On-demand or Spot? Selling the cloud to risk-averse customers

Darrell Hoy[*] Nicole Immorlica[†] Brendan Lucier[‡]

July 29, 2016

Abstract

In Amazon EC2, cloud resources are sold through a combination of an on-demand market, in which customers buy resources at a fixed price, and a spot market, in which customers bid for an uncertain supply of excess resources. Standard market environments suggest that an optimal design uses just one type of market. We show the prevalence of a dual market system can be explained by heterogeneous risk attitudes of customers. In our stylized model, we consider unit demand risk-averse bidders. We show the model admits a unique equilibrium, with higher revenue and higher welfare than using only spot markets. Furthermore, as risk aversion increases, the usage of the on-demand market increases. We conclude that risk attitudes are an important factor in cloud resource allocation and should be incorporated into models of cloud markets.

1 Introduction

Cloud computing allows clients to rent computing resources over the internet to perform a variety of computing tasks, from hosting simple web servers to computing complex financial models. By offloading these tasks to the cloud, clients avoid the necessity of procuring and maintaining expensive servers and infrastructure. The current market leader in this industry is Amazon who launched its cloud platform, Amazon Elastic Compute Unit (EC2), in 2006. Amazon uses its cloud internally for many of its own computations. Additionally, Amazon contracts with large clients who reserve instances of cloud resources for long usage periods. Due to natural variation in the nature of computing tasks from Amazon and its large clients, EC2 has a varying amount of leftover computing resources. Amazon sells these resources to small clients.

This leads to a natural question: how should a cloud provider price its resources to these small clients? The pricing model adopted by Amazon has two main components: an on-demand market and a spot market. In the on-demand market, clients may buy an instance of cloud resources at a fixed reservation price.[1] After resources have been allocated internally, to large clients, and to clients in the on-demand market, extra supply might still remain. This supply is sold in the spot market. In a spot market, clients place bids for instances, and a price is set so that the available supply equals the total demand at that price.

[*] University of Maryland. Part of this work was completed while D. Hoy was an intern at Microsoft Research.

[†] Microsoft Research

[‡] Microsoft Research

[1] This might more naturally be called a “reservation market” and we switch to this terminology in the remainder of the paper; however we stick to the term “on-demand” for the current discussion as this is the term used by Amazon.

Viewed through the lens of microeconomic theory, the persistence of this dual market is a curiosity at first glance. Indeed, in standard economic environments, a risk-neutral, expected-utility-maximizing client who desires a cloud resource should simply buy it in whichever market is expected to have the lower price – typically the spot market. This suggests all sales should happen in the spot market, leaving the on-demand market defunct.

That this is not reflective of reality stems from several factors. Most apparent is that clients are rarely risk-neutral. For example, it is easy to imagine that a company would have a soft budget set aside for computational costs. They would then spend freely within the confines of this budget, and extend the budget cautiously when necessary to meet their computing needs. This type of behavior suggests a tendency towards risk aversion on the part of the clients. As the budget is freely available, clients might prefer to "overspend" to guarantee the required resources at the on-demand price.

We show in a stylized setting that the presence of heterogenous risk attitudes can explain the prevalence of a combined on-demand and spot market. Specifically, this dual market induces a unique equilibrium in which more risk-averse customers (e.g., those with higher budgets) buy resources in the on-demand market and the others bid in the spot market. We show that this equilibrium outcome outperforms the outcomes achieved by running only one of the two types of markets on its own in many key objectives.

1.1 Results and Techniques

In order to highlight the impact of risk aversion on the market, we focus our analysis on a simple setting in which there is only one type of computational resource being sold (e.g., a server with one core for one hour), and each buyer demands only a single instance of this cloud resource at any given time[2]. Formally, we assume a continuum of buyers, where the type of a buyer consists of a value for an instance and a utility curve that maps outcomes (i.e., allocation and price paid) to payoffs. As alluded to above, the utility curve describes a buyer's attitude toward risk: for example, a buyer with a soft budget (as described above) would likely prefer to spend all of their budget all the time than to spend twice their budget half the time, and this preference is captured by a non-linear utility curve. The buyer types are described by a joint common prior. We assume the market is large; i.e., no single buyer has significant impact on the market outcome.

The market works as follows. First, the seller sets a price for on-demand instances. This price should be high enough to guarantee that supply exceeds demand, motivated by the fact that resources are always available for purchase in on-demand markets in practice. Buyers then realize their types (i.e., value/budget pairs) and choose whether to buy in the on-demand market. After these decisions have been made, the unsold supply receives an exogenous shock, modeling variation in the demand of large clients. Any remaining supply is then sold to the remaining buyers at a market-clearing price.

We prove that this system has a unique (subgame-perfect) equilibrium for each choice of the on-demand market price. We do this by analyzing the relationship between the spot price distribution and the distributions of clients' types and corresponding supply and demand. It turns out that the

[2] Of course, this model abstracts away from many reasonable sources of risk aversion in the cloud, such as clients with diminishing marginal returns for multiple instances, or the cost of prematurely terminating a long-running task. Even ignoring these factors, our model still generates heterogeneous preferences toward on-demand versus spot pricing.

distribution over spot prices up to a certain value v depends only on the distributions of clients' types in the range $[0, v]$, and hence one can explicitly solve for the price distribution recursively.

This equilibrium satisfies a monotonicity property: agents that are more risk-averse are more likely to purchase in the on-demand market, whereas agents that are less risk-averse are more likely to use the spot market. Furthermore, as the distribution shifts such that agents become more averse to risk (in the sense of first-order stochastic dominance), we show more clients end up buying instances in the on-demand market and the revenue correspondingly increases. This result is perhaps intuitive, but it is not obvious: as clients become more averse to risk, they shift towards the on-demand market and hence both decrease supply and demand in the spot market. This in turn could cause the spot price to shift either up or down, which would impact purchasing decisions of all clients. By further arguing about the equilibrium of the market, we show the shift towards the on-demand market in fact causes the spot price to increase thereby reinforcing the shift towards the on-demand market. Therefore, the equilibrium is monotone with such shifts in the value and budget distribution. This further illustrates the connection between the on-demand market and risk attitudes.

