L would therefore also strictly prefer the spot market, a contradiction.

Assume there exists a waiting cost $\bar{c} < c^U$ at which a user would prefer the spot market or be indifferent between spot and fixed-price market, and for which the waiting time is lower in the spot market than in the fixed-price market, i.e., for which $w(\mathcal{S}, \bar{c}, \rho, \sigma) < w(\mathcal{F}, \bar{c}, \rho, \sigma)$. Then similarly

$$
\begin{aligned}
& c^U w(\mathcal{S}, c^U, \rho, \sigma) + m(\mathcal{S}, c^U, \rho, \sigma) && (73)\\
\leq\ & c^U w(\mathcal{S}, \bar{c}, \rho, \sigma) + m(\mathcal{S}, \bar{c}, \rho, \sigma) && (74)\\
=\ & \bar{c} w(\mathcal{S}, \bar{c}, \rho, \sigma) + m(\mathcal{S}, \bar{c}, \rho, \sigma) + (c^U - \bar{c}) w(\mathcal{S}, \bar{c}, \rho, \sigma) && (75)\\
<\ & \bar{c} w(\mathcal{F}, \bar{c}, \rho, \sigma) + m(\mathcal{F}, \bar{c}, \rho, \sigma) + (c^U - \bar{c}) w(\mathcal{F}, \bar{c}, \rho, \sigma) && (76)\\
=\ & c^U w(\mathcal{F}, \bar{c}, \rho, \sigma) + m(\mathcal{F}, \bar{c}, \rho, \sigma) && (77)
\end{aligned}
$$

A user with c^U would therefore also strictly prefer the spot market, a contradiction.

Therefore, for all $c \in [c^L, c^U]$ the spot market can neither be faster nor cheaper than the fixed-price market. If any users with waiting cost $c \in [c^L, c^U]$ join the spot market they have to be indifferent between both markets. Thus, for any σ^* to be a BNE, this means that at most a set of measure zero of such users can join the spot market, and thus the total waiting time and payment stay constant over the whole interval. The statement of the lemma immediately follows. □

A.5 Proof of Proposition 4.4

It follows from Lemma A.2 that any equilibrium strategy profile is of the form $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$: Let the points c^L and c^U be as given by Lemma A.2 and let $\overrightarrow{c}^H$ denote the waiting costs above which users of each class cannot obtain a positive payoff anymore and balk. Define the cutoff vectors $\overrightarrow{c}^L, \overrightarrow{c}^U$ as $c_i^L = \min\left\{c^L, c_i^H\right\}$ and $c_i^U = \min\left\{c^U, c_i^H\right\}$. Note that this implies that $c_1^L = c^L$ and $c_1^U = c^U$ (because at least some users from class 1 go to the portion of the spot market that is faster than the fixed-price market). Then Equations (19), (20) and (21) immediately follow from Lemma A.2:

(1) Equation (19): The payoff at c^L has to be the same for joining the fixed-price or spot market.
(2) Equation (20): The payoff at c^U also has to be the same in the fixed-price and spot market.
(3) Equation (21): Users do not balk as long as their value for joining one of the markets is greater than 0.

We now show that this system of equations always has a unique solution using a constructive approach. For this, we first introduce some additional notation.

Given provider strategy ρ, we know that in order to satisfy Lemma A.2, jobs that pay more in the spot market than in the fixed-price marked need to arrive at a rate such that jobs with waiting cost c^L have to queue for exactly T. Denote this arrival rate by $\lambda(T, \rho)$. We now further overload our previous notation for a user strategy profile: for any vector $\hat{c} = (\hat{c}_1, \dots, \hat{c}_n)$ with $\hat{c}_i \geq \hat{c}_j$ for $i < j$, we let $\sigma = (\hat{c}, \lambda(T, \rho))$ denote a user strategy profile where every user of class i with waiting cost $c < \hat{c}_i$ joins the spot market, but everyone else balks even if he could obtain a positive payoff in one of the markets. Additionally, we assume that *dummy jobs* of maximal priority arrive with rate $\lambda(T, \rho)$ into the spot market. Thus, $w(\mathcal{S}, \hat{c}_1, \rho, (\hat{c}, \lambda(T, \rho))) = T$ by definition. Combined with Lemma A.2, this notational trick allows us to "simulate" the impact users with waiting costs between $\overrightarrow{c}^U$ and $\overrightarrow{c}^H$ have on all other users, without yet knowing $\overrightarrow{c}^U$ and $\overrightarrow{c}^H$. We now need to determine which classes of users join the fixed-price market (in the sense that there exists a c such that a user from that class with waiting cost c joins the fixed-price market) and which do not. Once we know that, we can split the system of equations into two parts that can be solved consecutively.

To check whether the k'th class joins the fixed-price market, i.e., whether $c^H_k > c^L$, we denote by $\hat{c} = (\hat{c}_1, \dots, \hat{c}_n)$ (where each $\hat{c}_i \in [0, \mu v_i]$) the cutoff vector solving the following:

$$0 = v_i - \int_0^{\hat{c}_i} w(\mathcal{S}, x, \rho, (\hat{c}, \lambda(T, \rho)))dx \quad \forall i \geq k \tag{78}$$

$$\hat{c}_i = \hat{c}_k \quad \forall i < k \tag{79}$$

This has a unique solution according to Lemma A.1. Note that the cutoff vector $\hat{c}$ here carries an implicit dependence on k, while $\hat{c}_k$ denotes its k'th element.
If it now holds that

$$p_F \frac{1}{\mu} > m(\mathcal{S}, \hat{c}_k, \rho, (\hat{c}, \lambda(T, \rho))) \tag{80}$$

$$= \int_0^{\hat{c}_k} w(\mathcal{S}, x, \rho, (\hat{c}, \lambda(T, \rho)))dx - w(\mathcal{S}, \hat{c}_k, \rho, (\hat{c}, \lambda(T, \rho))), \tag{81}$$

then $c^L \leq \hat{c}_k$ would mean that not enough users join the spot market to reach the price of the fixed-price market at the cutoff point c^L. This means that for any such c^L, the system of equations defining the equilibrium cutoff vectors (i.e., Equations (19), (20) and (21)) cannot be satisfied for any choice of c^U and $\overrightarrow{c}^H$. It follows that in equilibrium, $c^L > \hat{c}_k$. Thus, in equilibrium, no user of class k joins the fixed-price market, i.e., $c^H_k < c^L$.

Conversely, if Equation (80) does not hold, then setting $c^L > \hat{c}_k$ would mean that too many users join the spot market, and the payment in the spot market at the cutoff point c^L is larger than the payment in the fixed-price market. Thus, in any equilibrium, some users of class k join the fixed-price market and $c^L \leq \hat{c}_k$. As $m(\mathcal{S}, \hat{c}_k, \rho, (\hat{c}, \lambda(T, \rho)))$ is monotone decreasing in k, it follows that either there exists a lowest class k^* such that (81) is satisfied and for which no user joins the fixed-price market, or all classes join the fixed-price market in which case we set $k^* = n + 1$. Splitting the system of equations that defines the equilibrium strategy profile at this k^*, Lemma A.1 yields that

$$0 = \hat{c}_i(T + \frac{1}{\mu}) + p_F \frac{1}{\mu} - \int_0^{\hat{c}_i} w(\mathcal{S}, x, \rho, (\hat{c}, \lambda(T, \rho)))dx \quad \forall i < k^* \tag{82}$$

$$0 = v_i - \int_0^{\hat{c}_i} w(\mathcal{S}, x, \rho, (\hat{c}, \lambda(T, \rho)))dx \quad \forall i \geq k^* \tag{83}$$

has a unique solution with $\hat{c}_{k^*} < \hat{c}_{k^*-1}$ and for any equilibrium $(\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$ it holds that $c^L = \hat{c}_{k^*-1}$ and $c^H_i = \hat{c}_i$ for all $i \geq k^*$.

