Manuscript - Fixed and Market Pricing for Cloud Services

by Vineet Abhishek, Ian Kash, Peter Key

Mirko Richter, University of Zurich

April 21, 2017

1 Introduction

Cloud Services provide a customer with infrastructure as a service (IaaS). Current provider differentiate mainly between two different pricing mechanism. Either the customer pays a fixed price for his usage known as pay as you go (PAYG), where the customer pays a fixed price per time and cloud instance. In contrast to the PAYG mechanism, the customer has the option to participate in a spot market that sells unused resources via an auction. The main difference between those options, besides the price, is that in the PAYG mechanism the customer can be certain that he can use his instance whenever he wants without interruptions or having to wait. In the spot market, the customer may has to wait to use the instance and it is possible that the resource becomes unavailable causing an interruption.

(Abhishek et al. 2012) takes the view of a provider of cloud services and wants to find the better mechanism in terms of revenue. They compare the revenue of operating only a PAYG mechanism with a hybrid market of a PAYG mechanism and a spot market. Intuitively it is not clear what option produces more revenue. On one hand, the spot market can create a price discrimination which allows the provider to serve customers who would not participate in the isolated PAYG mechanism, but on the other hand, customer already participating in the PAYG mechanism might switch to the spot market when they can choose and cause a loss of revenue for the provider.

On of the main contributions of the paper is the characterization and analysis of the BNE in the spot and hybrid market that is independent from the auction design. So the results are not bound to a specific auction mechanism.

As one of the main provider of cloud services, you can see in figure 1, that Amazon in fact provides a hybrid version, so you can choose between a fixed price service or participate in an auction in the spot market. We will come back in the Conclusion with ideas why Amazon uses the hybrid version.

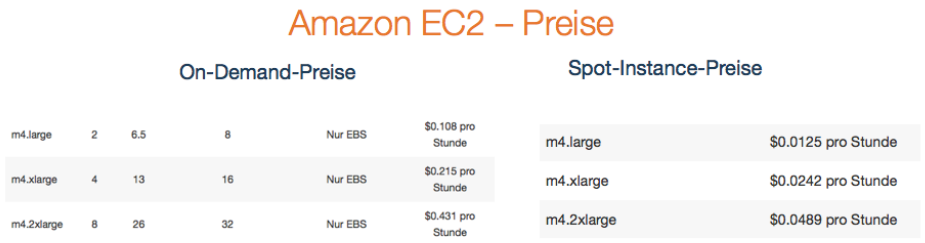

Figure 1: Amazon prices for PAYG and spot market.

2 Model

We categorise jobs in classes according to their values. So all jobs in class i have the same value v_i. The waiting cost is based on a random variable with a cumulative density function. The value and waiting cost are the job's private information, while the class values and the cumulative density function are common knowledge. The arrival rate and expected service time are the same for PAYG and spot market.

Jobs are rational and are trying to maximize their payoff. The cost of a job consist of the payment m and the waiting cost cw, which consist of the waiting and service time times waiting cost. So the payoff of a job in class i with waiting cost c is $v_i - cw - m$.

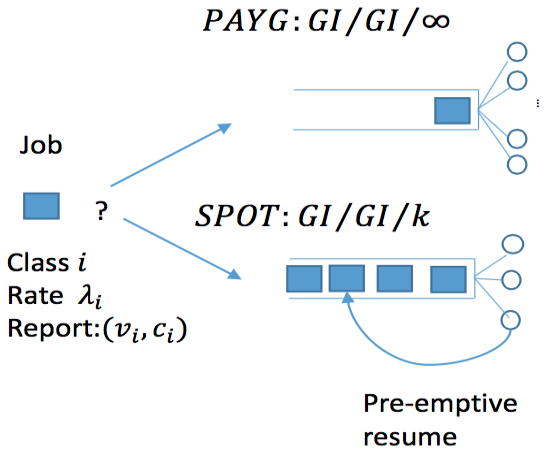

Figure 2: Queue Model for Cloud Service.

We assume that the system has enough capacity to handle the incoming jobs, thus we consider a $GI/GI/\infty$ queue. There is no waiting time in the PAYG model, the jobs are served directly and with no interruptions. Every job pays price $p > 0$ for instance and time unit and the payoff is $v_i - (p + c)/\mu$ with μ being the service rate. This leads to $c \leq \mu v_i - p$ as participation constraint.

The spot market is modelled as a GI/GI/k queue with priority order over the jobs in the queue. The auction underlying the spot market defines what priority a job gets as

well as the expected payment of a job. In contrast to the PAYG mechanism, the job with the lowest priority can be preempted and waits in queue for continuation.

3 Spot Pricing

One of the main contributions of (Abhishek et al. 2012) is that the revenue comparison of the PAYG and the Spot market is not tied to a particular pricing rule of the auction in the spot market.

As long as there are at least two classes and incomplete information, the revelation principle allows us to focus solely on direct truthful mechanism. So we can assume without loss of generality that agents report their true waiting costs. Furthermore, these waiting costs are independent from the agents value or class. This leads to the following lemma.