We leverage our equilibrium characterization to compare the dual market outcome to the outcome of a spot-only or on-demand-only market. We are interested in the welfare, efficiency, and revenue properties of these markets. The revenue of a market outcome is simply the sum of the payments, and is equal to the cloud provider's utility. The welfare of an outcome is the total utility of all market participants including the cloud provider. The efficiency is the total value of the cloud clients, ignoring payments. In risk neutral environments, the welfare and efficiency are equal, but with risk-averse clients the welfare can be less than the efficiency.

It is easy to see that a spot-only market is more efficient than a dual market, which in turn is more efficient than an on-demand-only market. This is because the spot market precisely generates the efficient outcome, even with exogenous supply uncertainty: in a spot market, allocation is always monotone in value and so it is never the case that a lower-valued client is served at the expense of a higher-valued one.

Surprisingly, these efficiency comparisons do not extend to welfare. As we show, the welfare of the dual market is better than the welfare of the spot market alone, regardless of the price set for the on-demand instances. In particular, this is true even when the on-demand market price is set to maximize the revenue of the cloud provider. This is not trivial: the on-demand market adds inefficiency, since clients with high value but low aversion to risk may not wish to purchase on-demand, whereas clients with lower value but higher risk-aversion might. This leads to circumstances where lower-valued clients win but higher-valued clients lose. However, since the clients that are winning in this scenario are actually more risk-averse, the transfer of payments to the cloud provider increases welfare. We show that the welfare increase due to additional transfers from risk-averse clients offset any inefficiencies in the allocation. Moreover, since this welfare comparison holds at every setting of the on-demand price, it applies in particular to the price that maximizes revenue. We show that this price must also generate more revenue than a spot-only market, leading to a simultaneous increase of both welfare and revenue.

In summary, a dual spot/on-demand market simultaneously improves both the revenue and welfare of a spot-only market. We also show by example that while an on-demand market is revenue-optimal for risk-neutral buyers, a dual market can generate strictly higher revenue when buyers are risk-averse. Furthermore, while a dual market may sometimes generate less revenue than an on-demand-only market, a dual market is always more efficient. This suggests that a cloud provider, especially one that holds a dominant position in the marketplace, might prefer a dual market system. This phenomenon is driven by heterogeneous risk attitudes that arise naturally in the context of cloud computation, leading us to posit that risk aversion is an important element to consider when one models the cloud marketplace.

1.2 Related Work

A number of papers explore cloud-computing market design. Zhang et al. [2013] consider designing a truthful auction where uncertainty lies in the arrival of demand and value profiles of bidders, whether they have a large job with deadline or general demand over time. An et al. [2010] design a negotiation-based mechanism for setting price contracts in the presence of demand uncertainty. Borgs et al. [2014] consider the pricing problem faced by a seller setting on-demand prices over multiple time periods and uncertain supply, with agents who arrive and have different deadlines for their tasks. The paradigm of a dual spot+reserve mechanism has also received a lot of attention. Wang et al. [2012] uses a Markov decision process to model the designer's choice of how to partition supply between the reserve and spot markets. Abhishek et al. [2012] models the cloud market as a queuing model, in which a continuum of jobs arrive and have (private) waiting costs. They find that a fixed cost model provides greater expected revenue than a spot market. Additionally, recent works [Menache et al., 2014; Ma and Huang, 2012] have focused on the problem faced by bidders in such a market: when to use the spot market and when to reserve. Ben-Yehuda et al. [2013] analyzes the expected spot prices in comparison to their reservation prices, and find that it is very likely that Amazon is intentionally manipulating the price or supply distribution so as to provide users with more uncertainty in the spot market. In all of the models described above, agents are risk-neutral and do not have budgets. As far as we are aware, our work is the first to use budget heterogeneity, or risk-aversion more generally, to explain the prevalence of a spot+reserve market.

Auctions for cloud computation resources share similarities to electricity markets, where the split between a spot market and a so-called "futures" market is common. Indeed, the use of both markets has been advocated to account for risk-averse buyers and sellers (see e.g., Ausubel and Cramton [2010]). One difference is that, unlike the on-demand market for cloud computation, futures markets for electricity are typically resolved years in advance.

While most work in auction theory assumes risk-neutral agents, some work has been done for auctions with risk-averse bidders. Optimal auctions have been characterized for simple settings [Maskin and Riley, 1984; Matthews, 1983], but the solutions are generally not expressible in closed form. It is therefore more common to study the simple auctions used in practice, with the general finding that second-price or spot-like auctions do poorly for revenue when compared with first-price auctions [Riley and Samuelson, 1981; Hu et al., 2010; Fu et al., 2013]. Matthews [1987] has looked at the preferences of bidders in the auctions, and showed that first-price auctions not only can get more revenue than the second-price auction, but also can be preferred by bidders due to reduction in uncertainty around the payment. Our model of risk-aversion as the presence of a soft budget is closely related to the capacitated utilities model of Fu et al. [2013], where the capacity in their model corresponds to value minus budget in our model. They show that with capacitated agents, a simple first-price auction with reserve has revenue that approximates the revenue of the optimal mechanism.

Our results are of a similar flavor to the eBay-style buy-it-now auction considered by Mathews and Katzman [2006]. As in our model, they find that adding a buy-it-now option increases revenue, and as agents become more risk-averse, the optimal price increases. Their model differs from ours in that they consider an explicitly randomized allocation rule, rather than clearing the market at a spot price, in order to incentivize use of the buy-it-now (i.e., reservation) option.

2 Preliminaries

In our model, a single cloud provider (the seller) is selling compute resources to a continuum of clients (the bidders).

Utility structure Each bidder i has value $v_i \in [0, 1]$ for a single compute instance. As is standard in the economics literature, we model the risk attitude of bidder i through a utility function $u_i : \mathcal{R}_{\geq 0} \to \mathcal{R}_{\geq 0}$. If the bidder obtains an instance and pays p, then her utility is $u_i(v_i - p)$. We will assume that $u_i(0) = 0$, that u_i is continuous and non-decreasing, and that u_i is not identically 0. Note that since $u_i(0) = 0$, we can think of v_i as the maximum price at which bidder i is willing to purchase an instance. A bidder that does not obtain an instance will pay nothing and have utility 0.