Given the solution to (82) and (83), we can now equivalently find the highest class k^{**} that joins the upper portion of the spot market (i.e. for which $c^U_{k^{**}} < c^H_{k^{**}}$). To this end, fix any $k < k^*$. Again carrying an implicit dependence on k, we define temporary cutoff vectors $\hat{c}^U$ and $\hat{c}^H$. Set $\hat{c}^H_i = c^L$ for all $k < i < k^*$ and $\hat{c}^H_i = \overrightarrow{c}^H_i$ for all $i \geq k^*$. Further let $\hat{c}^U$ and $\hat{c}^H_i$ for $i \leq k$ be given as the solution to

$$0 = v_{k+1} - \hat{c}^U_{k+1}(T + \frac{1}{\mu}) + p_F \frac{1}{\mu} \tag{84}$$

$$0 = v_i - \int_0^{\hat{c}^H_i} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^L, \hat{c}^U, \hat{c}^H))dx \quad \forall i \leq k \tag{85}$$

$$\hat{c}^U_i = min(\hat{c}^U_k, \hat{c}^H_i) \quad \forall i \neq k + 1. \tag{86}$$

This system of equations has a unique solution for every $k < k^*$ according to Lemma A.1. Intuitively, $(\overrightarrow{c}^L, \hat{c}^U, \hat{c}^H)$ can be seen as the strategy profile where users with waiting costs below c^L play the equilibrium strategy, no user joins the fixed-price market, and users with waiting costs above the point where class $k+1$ would obtain zero payoff in the fixed-price market join the spot market (if their payoff for doing so is positive). This means that under this strategy profile *more* users join the spot market than would under any potential equilibrium strategy profile where $c^U > \hat{c}^U_{k+1}$. Analogous to k^*, there now exists a lowest class $k^{**} < k^*$, such that if only jobs of classes $i \leq k^{**}$ join the upper part of the spot market, there are still enough users that potentially (i.e., as long as the fixed-price market isn't better) join the spot market, such that the waiting time in the spot market at $\hat{c}^U_k$ is at least as high as the waiting time in the fixed-price market, i.e., k^{**} is the smallest k for which it holds that

$$T + \frac{1}{\mu} \quad \leq \quad w(\mathcal{S}, \hat{c}^U_k, \rho, (\overrightarrow{c}^L, \hat{c}^U, \hat{c}^H)). \tag{87}$$

Conversely, if users of classes higher than k^{**} would join the upper portion of the spot market (i.e. $c^U \leq \hat{c}^U_{k^{**}+1}$) then the waiting time in the spot market at c^U is always above $T + \frac{1}{\mu}$. Consequently, we can calculate $\overrightarrow{c}^H_i$ for $k^{**} < i < k^*$ as the solution to

$$0 \quad = \quad v_i - \overrightarrow{c}^H_i(T + \frac{1}{\mu}) + p_F\frac{1}{\mu}. \tag{88}$$

Then finally, we can calculate c^U and $\overrightarrow{c}^H_i$ for $i \leq k^{**}$ as the solution to

$$0 \quad = \quad c^U(T + \frac{1}{\mu}) + p_F\frac{1}{\mu} - \int_0^{c^U} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H))dx \tag{89}$$

$$0 \quad = \quad v_i - \int_0^{c_i^H} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H))dx \quad \forall i \leq k^{**}, \tag{90}$$

which, given c^L and c^H_i for all $i > k^{**}$ now also has a unique solution according to Lemma A.1.

As each of the successively solved subsystems of equations was, at the time it was solved, independent of the then unsolved parts, $(\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$ solves the whole system of equations.

A.6 Proof of Lemma 4.6

For $l_S > 0$, all users with waiting costs in some neighborhood around zero prefer the spot market. Let ρ be a provider strategy with $l_S > 0$ that is proper. Assume we have an equilibrium where no one joins the fixed-price 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 c^S_1 (if $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) \geq T$) or c' (if $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) < T$) could then obtain a better payoff by switching to the fixed-price market, leading to a contradiction. Any BNE therefore has some users joining the fixed-price market.

Now assume ρ is not proper. We first show the statement for $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) \geq T$ (i.e., when the spot market is always slower than the fixed-price market). Proposition 4.3 gives us that any equilibrium where some users join the fixed-price market is of the form $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$. At any waiting cost c for which users of some class i balk under $\sigma = (\overrightarrow{c}^S)$ but join the spot market under $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$, their payoff in the spot market needs to be higher under $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$, i.e.,

$$\pi^c_i(\mathcal{S}, c, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B)) \geq 0 \geq \pi^c_i(\mathcal{S}, c, \rho, (\overrightarrow{c}^S)). \tag{91}$$

The payment and waiting time in the spot market are only *changing* in the bid c at bids for which any users go into the spot market. It directly follows that a user's payoff in the spot market (if he were to choose it) is (weakly) higher for *any* user (and thus also for users with waiting cost c_1^S) under $\sigma = (\overrightarrow{c}^S)$ than under $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$. If ρ is not proper, it follows

$$\pi_1^{c_1^S}(\mathcal{S}, c_1^S, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B)) \geq \pi_1^{c_1^S}(\mathcal{S}, c_1^S, \rho, (\overrightarrow{c}^S)) > \pi_1^{c_1^S}(\mathcal{F}, c_1^S, \rho, (\overrightarrow{c}^S)) = \pi_1^{c_1^S}(\mathcal{F}, c_1^S, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B)), \tag{92}$$

i.e., users of class 1 with waiting cost c_1^S would deviate from $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$, contradicting that $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ is a BNE.

Now we show the statement for when ρ is not proper and when $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) < T$ (i.e., when the spot market is faster for the highest bids). If no c' exists for which the expected waiting time in both queues is equal (i.e., for which condition (a) from Definition 4.5 holds), the spot market is trivially faster for every user (and consequently also cheaper) and the statement follows by the same argument as for $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) \geq T$.

Now assume there exists a c' satisfying condition (a), but it does not satisfy condition (b), i.e.

$$c'(T + \frac{1}{\mu}) + p\frac{1}{\mu} \quad \geq \quad \int_0^{c'} w(\mathcal{S}, x, \rho, \sigma)dx \tag{93}$$

for c' such that $T + \frac{1}{\mu} = w(\mathcal{S}, c', \rho, \sigma)$. It follows that

$$p_F\frac{1}{\mu} \quad \geq \quad \int_0^{c'} w(\mathcal{S}, x, \rho, \sigma)dx - c'w(\mathcal{S}, c', \rho, \sigma). \tag{94}$$

This means that even with bid c', for which 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, in this case, there only exist BNEs where no user joins the fixed-price market. Since the payoff in the spot market for every user is monotone decreasing in the number of users that join, it is enough to show that when playing $\sigma = \overrightarrow{c}^S$, no user has an incentive to switch to the fixed-price market.

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

$$\begin{aligned}
\pi_i^c(\mathcal{S}, c', \rho, \overrightarrow{c}^S) &= v_i - \int_0^{c'} w(\mathcal{S}, x, \rho, \sigma)dx + c'w(\mathcal{S}, c', \rho, \sigma) - cw(\mathcal{S}, c', \rho, \sigma) && (95)\\
&= v_i - \int_0^{c'} w(\mathcal{S}, x, \rho, \sigma)dx + c'w(\mathcal{S}, c', \rho, \sigma) - c(T + \frac{1}{\mu}) && (96)\\
&\geq v_i - p_F\frac{1}{\mu} - c(T + \frac{1}{\mu}) && (97)\\
&= \pi_i^c(\mathcal{F}, c, \rho, \overrightarrow{c}^S) && (98)\\
&> \pi_i^c(\mathcal{S}, c, \rho, \overrightarrow{c}^S) && (99)
\end{aligned}$$

Misreporting in the spot market would therefore be beneficial over reporting truthfully, contradicting the pricing rule being Bayes-Nash incentive compatible. Consequently, no user prefers the fixed-price market and by Theorem 4.2 it holds that $\sigma^* = \overrightarrow{c}^S$.

A.7 Lemma A.3

The proof of Theorem 5.4 requires the introduction of an additional technical Lemma. The following Lemma establishes that the average payments in the spot market approach the payments in the fixed-price market for small enough l_S.

LEMMA A.3. *For any strategy* $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ *or* $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$, *denote by* $\bar{m}(\mathcal{S}, \rho, \sigma^*)$ *the average payment in the spot market, i.e., respectively*

$$\bar{m}(\mathcal{S}, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B)) := \frac{\sum_i \lambda_i \int_0^{c_i^P} m(\mathcal{S}, x, \rho, \sigma^*) f_i(x) dx}{\sum_i \lambda_i \int_0^{c_i^P} f_i(x) dx} \tag{100}$$

and

$$\bar{m}(\mathcal{S}, \rho, (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)) := \frac{\sum_i \lambda_i \left[\int_0^{c_i^L} m(\mathcal{S}, x, \rho, \sigma^*) f_i(x) dx + \int_{c_i^U}^{c_i^H} m(\mathcal{S}, x, \rho, \sigma^*) f_i(x) dx\right]}{\sum_i \lambda_i \left[\int_0^{c_i^L} f_i(x) dx + \int_{c_i^U}^{c_i^H} f_i(x) dx\right]}. \tag{101}$$

For every setting, $p_F < \mu v_1$ *and* $\varepsilon > 0$, *there exists a (possibly fractional) number of spot instances* $l_S \leq l$ *such that for* $\rho = (p_F, l_S)$ *it holds that* $\bar{m}(\mathcal{S}, \rho, \sigma^*)$ *is greater than the expected payment in the fixed-price market minus* ε, *i.e.*

$$\bar{m}(\mathcal{S}, \rho, \sigma^*) \geq \frac{p_F}{\mu} - \varepsilon. \tag{102}$$

PROOF. For any fixed p_F there exists some number of spot instances l' such that all provider strategies $\rho = (p_F, l_S)$ with $0 < l_S \leq l'$ result in an equilibrium that is either of the form $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ or of the form $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$. We first present the case where $T \leq \frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right)$ for all $0 < l_S \leq l'$ and therefore $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$. To keep the proof readable, we introduce new notation to draw all waiting costs in the spot market from a single distribution instead of first drawing a job's class and then its waiting cost. Note that this does not change the number of jobs or their bids nor their waiting costs in the market. For provider strategy ρ and equilibrium strategy profile $\sigma^* = (\overrightarrow{c}^L, \overrightarrow{c}^U, \overrightarrow{c}^H)$, we define the distribution

$$F_S(c) := \frac{\sum_i \lambda_i \left[\int_0^{\min\{c, c_i^L\}} f_i(x) dx + \int_{\min\{c, c_i^U\}}^{c} f_i(x) dx\right]}{\sum_i \lambda_i \left[\int_0^{c_i^L} f_i(x) dx + \int_{c_i^U}^{c_i^H} f_i(x) dx\right]} \tag{103}$$

and the arrival rate

$$\lambda_S := \sum_i \lambda_i \left[\int_0^{c_i^L} f_i(x) dx + \int_{c_i^U}^{c_i^H} f_i(x) dx\right] \tag{104}$$

with similarly constructed PDF $f_S(c)$. Now consider an artificial spot market with arrival rate λ_S, where every arriving job's waiting cost is drawn from F_S and everyone joins. From the provider's point of view, this market is the same as the normal spot market that would result from her playing ρ, including users on average having the same expected payments. To analyze the provider's profit from the spot market when playing ρ, we can thus instead analyze this artificial market.

The per-user-average profit $\bar{m}(\mathcal{S}, \rho, \sigma^*)$ of the artificial spot market is given by taking the expectation of the payment $m(\mathcal{S}, c, \rho, \sigma^*)$, where the expectation is taken over c drawn from the

PDF $f_S(c)$:

$$\begin{aligned}
\bar{m}(\mathcal{S}, \rho, \sigma^*) &= \frac{\lambda_S \left[\int_0^{c_1^H} m(\mathcal{S}, x, \rho, \sigma^*) f_S(x) dx\right]}{\lambda_S} && (105)\\
&= \int_0^{c_1^H} m(\mathcal{S}, x, \rho, \sigma^*) f_S(x) dx && (106)\\
&= \int_{-\infty}^{\infty} m(\mathcal{S}, x, \rho, \sigma^*) f_S(x) dx && (107)\\
&= E_{c \sim f_S}[m(\mathcal{S}, c, \rho, \sigma^*)] && (108)
\end{aligned}$$

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

$$\begin{aligned}
\bar{m}(\mathcal{S}, \rho, \sigma^*) &\geq (1 - F_S(c_\xi^{l_S})) m(\mathcal{S}, c_\xi^{l_S}, \rho, \sigma^*) && (109)\\
&= (1 - \xi) m(\mathcal{S}, c_\xi^{l_S}, \rho, \sigma^*). && (110)
\end{aligned}$$

Further, because the total waiting time is monotone increasing, by the integral upper bound, the following holds:

$$\begin{aligned}
\frac{p_F}{\mu} &= \int_0^{c^L} w(\mathcal{S}, x, \rho, \sigma^*) dx + c^L(\frac{1}{\mu} - T) && (111)\\
&\leq c^L \, max_{c \in [0, c^L]} \, w(\mathcal{S}, c, \rho, \sigma^*) + c^L(\frac{1}{\mu} - T) && (112)\\
&= c^L w(\mathcal{S}, 0, \rho, \sigma^*) + c^L(\frac{1}{\mu} - T) && (113)
\end{aligned}$$