We will focus our attention on agents that are *risk averse*. That is, we will assume that utility curves are weakly concave, as is standard when modeling risk aversion. Note that we allow u_i to be linear, in which case bidder i is said to be risk-neutral.

Roughly speaking, an agent with a utility curve that is "more concave" will be more risk averse, in the sense that they are more likely to prefer guaranteed outcomes to uncertain lotteries. More formally, we say that utility function u is more risk averse than u^*, and write $u \preceq u^*$, if for every distribution L over non-negative real values and every fixed value $d \geq 0$, if $\mathbf{E}_{x \sim L}[u(x)] \geq u(d)$, then $\mathbf{E}_{x \sim L}[u^*(x)] \geq u^*(d)$. In other words, if an agent with utility curve u prefers a lottery L over a guaranteed payout of d, then an agent with utility curve u^* would prefer the lottery as well. This defines a partial order over utility curves. Note that, under this definition, all (weakly) concave utility curves are (weakly) more risk-averse than a risk-neutral (i.e., linear) curve. Note also that for twice-differentiable utility curves, u is more risk-averse than u^* if and only if the standard Arrow-Pratt measure of risk-averson is nowhere lower for curve u than for curve u^* Pratt [1964].[3]

Demand structure Types are distributed according to a joint distribution F on pairs (v, u). For ease of exposition, we will assume throughout that F is supported on a finite collection of (v, u) pairs. Write V and U for the (finite) sets of values and utility curves that support F, and for $(v, u) \in V \times U$ we will write $f(v, u)$ for the probability that an agent has type (v, u).

We will use $F(v)$ to refer to the induced distribution over values; that is, $F(v)$ is the probability that an agent's value is at most v. We will assume a large-market condition, which is that the aggregate demand is distributed exactly according to the type distribution F.

Supply structure The supply of instances, q, is unknown to the bidders and seller until the instances are to be allocated. The supply is then drawn from a distribution, Q. We will normalize the supply so that q represents the fraction of the market that can be simultaneously served, hence $q \in [0, 1]$.

[3] We define risk aversion with respect to agent preferences directly, rather than via the Arrow-Pratt measure, to avoid the requirement that utility curves be twice differentiable.

2.1 Auction Formats

We will consider three auction formats in this paper: spot auctions, reservation auctions (previously referred to as on-demand), and combination (or spot+reservation) auctions.

Spot (M^s) One type of auction to run is a *market-clearing* auction, or a spot auction. In this auction, buyers submit bids. A market-clearing price p_s is chosen such that the quantity of bids exceeding p_s is equal to the supply. Under our unit-demand and large market assumptions, it is a dominant strategy for a bidder to submit a bid equal to her value; henceforth we assume the bids in the spot market equal the values. We observe that a market-clearing price always exists in our market, even in the presence of non-linear utilities: with available supply q, and distribution over values F, the market clearing price, written $p_s(q)$, is precisely[4] the value for which $q = 1 - F(p_s(q))$.

Reservation (M^r) In a reservation-only (or "on-demand") market, the auctioneer sets a fixed price p_r per instance, in advance of seeing the realization of supply. Price p_r need not be a market-clearing price. If there is not enough supply to satisfy the demand for instances at this price, then the winning bidders are chosen uniformly at random from among those who wish to purchase.

Spot+Reservation (M^{s+r}) In a spot and reservation market, the auctioneer first sets a fixed price p_r and runs a reservation auction. The remaining inventory of supply (if any) is then sold via a spot auction.

The exact timeline of events in the spot and reservation auction M^{s+r} is as follows:

1. Auctioneer announces reservation price p_r.
2. Bidders realize types $(v_i, u_i) \sim F(v, u)$.
3. Each bidder decides whether to purchase an instance in the reservation auction, indicated by $a^r \in \{0, 1\}$. Let $T = \sum_{v,u} a^r(v, u)\ f(v, u)$ be the total volume of reserved instances.
4. Auctioneer realizes supply $q \sim Q$, and reserved instances are allocated as in the reservation market described above.
5. If $q > T$, the auctioneer runs a market-clearing auction to clear the excess capacity. Let $p_s(q)$ be the resulting market-clearing spot price.

Note that our specification does not ask bidders to decide whether or not to participate in the spot market. The fact that bidders are unit demand, and that the spot auction is truthful under our large market assumption, implies that in equilibrium bidders will bid (truthfully) in the spot auction if (and only if) they don't buy an instance in the reservation auction.

For a given (implicit) strategy profile for mechanism M^{s+r}, we will write $S(p_s)$ for the cumulative distribution function of the resulting spot prices.

[4] This price may not be unique if $q = 0$ or $q = 1$. In these cases we define $p_s(q)$ to be the supremum of prices satisfying the written condition, which will be ∞ for $q = 0$.

Solution concept: subgame-perfect equilibrium For each of these auctions, the solution concept we apply is *subgame-perfect equilibrium.* A strategy profile for a multi-stage game forms a subgame-perfect equilibrium (SPE) if, at every stage t of the game and every possible history of actions by players in previous stages, no agent can benefit by unilaterally deviating from her prescribed strategy from stage t onward.

For the spot auction and reservation auction, there is only one stage of the resulting game and hence equilibria are straightforward: each agent chooses to purchase her utility-maximizing quantity of instances given the specified price.

For mechanism M^{s+r}, we can characterize the SPE as follows. In the second (i.e., spot) stage of the mechanism, the equilibrium condition implies that agents always purchase instances if and only if their value is above the realized spot price. Thus, the only strategic choice to be made by agents is in the first (i.e., reservation) stage of the mechanism, where each agent must select whether to purchase an instance in the reservation market. We will therefore define a strategy profile $\mathbf{s}$ to be a mapping from a type (v, u) to an action $\{0, 1\}$, where $\mathbf{s}(v, u)$ is interpreted as the number of instances to purchase in the reservation market. Note that the distribution over market-clearing prices in the second stage is completely determined by the actions of agents in the first stage, and hence is determined by $\mathbf{s}$. An equilibrium is then a strategy profile such that no agent can benefit by unilaterally deviating from strategy $\mathbf{s}$ (i.e., by reserving more or fewer instances), given the distribution of spot prices implied by $\mathbf{s}$.