Now observe that the cutoff point c^L goes to zero as the spot market becomes sufficiently small (i.e., $c^L \xrightarrow[l_S \to 0]{} 0$). Combined with Equation (113), it follows that the waiting time goes to infinity for users with bid 0, i.e.:

$$w(\mathcal{S}, 0, \rho, \sigma^*) \xrightarrow[l_S \to 0]{} \infty. \qquad (114)$$

As a job with bid 0 is served exactly when there is an idle instance in the spot queue (i.e., there are fewer than l_S jobs of higher priority in the spot queue), the instance utilization of the spot queue has to go to full as the size of the spot market becomes sufficiently small, i.e.

$$\frac{\lambda_S}{l_S \mu} \xrightarrow[l_S \to 0]{} 1. \qquad (115)$$

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

$$\frac{(1 - F_S(c_\xi^{l_S}))\lambda_S}{l_S \mu} \xrightarrow[l_S \to 0]{} (1 - \xi)1 \qquad (116)$$

i.e., as the size of the spot market goes towards zero, the (average) utilization of the spot instances by jobs with priority over $c_\xi^{l_S}$ will always at most be $(1 - \xi)$. For a given ξ (but independent of l_S), this limits the total waiting time at $c_\xi^{l_S}$ to some possibly very high but finite value $\bar{w}_\xi$. For any $c_\xi^{l_S}$ it further either holds $c_{\xi}^{l_S} > c_1^L$ (and $m(\mathcal{S}, c_1^L, \rho, \sigma^*) < m(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*)$ trivially) or the following holds:

$$m(\mathcal{S}, c_1^L, \rho, \sigma^*) \tag{117}$$

$$= \int_0^{c_{\xi}^{l_S}} w(\mathcal{S}, x, \rho, \sigma^*)dx + \int_{c_{\xi}^{l_S}}^{c_1^L} w(\mathcal{S}, x, \rho, \sigma^*)dx - c_1^L w(\mathcal{S}, c_1^L, \rho, \sigma^*) \tag{118}$$

$$\leq \int_0^{c_{\xi}^{l_S}} w(\mathcal{S}, x, \rho, \sigma^*)dx + (c_1^L - c_{\xi}^{l_S})w(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) - c_1^L w(\mathcal{S}, c_1^L, \rho, \sigma^*) \tag{119}$$

$$= \int_0^{c_{\xi}^{l_S}} w(\mathcal{S}, x, \rho, \sigma^*)dx - c_{\xi}^{l_S} w(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) + c_1^L(w(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) - w(\mathcal{S}, c_1^L, \rho, \sigma^*)) \tag{120}$$

$$\leq \int_0^{c_{\xi}^{l_S}} w(\mathcal{S}, x, \rho, \sigma^*)dx - c_{\xi}^{l_S} w(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) + c_1^L(\bar{w}_{\xi} - w(\mathcal{S}, c_1^L, \rho, \sigma^*)) \tag{121}$$

$$= m(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) + c_1^L(\bar{w}_{\xi} - w(\mathcal{S}, c_1^L, \rho, \sigma^*)) \tag{122}$$

As $c_1^L \xrightarrow[l_S \to 0]{} 0$, it follows that, for all $0 < \xi < 1$ and all $\delta > 0$ there exists an l_S with $m(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) \geq m(\mathcal{S}, c_1^L, \rho, \sigma^*) - \delta$ and therefore

$$\bar{m}(\mathcal{S}, \rho, \sigma^*) \geq (1-\xi)m(\mathcal{S}, c_{\xi}^{l_S}, \rho, \sigma^*) \tag{123}$$

$$\geq (1-\xi)(m(\mathcal{S}, c_1^L, \rho, \sigma^*) - \delta) \tag{124}$$

Choosing ξ and δ such that $\frac{1}{2}\varepsilon \geq \xi m(\mathcal{S}, c_1^L, \rho, \sigma^*) + (1-\xi)\delta$ and noting that by Lemma A.2 it holds $m(\mathcal{S}, c_1^L, \rho, \sigma^*) = \frac{p_F}{\mu}$ then yields

$$\bar{m}(\mathcal{S}, \rho, \sigma^*) \geq m(\mathcal{S}, c_1^L, \rho, \sigma^*) - \frac{1}{2}\varepsilon = \frac{p_F}{\mu} - \frac{1}{2}\varepsilon \tag{125}$$

and the statement of the lemma follows.

When $T > \frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right)$ and $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$, we analogously (only replacing the relevant notation) obtain

$$\bar{m}(\mathcal{S}, \rho, \sigma^*) \geq m(\mathcal{S}, c_1^P, \rho, \sigma^*) - \frac{1}{2}\varepsilon. \tag{126}$$

Because users with waiting cost $\overrightarrow{c}_1^P$ are indifferent between both markets, it has to hold that $\frac{p_F}{\mu} + c_1^P(\frac{1}{\mu} + T) = c_1^P \frac{1}{\mu}\frac{1}{1-\tau\psi_E(l_S)} + m(\mathcal{S}, c_1^P, \rho, \sigma^*)$. Solving this for $m(\mathcal{S}, c_1^P, \rho, \sigma^*)$ and substituting it into Equation (126) yields

$$\bar{m}(\mathcal{S}, \rho, \sigma^*) \geq \frac{p_F}{\mu} - c_1^P\left(\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) - T\right) - \frac{1}{2}\varepsilon. \tag{127}$$

Lastly note that $c_1^P\left(\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) - T\right) \xrightarrow[l_S \to 0]{} 0$ because $\frac{1}{\mu}\left(\frac{1}{1-\tau\psi_E(l_S)} - 1\right) - T$ increases in l_S and it holds that $c_1^P \xrightarrow[l_S \to 0]{} 0$. For any $\varepsilon > 0$, l_S small enough therefore yields the statement of the lemma. □

B BASIS OF THE NUMERICAL EXAMPLE

In this section, we give a precise description of how we calculate waiting times and preemption probabilities for the numerical examples. In the following we let $\varphi(l, \frac{\lambda}{\mu})$ denote the probability that fewer than l jobs are currently in a queue with l instances, arrival rate λ and with expected job service time $\frac{1}{\mu}$. For memoryless queues, $\varphi(l, \frac{\lambda}{\mu})$ is given by the well-known Erlang C formula:[20]

$$\varphi(l, \frac{\lambda}{\mu}) = 1 - \left(1 + (1 - \frac{\lambda(c)}{l\mu}) \frac{l!}{\frac{\lambda(c)}{\mu}^{l}} \sum_{k=0}^{l-1} \frac{\frac{\lambda(c)}{\mu}^{k}}{k!}\right)^{-1} \tag{128}$$

Given this, the required calculations of the numerical example for the fixed-price queue are straightforward, while we need to make additional simplifying assumptions for the spot queue.

B.1 Fixed-price Queue

For a fixed-price queue, the total waiting time $w(\mathcal{F}, c, \rho, \sigma)$ and payment $m(\mathcal{F}, c, \rho, \sigma)$ are directly determined by the parameters of the setting. The only thing left to calculate is the minimal number of instances $l_F(\rho, \sigma)$ required to serve all users in the fixed-price market while observing the upper bound on the queuing time T. This can easily be done using the Erlang C formula, as it is well known that the expected queuing time $q(l, \frac{\lambda}{\mu})$ of a user joining a FIFO queue with l instances, arrival rate λ and service time $\frac{1}{\mu}$ is given by

$$q(l, \frac{\lambda}{\mu}) = \frac{1 - \varphi(l, \frac{\lambda}{\mu})}{l\mu - \lambda}. \tag{129}$$

See (Cooper 1981) for a proof. Plugging in the arrival rate into the fixed-price market for any pair of strategies (ρ, σ) and solving $q(l_F(\rho, \sigma)) = T$ then yields $l_F(\rho, \sigma)$.

B.2 Spot Queue

For the spot queue, the total waiting time $w(\mathcal{S}, c, \rho, \sigma)$ and payment $m(\mathcal{S}, c, \rho, \sigma)$ are not directly determined by the setting because they depend on the dynamics of the queue. Unfortunately, we cannot directly use the Erlang C to derive those terms because this would require that the running times and queuing times for all users are equal (Buzen and Bondi 1983). Since with priorities, and especially when costly preemptions are present, they are *not* equal, we make the following two simplifying assumptions (for the numerical examples only). This then allows us to calculate waiting times and preemption probabilities.

ASSUMPTION. *For calculating $w(\mathcal{S}, c, \rho, \sigma)$, we assume that jobs, while running, see the steady state distribution over states of the spot queue (when looking at the queue not including itself).*

Note that Assumption B.2 would be exactly satisfied if a given job ran for an infinite amount of time. Since jobs start in a random state but end in a state in which the queue has free capacity for a job of the same priority, jobs in practice see more "busy" states than in steady state and consequently take slightly longer to run. Importantly, we only make this assumption when calculating any single jobs's runtime, but we still calculate the steady state of the queue exactly to avoid the accumulation of approximation errors.

ASSUMPTION. *Any additional running time above and beyond a job's service time $\frac{1}{\mu}$ is run on "abstract" additional instances and does not influence the spot queue's steady state. However, while run on these abstract instances, a job still causes load-dependent costs for the provider and is still (internally and externally) preempted as if it was in the queue, as denoted by $\psi_I(c, \rho, \sigma)$ and $\psi_E(l_S)$.*

[20] See for example (Cooper 1981); a proof of the Erlang C formula can be found in (Takagi 2008).

Effectively, Assumption B.2 dynamically gives the spot market more instances than it actually has, to accommodate the additional running time needed due to preemptions.

Taken together, these two assumptions give us the following very useful Lemma.

LEMMA B.1. *During its time in the spot queue, a job with bid c sees the steady state distribution of a FIFO queue with arrival rate $\lambda(c)$ and service time $\frac{1}{\mu}$. Here, $\lambda(c)$ denotes the arrival rate of jobs with a higher priority; i.e., during any time unit, on average, $\lambda(c)$ jobs with a higher priority than c arrive into the queue.*

PROOF. Note that jobs with a lower bid do not influence the total waiting time of a user and can thus be ignored. As the probability that any other user in the system also has 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. Combining this with Assumption B.2, we can assume that the job, while running, sees the steady state probabilities of the queue consisting of only those users with higher priorities than itself. Furthermore, by Assumption B.2, these steady state probabilities are the same as the steady state probabilities with zero preemption costs, which in turn are the same as the steady state probabilities of a FIFO queue consisting of all users with higher priority (see Buzen and Bondi (1983)). Taken together, the statement follows. □

Given Lemma B.1, we can now derive expressions for the waiting time and the expected number of internal preemptions per time unit.

PROPOSITION B.2. *Given Assumptions B.2 and B.2, provider strategy $\rho = (p_F, l_S)$, and user strategy profile σ, the total waiting time of a user with bid c is given by*

$$w(\mathcal{S}, c, \rho, \sigma) \quad = \quad \frac{r(\mathcal{S}, c, \rho, \sigma)}{\varphi(l_S, \frac{\lambda(c)}{\mu})}, \tag{130}$$

where $\lambda(c)$ denotes the arrival rate of jobs with a bid higher than c into the spot queue (given σ).

PROOF. Recall that the waiting time is defined as

$$w(\mathcal{S}, c, \rho, \sigma) \quad = \quad q(\mathcal{S}, c, \rho, \sigma) + r(\mathcal{S}, c, \rho, \sigma). \tag{131}$$

Observe that when a job with bid c is in the spot queue, it is run whenever there are fewer than l_S jobs with a higher bid in the system. By Lemma B.1, during its runtime, a jobs sees the steady state distribution of a FIFO queue with arrival rate $\lambda(c)$ and service time $\frac{1}{\mu}$. Thus, to be running for one full time unit, the job with bid c will, on average, have a waiting time of $\frac{1}{\varphi(l_S, \frac{\lambda(c)}{\mu})}$ time units. The statement of the Proposition now follows by noting that the running time of a job is given by $r(\mathcal{S}, c, \rho, \sigma)$. □

PROPOSITION B.3. *Given Assumptions B.2 and B.2, provider strategy $\rho = (p_F, l_S)$, and user strategy profile σ, the expected number of internal preemptions per time unit of a user with bid c is given by*

$$\psi_I(c, \rho, \sigma) \quad = \quad \frac{(\mu l_S - \lambda(c))(1 - \varphi(l_S, \frac{\lambda(c)}{\mu}))}{\varphi(l_S, \frac{\lambda(c)}{\mu})}, \tag{132}$$

where $\lambda(c)$ denotes the arrival rate of jobs with a bid higher than c into the spot queue (given σ).

PROOF. While a job with bid c is running, it will be internally preempted whenever the system contains exactly $l_S - 1$ jobs of higher priority and another job of higher priority arrives. By Lemma B.1, during its runtime, the job sees the steady state distribution of a FIFO queue with arrival rate $\lambda(c)$ and service time $\frac{1}{\mu}$. Thus, given a newly-arriving job (with priority higher than c), the probability that this job preempts the job with bid c is equal to the probability that the FIFO queue contains exactly $l_S - 1$ jobs, which is given by

$$\frac{(1-\frac{\lambda(c)}{\mu l_S})}{\frac{\lambda(c)}{\mu l_S}}(1-\varphi(l_S,\frac{\lambda(c)}{\mu})) \tag{133}$$

(see (Cooper 1981)). Since we are interested in the preemption rate taken over the running time of the job (as opposed to the total time the job is in the system), we normalize this term by the probability that less than l_S jobs of higher priority are in the system. Because $\lambda(c)$ jobs with higher priority arrive per time unit, we also multiply with $\lambda(c)$, which yields

$$\psi_I(c,\rho,\sigma) = \lambda(c)\frac{\frac{(1-\frac{\lambda(c)}{\mu l_S})}{\frac{\lambda(c)}{\mu l_S}}(1-\varphi(l_S,\frac{\lambda(c)}{\mu}))}{\varphi(l_S,\frac{\lambda(c)}{\mu})} \tag{134}$$

$$= \frac{(\mu l_S - \lambda(c))(1-\varphi(l_S,\frac{\lambda(c)}{\mu}))}{\varphi(l_S,\frac{\lambda(c)}{\mu})}. \tag{135}$$

□

Taking the expression for the waiting time from Proposition B.2, plugging in the expression for the running time from Proposition 3.1, and lastly plugging in the expression for the internal preemptions from Proposition B.3, we can now write the expected total waiting time $w(\mathcal{S},c,\rho,\sigma)$ of a user joining the spot queue as

$$w(\mathcal{S},c,\rho,\sigma) = \frac{1}{\mu\varphi(l_S,\frac{\lambda(c)}{\mu})}\frac{1}{1-\tau(\psi_I(c,\rho,\sigma)+\psi_E(l_S))} \tag{136}$$

$$= \frac{1}{\mu\varphi(l_S,\frac{\lambda(c)}{\mu})}\frac{1}{1-\tau(\frac{(\mu l_S-\lambda(c))(1-\varphi(l_S,\frac{\lambda(c)}{\mu}))}{\varphi(l_S,\frac{\lambda(c)}{\mu})}+\psi_E(l_S))}. \tag{137}$$

Using the iterative approach described in the proof of Proposition 4.4, we can calculate the cutoff vectors of the user equilibrium strategies by solving a number of non-linear root searches. This allows us to calculate payments and profits and search for the optimal provider strategy $\rho = (p_F, l_S)$.