2.2 Objectives

We consider three objectives when evaluating mechanisms: revenue (REV), welfare (WEL) and efficiency (EFF).

The revenue of a mechanism M, $REV(M)$, is the sum of the payments made to the auctioneer. Note that for the spot+reservation mechanism,

$$REV(M^{s+r}) = p_r T + \mathbf{E}_{q \sim Q} \left[p_s(q)(q - T) \right] \tag{1}$$

where we used the fact that $T \leq q$ with probability 1.

The welfare of a mechanism $WEL(M)$ is the sum of utilities of all agents, including the auctioneer (whose utility is precisely the revenue of the mechanism).

The efficiency $EFF(M)$ of a mechanism measures the value created, without considering the welfare lost due to the (non-linear) utility functions of agents. For the spot+reservation mechanism:

$$EFF(M^{s+r}) = \mathbf{E}_{(v_i, u_i) \sim F} \left[v_i \cdot s(v_i, u_i) + v_i \cdot (1 - s(v_i, u_i)) \cdot S(v_i) \right]$$

If an agent reserves, her value generated is v_i. If she does not reserve, her value generated is $v_i \cdot S(v_i)$, which is her value times the probability that the spot price is below her value. Note that if agents are risk-neutral (i.e., have the identity function as their utility functions), then $EFF(M) = WEL(M)$.

3 Equilibrium Behavior & Analysis

In this section, we analyze the choices of bidders and use this to characterize equilibrium of the spot+reservation market. We begin by noting the relationship between the spot price distribution and the distributions of supply, type, and reservation demand in equilibrium. Recall that Q denotes the CDF of the supply distribution.

Lemma 1. *Fix strategy profile* $\mathbf{s}$*, let* S *be the distribution of spot prices under* $\mathbf{s}$*, and let* $T(p)$ *be the volume of reserved instances demanded from agents with value at most* p*, under* $\mathbf{s}$*. Then* $\mathbf{s}$ *forms an equilibrium if and only if, for all* p*,*

$$S(p) = 1 - Q\left(1 - F(p) + T(p)\right), \text{ and} \tag{2}$$

$$T(p) = \sum_{\substack{v \in V \\ v \leq p}} \sum_{u \in U} f(v, u) \cdot \mathbb{1}\left[u(v - p_r) \geq \mathbf{E}_{p \sim S}\left[u(\max\{v - p, 0\})\right]\right]. \tag{3}$$

Proof. The probability that the spot price is at most p is exactly the probability that the supply is greater than necessary to satisfy all of the demand for resources from bidders with higher marginal values than p, plus all reservation demand for resources with lower marginal values than p. On the other hand, the volume of reserved resources demanded from agents with value at most p, at equilibrium, is precisely the probability that such an agent will prefer the deterministic reservation outcome to the lottery over outcomes determined by the distribution over spot market prices. □

In light of Lemma 1, we will tend to equate equilibria with the resulting distributions S and T, rather than with an explicit strategy profile $\mathbf{s}$.

Lemma 2. *Purchasing in the reservation stage is monotone in the risk-aversion of* u_i*: if a* (v_i, u_i) *bidder (weakly) prefers to reserve an instance, then a* (v_i, u_i^*) *bidder with* $u_i^* \preceq u_i$ *(weakly) prefers reserving.*

Proof. We begin by considering the special event in which the agent is not allocated an instance even if they reserve, due to the supply being insufficient to honor all reservations and the agent not being selected randomly as a winner. If this event occurs, the bidder's utility will necessarily be 0, and this is independent of their utility curve and their chosen action (since, if $q < T$, no agent that enters the spot market will obtain an instance).

It therefore suffices to consider the agent's expected utility conditional on the event that the agent will be allocated an instance with certainty if they choose to reserve. With this in mind, the utilities of a unit demand agent from reserving or participating only in the spot market, respectively, are

$$u^r(v_i, u_i) = u_i(v_i - p_r),$$
$$u^s(v_i, u_i) = \mathbf{E}_{p_s \sim S}\left[u_i(v_i - p_s) \cdot \mathbf{1}_{p_s \leq v_i}\right].$$

We want to show that if $u^r(v_i, u_i) \geq u^s(v_i, u_i)$, then we have $u^r(v_i, u_i^*) \geq u^s(v_i, u_i^*)$ as well.

Note that, fixing v_i, the spot market generates a certain lottery L over values $(v_i - p_s)$, and the reservation market generates a certain value $v_i - p_r$. Thus, from the definition of risk aversion, if an agent with utility curve u_i prefers the certain outcome to the lottery L, and $u_i^* \succeq u_i$, then an agent with utility curve u_i^* prefers the certain outcome as well. □

3.1 Equilibrium Existence and Uniqueness

We are now ready to establish uniqueness of equilibrium. One subtlety about equilibrium uniqueness is the manner in which buyers break ties. If a positive mass of agents is indifferent between the spot and reservation markets, there may be multiple market outcomes consistent with those preferences. We will therefore fix some arbitrary manner in which bidders break ties, which could be randomized and heterogeneous across bidders. Our claim is that for any such tie-breaking rule, the resulting equilibrium will be unique.

Lemma 3. *There is a unique equilibrium of M^{s+r}. Moreover, this equilibrium can be computed in time proportional to the size of the support of type distribution f.*

Proof. As shown in Lemma 1, the challenge of characterizing equilibrium essentially reduces to characterizing the fraction of bidders who reserve at a given price, $T(p)$. This is because T determines the distribution S over spot prices, and S (together with an arbitrary tie-breaking rule) uniquely determines the strategy profile s, since this can be inferred from the expected utility when choosing the spot market. Thus, to show uniqueness and existence of equilibrium, it suffices to show uniqueness of the functions S and T.

We will prove that, for all $v \in V$, $T(v)$ and $S(v)$ are uniquely determined by the functions T and S restricted to values less than v. The result will then follow by induction on the elements of V.

Consider first an agent with value $v = \min V$. Recall that the spot price is always at least v. Thus, if $p_r < v$ then the agent will always reserve, if $p_r > v$ then the agent will always choose the spot market, and if $p_r = v$ the agent will be indifferent and apply the fixed tie-breaking rule. In each case, the value of $T(v)$ is uniquely determined, and thus $S(v)$ is as well.

Now choose $v > \min V$, and suppose T and S are determined for all elements of $V \cap [0, v)$. We claim that the distribution of the random variable $\max\{v - p_s, 0\}$, where p_s is distributed according to S, is then uniquely determined. This is because the non-zero values of this random variable are distributed according to S restricted to values in $V \cap [0, v)$. But, by Lemma 1, this random variable determines the value of $T(v)$, which in turn determines the value of $S(v)$. Thus $T(v)$ and $S(v)$ are uniquely determined by S and T on $V \cap [0, v)$, as required. Moreover, these quantities can be explicitly computed by evaluating the summation in Lemma 1. □

3.2 An Example: Soft Budgets

In this section we present a special case of risk-aversion, driven by soft budgets, and give an interpretation of our equilibrium characterization for this case.

Suppose that each buyer i is characterized by their value v_i for a compute instance and a soft budget $b_i \in [0, v_i]$. We think of b_i as a budget of funds that has been allocated to acquiring a compute instance. If the buyer obtains an instance but pays less than b_i, the residual budget is lost: it is as if they had paid b_i. On the other hand, if the buyer pays more than b_i, they suffer no additional penalty; they simply incur the cost of their payment.

This scenario is captured by the following utility curve, which is piecewise linear with two pieces:

$$u_i(z) = \begin{cases} z & \text{if } z \leq v_i - b_i \\ v_i - b_i & \text{otherwise.} \end{cases}$$

To see that this utility curve is equivalent to the soft budget scenario described above, note that if a buyer with budget b_i obtains an instance and pays $p_i > b_i$ their utility is $u_i(v_i - p_i) = v_i - p_i$, whereas if they pay $p_i < b_i$ their utility is $u_i(v_i - p_i) = v_i - b_i$.

Note that for a fixed value v_i, an agent with a higher budget is more risk-averse. To see this, note that if a buyer *strictly* prefers a lottery L to a deterministic outcome d, then it must be that $d < v_i - b_i$ (since otherwise d must be utility-maximizing). In this case, decreasing the budget can only make the lottery more valuable, while not affecting the utility from the deterministic outcome. Thus, a decreased budget can only increase the propensity to select a lottery over a deterministic outcome.

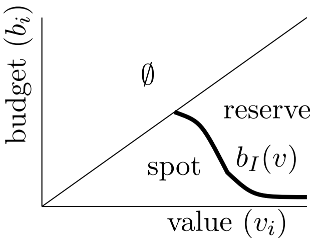

Figure 1: With soft-budgets, a monotone-decreasing indifference curve partitions agents into those that reserve an instance and those who rely on the spot market.

The monotonicity result in Lemma 2 therefore results in a partitioning of agents that prefer the spot market to the reservation market and vice versa by an indifference curve over budgets. See Figure 1 for an illustration. Any agent with (v_i, b_i) below the curve prefers the spot market; any agent above the curve (such that $v_i \geq b_i$), prefers the reservation market. Lemma 3 shows that this indifference curve is unique, given the distribution over agent types, and precisely specifies which agents will choose to reserve and which will choose to enter the spot market

4 Comparative Statics

In this section, we first consider the impact of changes to buyer risk attitudes. We show that as agents become more risk averse, more agents use the reservation market and revenue increases, for every setting of the reservation price. Second, we compare the reservation+spot mechanism to the spot market and the reservation market. We first show that the combination mechanism's outcomes are more efficient than running a reserve market on its own. We then show that it generates both more revenue and more welfare than running only a spot market.

The results in this section hold under two assumptions on the reservation price set by the seller. First, we will make the assumption that p_r is set high enough so that, in the resulting equilibrium, $\Pr_{q \sim Q}[q < T] = 0$. That is, over the uncertainty in supply, the mechanism can serve the reserved instances with certainty. This assumption is motivated by the fact that these instances are typically viewed as guaranteed by the mechanism.

Another natural and practical property is that the reservation price p_r be set high enough that it will be greater than the expected spot price. That is, p_r is large enough that it is more costly, in expectation, to reserve a guaranteed instance than to bid for an instance in the spot market.

4.1 Effect of Increased Risk Aversion

We consider the impact of an increase in customer risk aversion. Consider type distribution F and a type distribution F^+ induced by a pointwise transformation $g^+(U, V) \rightarrow (U, V)$ applied to each point in F which weakly increases risk aversion and does not affect values. Specifically, for any $(u^+, v^+) = g^+(u, v)$, $v^+ = v$ and $u^+ \preceq u$. In the following lemma, proved in the appendix, we show that such a change can only increase the fraction of agents who choose to reserve, and can only increase revenue.

Lemma 4. *For mechanism M^{s+r}, and for any reserve price p_r, if risk aversion of agents increases then the fraction of agents who purchase in the reservation stage increases, as does the expected* *revenue of the mechanism.*

The intuition underlying Lemma 4 is as follows. The first order effect from a change in risk aversion is an increase in T, the fraction of users who choose to reserve at a given price. This increase in reservations translates into higher spot prices, since it reduces the quantity sold in the spot market. Higher spot prices in turn cause more users to prefer to reserve, which can only increase spot prices further. This can be shown by induction over agent values.

4.2 Comparing Mechanisms

We now compare welfare and revenue of M^{s+r} to M^s and M^r. Here we make use of the two assumptions discussed in Section 2.1: 1) the reservation phase is not oversubscribed, i.e., the reservation price is set sufficiently high that there will be sufficient supply to fulfill the demand for reserved instances; and 2) the reservation price is sufficiently high to be above the expected spot price.