C USER WELFARE IN EXAMPLE 1

To help us better understand how the hybrid strategy with $p_F = p_F^{*0}$ affects the users, Figure 5 shows the user welfare for this strategy and compares it against the fixed-price strategy (we omit the other two strategies because plotting all four strategies makes the figure very hard to read). As we can see, for the fixed-price strategy (dashed blue line), the user welfare monotonically decreases in the instance costs κ. While the instance costs do not

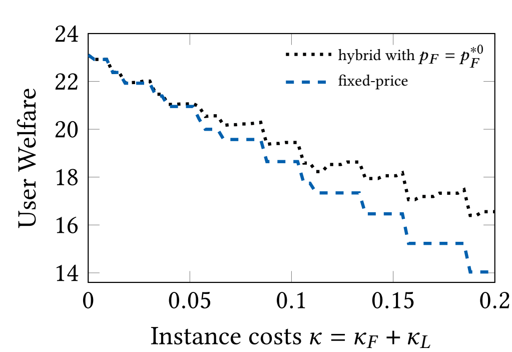

Fig. 5 User welfare under different strategies while varying the instance costs κ

directly affect the users, they influence the optimal provider strategies. This leads to the observed discontinuities in the welfare, since instances can only be bought in discrete units and the provider consequently changes her strategy in discrete steps.

It is clearly visible that the hybrid strategy with $p_F = p_F^{*0}$ (the dotted black line) shares many large discontinuity points with the fixed-price strategy (the dashed blue line). This happens because both strategies share the same price p_F^{*0}, which at these points increases, causing some users to balk and making everyone that remains in the fixed-price market worse off. However, the *size* of the discontinuities may differ, because under the hybrid strategy, some users might additionally move from the fixed-price to the spot market. Perhaps surprisingly, *between* these shared discontinuity points, the welfare corresponding to the hybrid strategy often increases. This happens because the provider is restricted to $p_F = p_F^{*0}$ and cannot optimally increase her price, so has to change l_S instead. To understand this in more detail, consider a situation where a cost increase does not lead to a change of p_F, but where the provider still wants users to move out of the relatively less profitable fixed-price market and into the spot market. In this situation, she then has to increase the attractiveness of the spot market, which increases the user welfare. Finally, as expected, the hybrid strategy with $p_F = p_F^{*0}$ always leads to a higher user welfare than the fixed-price strategy, which can be seen by the dotted black line always lying above the dashed blue line.

D USING IDLE FIXED-PRICE INSTANCES FOR THE SPOT MARKET

In this section, we consider an alternative model where the provider takes the spot instances from the pool of currently idle *fixed-price instances* instead of using other idle capacity (long-term reserved instances, maintenance capacity, etc.). In Section D.1., we first discuss the importance of using instances that are *reliably idle* for the spot market and why the fixed-price market is typically *not* the best source for such instances. Nevertheless, as some providers may want to use idle instances from the fixed-price market, we then show how the new cross-channel interactions change our results compared to our main model. In Section D.2, we first present the required changes to the model. In Section D.3, we then show how the equilibria of the user game change. Finally, in Section D.4, we then derive a similar condition for how a provider can simultaneously achieve a profit increase and a Pareto improvement for the users (as we did for our main model).

D.1 The Availability of Reliably Idle Instances from the Fixed-Price Market

As preemptions are costly for the users, the "usefulness" of an idle instance critically relies on how *reliably idle* it is (i.e., for how long the instance will remain idle). Therefore, the provider should use the most reliably idle instances for the spot market. While fixed-price markets of large cloud providers do contain a reasonable number of idle instances *on average*, only few of those instances are reliably idle and providers should not simply put any idle fixed-price instance on the spot market. This is due to two effects: larger markets require relatively smaller buffers and these buffers will be used more frequently the larger the market (but for increasingly shorter durations). To see this, we now provide a simple but striking numerical example.

Example D.1. Consider two M/M/l queues, one with arrival rate 100 and one with arrival rate 1000. Assume for both an expected service time of $\frac{1}{\mu} = 1$ and an SLA of $T = 0.0001$. We can use the Erlang C formula to derive that we need a buffer of 22 instances to satisfy the SLA for the first queue with arrival rate of 100 (i.e., the provider needs $l_F = 122$ fixed-price instances). For the second queue with arrival rate 1000 the required buffer only grows from 22 to 55 (i.e., the provider now needs $l_F = 1055$ fixed-price instances).

Now assume that the provider uses up to 1 idle instance for a spot queue, i.e., $l_s = 1$. Assume that this idle instance comes from the first queue (with arrival rate 100) and whenever the provider has at least 1 idle fixed-price instance and no spot job is running, he starts a new spot job. Then a job in the spot market would in expectation get externally preempted at a rate of $\psi_E = 0.47$. For the second queue (with arrival rate 1000), the rate of external preemptions for spot jobs rises to $\psi_E = 3.07$. For both queues, the high preemption rate occurs because a large queue can frequently reach zero idle capacity while still satisfying the SLA, as it is highly likely that another instance becomes free shortly thereafter (and because newly-arriving users are willing to wait a little bit). However, any time this happens, the spot instance is immediately preempted. This effect becomes more pronounced the larger the fixed-price queue, which explains the large rise of the preemption rate for the second queue with arrival rate 1000.

In practice, the provider wants to keep the preemption rate reasonably low. To this end, she can decide to only start spot jobs whenever the expected time until any running spot job will be externally preempted is above some threshold t_E, resulting in $\psi_E < \frac{1}{t_E}$. While doing this reduces the number of external preemptions, this of course also further reduces the supply of instances for the spot market. For example, if the provider wanted to ensure that jobs (in expectation) run for at least $t_E = 20$ before they get preempted (leading to $\psi_E < \frac{1}{t_E} = 0.05$), then she would have to only start spot jobs when the fixed-price queue contains less than 105 jobs (for the queue with arrival rate 100 and $l_F = 122$ fixed-price instances) and less than 957 jobs (for the queue with arrival rate 1000 and $l_F = 1055$ fixed-price instances).[21]

Note that the size of the effects observed in the example are particularly large because the example considers memoryless service processes. While real-world service processes are usually heavy-tailed (which leads to larger and more reliably idle buffers), the observation that larger fixed-price markets lead to less reliably idle instances remains true in practice. Thus, a provider who wants to limit the number of external preemptions can only offer relatively few reliably idle fixed-price instances on the spot market. Additionally, the provider has to consider the cross-channel effects that occur when users move from the fixed-price market to the spot market, which decreases the number of fixed-price instances but not the number of users. While a provider can nevertheless choose to offer idle fixed-price instances on the spot market, most providers typically have access to many alternative instances that are more reliably idle than most idle instances from the fixed-price market. This includes instances from other business areas (e.g., long-term reserved instances), maintenance instances (which make up 5-10% of the capacity of a cloud computing center), etc. At least some of these alternatives are available for all current major cloud providers.