We begin by comparing the revenue of the combination mechanism with the expected revenue of a spot-only market. Note that, trivially, the best revenue of the combined mechanism is at least the revenue of a spot market; this is because, in the combined mechanism, the reserve price can be set sufficiently high that all customers buy in the spot market. We show something stronger: for *every* choice of reservation price, the revenue of the combined mechanism is at least that of a spot market run in isolation. The formal proof of the following lemma is deferred to the full version.

Lemma 5. *For any choice of the reservation price satisfying our assumptions, the expected revenue of the reservation and spot mechanism is weakly greater than the revenue of the spot-only market.*

Proof (sketch). As in Lemma 4, as risk aversion increases, revenue increases for a fixed reservation price. Fix a reservation price and consider starting with a distribution of risk-neutral agents. These agents will all bid in the spot market and thus the outcome (and in particular, the revenue) will be identical to the spot-only mechanism. By deforming the utility curves of the agents in a manner that only increases risk aversion, until they match the correct distribution, and applying Lemma 4, we can conclude that the revenue of the dual market only increases while the revenue of the spot-only mechanism, which is unaffected by the utility curves, remains the same. □

We next provide an example demonstrating that the expected revenue of the reservation+spot mechanism can be strictly greater than the revenue of the reservation market, when agents are not necessarily risk-neutral.

Example 1. *We will consider an example in which agent utility curves are specified by soft budgets, as in Section 3.2. Recall that a budget of* 0 *corresponds to risk-neutrality. Take* $\epsilon > 0$ *to be sufficiently small, and consider the following distribution over buyer types:*

- *with probability* $0.5 - \epsilon$, $(v, b) = (5, 0)$
- *with probability* 0.5, $(v, b) = (10, 10 - \epsilon)$
- *with probability* ϵ, $(v, b) = (20, 0)$

The supply is distributed such that $q = 1 - \epsilon$ *with probability* 0.8, *and otherwise* $q = 0.5 + \epsilon/2$.

The optimal reserve price for the reservation market is 10, *which generates a total revenue of* $(0.8) \times 10 \times (0.5 + \epsilon) + (0.2) \times 10 \times (0.5 + \epsilon/2) = 5 + 9\epsilon$.

Next consider the reservation and spot mechanism with reservation price $10-\epsilon$. At equilibrium, the buyers of type $(10, 10-\epsilon)$ strictly prefer to reserve (obtaining utility ϵ with probability 1, rather than utility ϵ with probability 0.8), whereas the buyers of type $(20, 0)$ strictly prefer to participate in the spot market (obtaining utility $(20-5)$ with probability 0.8, rather than utility $10+\epsilon$ with probability 1). The revenue from the reservation market is then $(0.5) \times (10-\epsilon)$, and the revenue from the spot market is $(0.8) \times 5 \times (0.5-\epsilon) + (0.2) \times 20 \times (\epsilon/2)$, for a total of $7 - \frac{5}{2}\epsilon$. This is greater than $5 + 9\epsilon$ for sufficiently small ϵ.

We next compare the efficiency of the dual mechanism to the efficiency of the reservation-only mechanism.

Lemma 6. *For any choice of the reserve price satisfying the assumptions above, the expected efficiency of the reservation+spot mechanism is weakly greater than the efficiency of the reservation market with the same reservation price.*

Proof. Recall that the efficiency of a mechanism is the expected value generated by the agents, ignoring the welfare lost due to the nonlinear utility functions. For any realized supply q then, weakly more people are served in M^{s+r}, hence efficiency is greater. □

Finally, we will show that the social welfare (sum of utilities) of the combined spot+reservation mechanism is only greater than the welfare of the spot-only mechanism. The formal proof of the following theorem appears in the appendix.

Theorem 7. *In any equilibrium of the the spot and reservation mechanism where the reservation price is set above the expected spot price, the expected welfare of the reservation and spot mechanism is weakly greater than the expected welfare of the spot-only mechanism.*

Proof (sketch). Note that, relative to a spot market, introducing a reservation price adds inefficiency. This is because if a bidder is willing to reserve to get a guaranteed instance, any time they would not have one in the spot market, there is a higher valued bidder than them who could be allocated.

However, welfare is increased when a bidder chooses to reserve. Consider an agent with value v_i who is willing to reserve at price p_r. Reserving increases his utility — because he prefers reserving — and the auctioneer is receiving more in revenue, because the reservation price is greater than the expected spot price (by assumption).

The full proof consists of three parts. First, we define a benchmark B^{s+r} that is just like M^{s+r} except agents who reserve pay the spot price instead of the reservation price. Then, we show that the welfare of M^{s+r} is greater than the welfare of B^{s+r}, which will follow largely because the expected reservation price is greater than the expected spot price. Finally, we show that spot prices increase when agents choose to reserve, which leads to the welfare of B^{s+r} being greater than the welfare of M^s, and hence the welfare of M^{s+r} is greater than M^s. □

References

Abhishek, V., Kash, I. A., and Key, P. (2012). Fixed and market pricing for cloud services. *CoRR*, abs/1201.5621.

An, B., Lesser, V., Irwin, D., and Zink, M. (2010). Automated negotiation with decommitment for dynamic resource allocation in cloud computing. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 981–988. International Foundation for Autonomous Agents and Multiagent Systems.

Ausubel, L. M. and Cramton, P. (2010). Using forward markets to improve electricity market design. *Utilities Policy*, 18(4):195–200.

Ben-Yehuda, O. A., Ben-Yehuda, M., Schuster, A., and Tsafrir, D. (2013). Deconstructing Amazon EC2 Spot Instance Pricing. *ACM TEAC*, 1(3):1–20.

Borgs, C., Candogan, O., Chayes, J., Lobel, I., and Nazerzadeh, H. (2014). Optimal Multiperiod Pricing with Service Guarantees. *Management Science*, 60:1792–1811.

Fu, H., Hartline, J., and Hoy, D. (2013). Prior-independent Auctions for Risk-averse Agents. In *EC'13*.

Hu, A., Matthews, S. A., and Zou, L. (2010). Risk aversion and optimal reserve prices in first- and second-price auctions. *Journal of Economic Theory*, 145(3):1188–1202.

Ma, D. and Huang, J. (2012). The Pricing Model of Cloud Computing Services. In *EC'12*, pages 263–269.

Maskin, E. and Riley, J. (1984). Optimal auctions with risk averse buyers. *Econometrica*, 52(6):1473–1518.

Mathews, T. and Katzman, B. (2006). The role of varying risk attitudes in an auction with a buyout option. *Economic Theory*, 27(3):597–613.

Matthews, S. A. (1983). Selling to risk averse buyers with unobservable tastes. *Journal of Economic Theory*, 30(2):370–400.

Matthews, S. A. (1987). Comparing auctions for risk averse buyers: A buyer's point of view. *Econometrica*, 55(3):633–646.

Menache, I., Shamir, O., and Jain, N. (2014). On-demand, Spot, or Both: Dynamic Resource Allocation for Executing Batch Jobs in the Cloud. In *ICAC'14*.

Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica*, 32(1):122–136.

Riley, J. and Samuelson, W. (1981). Optimal Auctions. *The American Economic Review*, 71(3):381–392.

Wang, W., Li, B., and Liang, B. (2012). Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing. In *ICDCS'12*, pages 425–434. Ieee.

Zhang, H., Li, B., Jiang, H., Liu, F., Vasilakos, A. V., and Liu, J. (2013). A framework for truthful online auctions in cloud computing with heterogeneous user demands. In *IEEE INFOCOM*, pages 1510–1518. Ieee.

A Appendix: full version proofs

A.1 Proof of Lemma 4

Proof. Let F be the original distribution over types, and let $\overline{F}$ denote a distribution in which the risk aversion of agents increases, for each fixed value v.Let $\overline{T}$ and $\overline{S}$ refer to the volume of reserved instances and distribution of spot prices under the alternate distribution $\overline{F}$. As before, b_I, T, and S correspond to the respective quantities for the original distribution F.

We will show that for any price $p \geq p_r$, it is the case that $\overline{T}(p) \geq T(p)$ and $\overline{S}(p) \leq S(p)$. To see how this implies the lemma, note that these inequalities imply that the increase in risk aversion of agents results in (a) more agents reserving, and (b) higher spot prices. Because the reservation price is greater than the expected spot price, the now-reserving agents pay more than they were paying before and total revenue increases.

We will prove the desired inequalities by induction on $p \in V$. Note first that for any $p < p_r$, an agent with value p will certainly choose to purchase in the spot market regardless of their risk attitudes, and hence $\overline{T}(p) = T(p)$ and $\overline{S}(p) = S(p)$. For any $p \geq p_r$, recall that

$$T(p) = \sum_{\substack{v \in V \\ v \leq p}} \sum_{u \in U} f(v,u) \cdot \mathbb{1}\left[u(v - p_r) \geq \mathbf{E}_{p \sim S}\left[u(\max\{v-p, 0\})\right]\right]$$

and

$$\overline{T}(p) = \sum_{\substack{v \in V \\ v \leq p}} \sum_{u \in U} \overline{f}(v,u) \cdot \mathbb{1}\left[u(v - p_r) \geq \mathbf{E}_{p \sim \overline{S}}\left[u(\max\{v-p, 0\})\right]\right]$$

As in Lemma 3, the dependence on S and $\overline{S}$, respectively, is limited to values in $V \cap [0, v)$. Since we assume inductively that $S(p) \geq \overline{S}(p)$ for all $p \in V \cap [0, v)$, the distribution over spot market utilities under $\overline{F}$ is stochastically dominated by the distribution under F, for agents with value v. Since $\overline{F}$ additionally increases risk aversion relative to F, fixing v, we conclude that weakly fewer agents with value v would choose to reserve under $\overline{F}$ relative to F. That is, $\overline{T}(p) \geq T(p)$. Then, since

$$S(p) = 1 - Q\left(1 - F(p) + T(p)\right)$$

and

$$\overline{S}(p) = 1 - Q\left(1 - F(p) + \overline{T}(p)\right)$$

we can immediately conclude $\overline{S}(p) \leq S(p)$ as well. The desired result then follows by induction. □

A.2 Proof of Theorem 7

We now proceed with the proof of Theorem 7: that the welfare of the spot+reservation mechanism is greater than the welfare of the spot-only mechanism. Let $WEL_q(M)$ be the welfare of mechanism M at fixed supply q. Then we can write the total welfare of M as

$$WEL(M) = \mathbf{E}_{q \sim Q}\left[WEL_q(M)\right] \tag{4}$$

Write $w_i^{v_i}(p)$ for the welfare generated by an agent paying price p, hence $w_i^{v_i}(p) = u_i(v_i - p) + p$. Note that $w_i(p)$ is monotone non-decreasing in p, since u_i is a concave non-decreasing function with $u_i'(0) = 1$.

Let $p_s^s(q)$ be the price in the spot-only mechanism M^s if the realization of supply is q, and $p_s^{s+r}(q)$ the spot price in spot and reservation mechanism M^{s+r}. Note that $p_s^{s+r}(q) \geq p_s^s(q)$ for all q, no matter how the buyers behave. This is because $p_s^s(q)$ is precisely the minimal price at which a q fraction of the buyers will purchase, and hence if $p_s^{s+r}(q) < p_s^s(q)$ then more than a q fraction of buyers must be purchasing in M^{s+r} (since they would buy in the spot market if they didn't reserve), which is impossible.

We now analyze the welfare of M^{s+r}, breaking down the welfare generated by those bidders who buy in the spot market and who reserve. Let $Y(v)$ be the fraction of bidders with value v who reserve. Then

$$WEL_q(M^{s+r}) = \sum_{v \leq p_s^{s+r}(q)} w_i^v(p_r)Y(v)f(v) \tag{5}$$

$$+ \sum_{v > p_s^{s+r}(q)} \left(w_i^v(p_s^{s+r}(q))(1-Y(v)) + w_i^v(p_r)Y(v)\right) f(v). \tag{6}$$

Note the split between the two summations at $p_s^{s+r}(q)$. Agents with values below $p_s^{s+r}(q)$ are served only if they reserve, whereas agents with values above $p_s^{s+r}(q)$ are served whether or not they reserve — the only question for their welfare is whether they pay p_r or p_s^{s+r}. Call each of these summations $WEL_q^-(M^{s+r})$ and $WEL_q^+(M^{s+r})$ respectively.