D.2 Required Model Changes

Even though most idle instances from the fixed-price market are typically not reliably idle, some cloud providers may still want to use them for a secondary spot market. Therefore, we now show how to adapt our model to using idle fixed-price instances. The most immediate change is that an *external preemption* now happens whenever a spot user is preempted in favor of a fixed-price user. Thus, while previously the number of external preemptions $\psi_E(l_S)$ was a function given by the setting that only depended on the provider's strategy ρ, the number of external preemptions $\psi_E(c, \rho, \sigma)$ now arises from the queueing system. Specifically, it now also depends on the strategies of all other users σ and, because lower bids get preempted first, also on a user's bid c. While in our main model, l_S denotes the (average) number of offered spot instances, it now denotes the maximum number of idle fixed-price instances the provider offers on the spot market, i.e., l_S is now an *upper bound* on the number of offered spot instances.

[21] The preemption rates and the expected time until the next preemption can be calculated by solving the difference equations of the corresponding Markov chains.

To control the number of external preemptions and only offer sufficiently reliably idle instances on the spot market, we introduce an additional strategy variable for the provider t_E, which denotes that the provider only starts a new spot job whenever, after starting this job, the expected time until the next external preemption for any running spot job is above the threshold t_E. Thus, given provider strategy $\rho = (p_F, l_S, t_E)$, for a job to be started in the spot market, four conditions have to be satisfied: (1) no job with a higher priority is waiting; (2) if the job started, there would be at most l_S spot jobs running, (3) there has to be an idle fixed-price instance or there is currently a spot job with a lower priority running, and (4) if the job started now, the expected time until the next external preemption for any running spot job would be at least t_E. Note that this implies that a spot job with low priority is *not* immediately preempted when a spot job with higher priority is waiting if the expected time until the next external preemption is currently too low. Due to the new cross-channel interactions that arise because the spot instances are now taken from the fixed-price market, both the running time $r(\mathcal{S}, c, \rho, \sigma)$ and the queuing time $q(\mathcal{S}, c, \rho, \sigma)$ in the spot market are now highly dependent on the number of users that join the fixed-price market. Additionally, if the threshold t_E is set too high (for a given user strategy profile σ), then the provider may never start a spot job (effectively not offering a spot market). A setting in our alternative model is now fully defined by $(n, v, \lambda, \mu, F, \tau, \kappa_F, \kappa_L, T)$ because the alternative model does not contain a maximum number of available spot instances l nor an exogenous function for the number of external preemptions.

Because larger fixed-price markets can have less reliably idle capacity than smaller markets (see Example D.1), we may observe the counterintuitive effect that the waiting time of the users with the highest priority in the spot market can decrease in the number of people that join the spot market. However, the overall costs of any user joining the spot market, i.e, $\int_0^c w(\mathcal{S}, x, \rho, \sigma)dx$ typically increase for any fixed c if more users move to the spot market. To see this, note that when users move from the fixed-price market to the spot market, the total number of instances decreases, but the number of users does not. Yet we cannot say with certainty that $\int_0^c w(\mathcal{S}, x, \rho, \sigma)dx$ *always* increases because the service discipline of the whole market is not work-conserving (i.e., there can be an idle instance even though jobs are waiting when the time until the next external preemption is too low) and these dynamics change whenever users move from the fixed-price to the spot market. While this effect is typically negligible compared to the reduction in the number of instances in the system, we cannot fully exclude the possibility that there could be some parameterizations for which there is a σ where $\int_0^c w(\mathcal{S}, x, \rho, \sigma)dx$ is *locally* decreasing in the number of users that choose the spot market. To avoid having to handle those cases (which do not change the form of the potential equilibria, but could in rare cases potentially lead to the existence of multiple equilibria) we therefore make the following assumption for the rest of the paper:

Assumption. *The overall cost of a user with any fixed bid c that joins the spot market, i.e.* $\int_0^c w(\mathcal{S}, x, \rho, \sigma)dx$*, increases if additional users (compared to* σ*) move to the spot market.*

D.3 Equilibria

Whenever some instances are actually offered on the spot market, we obtain an equilibrium structure similar to the one derived in Subsection 4.3.1:

Proposition D.2. *For any provider strategy* $\rho = (p_F, l_S, t_E)$*, in any BNE of the user game where any user joins the spot market, any equilibrium strategy profile is of the form* $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$*. Here,* $\sigma = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ *denotes that a user of class i with waiting cost c joins the spot market when* $c < c_i^P \leq c_i^B$ *and the fixed-price market when* $c_i^P < c < c_i^B$*; when* $c > c_i^B$*, he balks and does not join any market. The* *cutoff point* c_1^P *and the cutoff vector* $\overrightarrow{c}^B$ *are the unique solution to the following system of equations:*

$$0 = c_1^P(T + \frac{1}{\mu}) + \frac{p_F}{\mu} - \int_0^{c_1^P} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B))dx \tag{138}$$

$$0 = v_i - \min\left\{c_i^B(\frac{1}{\mu} + T) + \frac{p_F}{\mu}, \int_0^{c_i^B} w(\mathcal{S}, x, \rho, (\overrightarrow{c}^P, \overrightarrow{c}^B))dx\right\} \quad \forall i \in \{1, \ldots, n\} \tag{139}$$

The rest of the cutoff vector $\overrightarrow{c}^P$ *is given as* $c_i^P = min(c_1^P, c_i^B)$.

Proof. Even users with the highest bid in the spot market have to queue longer than users in the fixed-price market, because (by definition) a user only gets served in the spot market when the fixed-price market has idle capacity. Thus, all users in the spot market are willing to pay strictly less than what they would have to pay in the fixed-price market. The remainder of the proof is equivalent to the proof of Proposition 4.3. □

While this gives us the structure of the equilibrium when some spot instances are offered and utilized, the following proposition tells us when that is the case.

Proposition D.3. *For any provider strategy* $\rho = (p_F, l_S, t_E)$*, the equilibrium strategy profile of the users is*

(1) $\sigma^* = \overrightarrow{c}^F$ *(i.e., no user joins the spot market, as described in Proposition 4.1) if and only if* $l_S = 0$ *or* t_E *is "too high," i.e., the fixed-price queue arising from* $\sigma = \overrightarrow{c}^F$ *has no state for which the expected time until the next external preemption of a hypothetically starting spot job would be more than* t_E.
(2) $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$ *otherwise.*

Proof. Recall from Proposition 4.1 that $\sigma = \overrightarrow{c}^F$ is the equilibrium user strategy profile when no spot market is offered. We denote by $(\overrightarrow{x}, \overrightarrow{c}^F)$ a different strategy profile where any user of class i with waiting cost c joins the spot market if $c < x_i$. We now look at different provider strategies and classify the corresponding user equilibrium strategy profiles. If $l_S = 0$, then $\sigma^* = \overrightarrow{c}^F$ trivially. Now assume that the fixed-price queue arising from $\sigma = \overrightarrow{c}^F$ has no state for which the expected time until the next external preemption of a hypothetically starting spot job would be more than t_E. Then, as long as almost all users (i.e., all besides at most a null set) play $\sigma = \overrightarrow{c}^F$, the provider would never start a spot job, even if a single user deviated to the spot market and $l_S > 0$. Consequently, it holds that $\int_0^x w(\mathcal{S}, c, \rho, \sigma)dc = \infty$. By Assumption D.2, it immediately follows that $\int_0^x w(\mathcal{S}, c, \rho, (\overrightarrow{x}, \overrightarrow{c}^F))dc = \infty$ for any $\overrightarrow{x}$ and thus, in equilibrium, no user joins the spot market. On the other hand, if $l_s > 0$ and if the fixed-price queue arising from user strategy profile $\sigma = \overrightarrow{c}^F$ has a state for which the expected time until the next external preemption of a hypothetically starting spot job would be more than t_E, then users with waiting cost very close to 0 prefer the spot market and by Proposition D.2 it holds that $\sigma^* = (\overrightarrow{c}^P, \overrightarrow{c}^B)$. □

D.4 Well-behaved Settings: Increasing Provider Profit and User Welfare

We now show how the profit and welfare result from our main model translates to the alternative model. First note that the bound from Lemma 5.1 on the number of saved fixed-price instances per fixed-price user who moves to the spot market still holds, as the mechanics of the fixed-price market did not change. Next, we translate Lemma 5.2 to the alternative model.

Lemma D.4. *The average running time in the spot market (i.e., the left-hand side of the following inequality) is bounded above as follows:*

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} < (\frac{1}{\mu}+\tau)\frac{1}{1-\tau\frac{1}{t_E}} \tag{140}$$

Proof. Recall that $\psi_I(y,\rho,\sigma)r(\mathcal{S},y,\rho,\sigma)$ denotes the number of internal preemptions a job suffers in expectation. By the same arguments as in Lemma 5.2, it holds that

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} \psi_I(x,\rho,\sigma) r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} < 1 \tag{141}$$

and

$$r_I(\mathcal{S},c,\rho,\sigma) = \psi_I(c,\rho,\sigma)r(\mathcal{S},c,\rho,\sigma)\tau\frac{1}{1-\tau\psi_E(c,\rho,\sigma)} \tag{142}$$

$$\leq \psi_I(c,\rho,\sigma)r(\mathcal{S},c,\rho,\sigma)\tau\frac{1}{1-\tau\frac{1}{t_E}}, \tag{143}$$

where (143) follows from (142) because, whenever a job starts to run, the expected time until the next preemption is bounded by t_E. Combining these two inequalities we obtain

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r_I(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} f_i(x)dx} < \frac{\tau}{1-\tau\frac{1}{t_E}}. \tag{144}$$

Similar as in the proof of Lemma 5.2, we finally obtain

$$\frac{\sum_i \lambda_i \int_{x:\sigma_{i,1}(x)=\mathcal{S}} r(\mathcal{S},x,\rho,\sigma) f_i(x)dx}{\sum_i \lambda_i \int_{x:\sigma_{i,1}(xx)=\mathcal{S}} f_i(x)dx} < (\frac{1}{\mu}+\tau)\frac{1}{1-\tau\frac{1}{t_E}}. \tag{145}$$

□

Given these bounds, we can now state a well-behavedness condition analogous to Definition 5.3 for our main model, where the new parameter t^w corresponds to a lower bound on the strategy variable t_E (capturing the reliablity of the fixed-price instances):

Definition D.5. We say that a setting $(n,v,\lambda,\mu,F,\tau,\kappa_F,\kappa_L,T)$ is t^w-*well-behaved* if t^w is the infimum over the t_E for which the following holds:

$$\frac{1+\tau\mu}{1-\tau\frac{1}{t_E}} - 1 < \frac{\kappa_F}{\kappa_L} \tag{146}$$

With this definition in hand, we can now show a profit and welfare result analogous to Theorem 5.4 for our main model.

Theorem D.6. *Given a t^w-well-behaved setting $(n,v,\lambda,\mu,F,\tau,\kappa_F,\kappa_L,T)$ and any fixed-price strategy $\rho_0 = (p_F^0, 0, \infty)$ that results in a positive profit and for which the queue arising from the corresponding equilibrium user strategy profile σ_0^* has any state for which the expected time until the next external preemption of a hypothetically starting spot job would be more than t^w, then there exists a strategy $\rho = (p_F^0, l_S, t_E)$ with the same price p_F^0, with $0 < l_S$ and with $t_E \geq t^w$ that yields a higher profit for the provider, i.e.,*

$$\Pi((p_F^0, l_S, t_E), \sigma^*) > \Pi((p_F^0, 0, \infty), \sigma_0^*), \tag{147}$$

and the same strategy also yields a Pareto improvement for the users, i.e.,

$$\forall i \in \{1, \ldots, n\}\, \forall c \in [0, \mu v_i]: \quad \pi_i^c(\alpha, \beta, \rho, \sigma^*) \geq \pi_i^c(\alpha, \beta, \rho_0, \sigma_0^*), \textit{ and} \tag{148}$$

$$\exists i \in \{1, \ldots, n\}\, \exists c \in [0, \mu v_i]: \quad \pi_i^c(\alpha, \beta, \rho, \sigma^*) > \pi_i^c(\alpha, \beta, \rho_0, \sigma_0^*). \tag{149}$$

PROOF. By Proposition D.3, any such strategy $\rho = (p_F^0, l_S, t_E)$ leads to some users joining the spot market in equilibrium. The proof of the theorem is then equivalent to the proof of Theorem 5.4 after replacing the general well-behavedness bound on the running time with $b(l_S) = \frac{1+\tau\mu}{1-\tau\frac{1}{t_E}}$. □

Informally, Theorem D.6 says that if the provider's current fixed-price market has some instances that are sufficiently reliably idle, then she can obtain a profit increase and achieve a Pareto improvement for the users by offering a spot market alongside her existing fixed-price market (as in our main model). Note that, in contrast to our main model, executing the provider's strategy in practice is now more difficult, because it will typically be intractable to exactly calculate, for every possible state, whether t_E would be satisfied when starting a new job. However, in this case, the provider could still approximate t_E (e.g., by using historical or simulated data).

While our analysis shows that offering idle fixed-price instances on the spot market can (in principle) be advantageous for the provider, recall from Section D.1 that a provider typically only has relatively few fixed-price instances that are sufficiently reliably idle. In contrast, instances from other areas of the cloud computing center (e.g., long-term reserved instances, maintenance instances, or capacity buffers intended for hardware failure) usually offer a better stock of idle capacity. We therefore recommend using idle instances from the fixed-price market only to bolster the supply of instances for the spot market when the utilization of the fixed-price market is particularly low and to instead primarily use other sources of idle capacity for the spot market.