We now consider the welfare from a benchmark, B^{s+r}. For a given supply q, if agents reserve and have values above $p_s^{s+r}(q)$, we assume that they pay the spot price $p_s^{s+r}(q)$ instead of the reservation price p_r. For agents who reserve with values below the spot price $p_s^{s+r}(q)$, we assume they pay the spot price $p_s^{s+r}(q)$ and (magically) get no utility or disutility from doing so, hence the only welfare generated is the welfare the designer experiences from the payment: $w_i^z(p_s^{s+r}(q)) = p_s^{s+r}(q)$.

As we did for the combined mechanism, we can write the welfare of the benchmark as

$$WEL_q(B^{s+r}) = WEL_q^-(B^{s+r}) + WEL_q^+(B^{s+r}),$$

where $WEL_q^-(B^{s+r})$ denotes the contribution to welfare from bidders with values below $p_s^{s+r}(q)$, and $WEL_q^-(B^{s+r})$ is the contribution to welfare from bidders with values above $p_s^{s+r}(q)$. For bidders with values below the spot price, we have

$$\begin{aligned} WEL_q^-(B^{s+r}) = & \sum_{v \leq p_s^{s+r}(q)} p_s^{s+r}(q)Y(v)f(v) \\ = & p_s^{s+r}(q)T(p_s^{s+r}(q)) \end{aligned}$$

The last line followed because $T(p_s^{s+r}(q))$ is exactly the volume of agents with value at most $p_s^{s+r}(q)$, hence $T(p_s^{s+r}(q)) = \sum_{v \leq p_s^{s+r}(q)} Y(v)f(v)$. For the bidders with values above the spot price, the welfare satisfies

$$WEL_q^+(B^{s+r}) = \sum_{v > p_s^{s+r}(q)} w_i^v(p_s^{s+r}(q))f(v).$$

Whether or not the benchmark welfare for agents with values above the spot price is above or below the actual welfare depends on whether the spot price is above or below the reservation price. But, as the following claim shows, the welfare benchmark will be less in expectation than the actual welfare:

Claim 1.

$$\mathbf{E}_{q\sim Q}\left[WEL_q(M^{s+r})\right] \geq \mathbf{E}_{q\sim Q}\left[WEL_q(B^{s+r})\right] \tag{7}$$

Proof. The expected payment from every agent who reserves in B^{s+r} is $\mathbf{E}_{q\sim Q}[p_s^{s+r}(q)]$, which by assumption satisfies $\mathbf{E}_{q\sim Q}[p_s^{s+r}(q)] \leq p_r$. Thus the total revenue in the mechanism M^{s+r} is greater than in the benchmark B^{s+r}.

For all agents who do reserve, we know their utility from reserving is more than their utility if they had not reserved and only participated in the spot market, $(u^r(v_i, u_i) \geq u^s(v_i, u_i))$. The utilities of all agents who do not reserve are the same in both, so they are indifferent.

Summing the utilities of all the agents and the revenue of the designer gives

$$\mathbf{E}_{q\sim Q}\left[WEL_q(M^{s+r})\right] \geq \mathbf{E}_{q\sim Q}\left[WEL_q(B^{s+r})\right].$$

□

We now argue that the welfare of the benchmark is an upper bound on the welfare from the spot-only mechanism.

Claim 2.

$$WEL_q(B^{s+r}) \geq WEL_q(M^s) \tag{8}$$

Proof. The agents who purchase in M^s are precisely those with values above $p_s^s(q)$, the spot price for the spot-only mechanism. Consider separately the agents with values above and below the spot price $p_s^{s+r}(q)$ for the combined mechanism.

Agents with values above $p_s^{s+r}(q)$ are always allocated in the benchmark and the spot-only mechanism. As the spot price (and hence benchmark payment) is higher in the spot+reservation mechanism than the spot-only mechanism for a given supply, and w_i is non-decreasing in p, $w_i(p_s^{s+r}(q)) \geq w_i(p_s^s(q))$, thus

$$\begin{aligned}\sum_{v>p_s^{s+r}(q)} w_i^v(p_s^s(q))f(v) &\leq \sum_{v>p_s^{s+r}(q)} w_i^v p_s^{s+r}(q)f(v) \\ &= WEL_q^+(B^{s+r}). \end{aligned} \tag{9}$$

Consider now agents with values below $p_s^{s+r}(q)$. If the agent does not reserve in the spot+reservation mechanism, then we know that the price paid in the benchmark is higher than the welfare generated from receiving the item: $v_i \leq p_s^{s+r}(q)$, hence we know that $w_i^z(p_s^s(q)) \leq p_s^{s+r}(q)$. Recall that $T(p_s^{s+r}(q))$ is the volume of agents with values below $p_s^{s+r}(q)$ who reserve. Thus,

$$\begin{aligned}\sum_{v|p_s^s(q)<v\leq p_s^{s+r}(q)} w_i^z(p_s^s(q))f(v) &\leq \sum_{v|p_s^s(q)<v\leq p_s(q)} p_s^{s+r}(q)f(v) && (10)\\ &= p_s^{s+r}(q)(S(p_s^{s+r}(q)) - S(p_s^s(q))) \\ &\leq p_s^{s+r}(q)T(p_s^{s+r}(q)) && (11)\\ &= WEL_q^-(B^{s+r}). && (12)\end{aligned}$$

Combining equations (12) and (9) gives:

$$\begin{aligned}WEL_q(M^s) &\leq WEL_q^+(B^{s+r}) + WEL_q^-(B^{s+r}) \\ &= WEL_q(B^{s+r}),\end{aligned}$$

our desired result. □

We can now combine the bounds from Claim 1 and Claim 2 to show that the welfare of the spot and reservation mechanism M^{s+r} is greater than the welfare of the spot-only mechanism, M^s, completing the proof of Theorem 7.

Proof of Theorem 7. Taking expectation over the supply and using the claims above, we have

$$WEL(M^s) \leq WEL(B^{s+r}) \leq WEL(M^{s+r})$$

as desired. □