Lemma 2. [1] *In equilibrium, the expected waiting time $\tilde{w}(c)$ is non increasing in c and the expected payment $\tilde{m}(c)$ is non-decreasing in c.*

Proof. Consider $\hat{c} > c$. Equilibrium with truthful implies:

$$\hat{c}\tilde{w}(\hat{c}) + \tilde{m}(\hat{c}) \leq \hat{c}\tilde{w}(c) + \tilde{m}(c) \tag{1}$$

$$c\tilde{w}(c) + \tilde{m}(c) \leq c\tilde{w}(\hat{c}) + \tilde{m}(\hat{c}) \tag{2}$$

Adding (1) and (2) gives us $\tilde{w}(\hat{c}) \leq \tilde{w}(c)$. This and (2) then give us $\tilde{m}(c) \leq \tilde{m}(\hat{c})$. □

Then (Abhishek et al. 2012) uses two assumption that are not without loss of generality. First it is now assumed that $\tilde{m}(0) = 0$, so jobs with waiting cost equal to 0 are served for free in the spot market. Secondly, in equilibrium in the spot market, jobs with higher waiting cost have strictly higher priority.

We will later see that the expected payment in the spot market is independent from the job's value. But the job's value determines its class and every class has a waiting cost cutoff. Every agent in this class with lower waiting cost than the cutoff participates in the spot market, while for waiting costs above the cutoff, the agents will not participate.

Since we want to characterize an equilibrium using the waiting time function with cutoffs w(c;**c**), with **c**= $(c_1, c_2, .., c_n)$ marking the cutoff for every class, we need to characterize some properties of w(c;**c**).

(i) w(c;**c**) is decreasing in c for $c \in [0, \max_i c_i]$, $w(c; \mathbf{c}) > 1/\mu$ if $c < \max_i c_i$ and $w(c; \mathbf{c}) = 1/\mu$ if $c \geq \max_i c_i$

(ii) w(c;**c**) is increasing in c_i for $c_i \in [0, \mu v_i]$

(iii) For any $t \geq \hat{c}_j > c_j : w(t; \mathbf{c}) = w(t; (\hat{c}_j, \mathbf{c}_{-j}))$

[1] Numbering of Lemma and Theorems are consistent with (Abhishek et al. 2012).

The first property assures that higher waiting cost leads to higher priority, thus lower waiting time. Further, the expected waiting time is greater than zero for waiting cost smaller than the maximum and the agent with the highest waiting cost does not have to wait. The second property says that more jobs increase the waiting time but then the third assures that only the waiting time of jobs with smaller waiting cost are affected by new jobs.

Now, suppose a BNE in the spot market with truthful reporting and cutoffs $\mathbf{c}$. Then the following incentive compatibility constraint must hold for all $\forall c, \hat{c} \leq \max_i c_i$ and any i,

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

m(c) being the expected payment by an agent with waiting cost c.

We are now able to proof a condition for the expected payment function that is necessary in order to satisfy the incentive compatibility constraint.

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

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

Hence, the expected payment by a job with waiting cost c is uniquely determined by the function w. Moreover, the properties of w(c;$\boldsymbol{c}$) together with (4) *satisfy the incentive compatibility* (3)*.*

Proof. Let

$$\pi(\hat{c}, c) \triangleq v_i - cw(\hat{c}, \mathbf{c}) - m(\hat{c})$$

We know that to satisfy (3) in a truthful BNE, the $\max_{\hat{c}} \pi(\hat{c}, c)$ must be achieved in $\hat{c} = c$. Otherwise, the agent would have an incentive to misreport $\hat{c}$ instead of c. Since π as a function is affine, it follows that $\pi(c, c) = \max_{\hat{c}} \pi(\hat{c}, c)$ is convex, and hence is differentiable almost everywhere, with right derivative:

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

Integrating over [0,c] and multiply by -1:

$$-\pi(c, c) = \int_0^c w(t; \mathbf{c})dt + c$$

Substitute $\pi(c, c)$ with $v_i - cw(c, \mathbf{c}) - m(c)$ and rearrange to the format "m(c) = ":

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

Now we have an expected payment of $v_i + c$ for waiting cost $c = 0$

$$m(0) = 0 + 0 + v_i + c$$

But we need to satisfy our assumption $m(0) = 0$, hence we get the condition we are looking for:

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

□

It is now proven that there is a necessary condition between the waiting time function and the expected payment. And this is not dependent on a particular auction mechanism for the spot market. Instead, it proofs the expected payment to be independent from a particular auction mechanism design, hence the later analysis using this expected payment is also not dependent on a particular auction mechanism design.

4 Revenue and Equilibria for isolated Markets

Considering the PAYG market in isolation. As stated before the expected payoff for a job from class i and price p is $v_i - (p + c)/\mu$. The participation constraint is $c \leq \mu v_i - p$. Then the effective arrival rate for class i jobs is $\lambda_i F_i(\mu v_i - p)$ where $F_i(\mu v_i - p) = 0$ for $p \geq \mu v_i$. F(x) being the cumulative density function. With expected job duration of $1/\mu$ and price p per time unit, we get the following revenue

$$R^{PAYG}(p) \triangleq \frac{p}{\mu}\left(\sum_i \lambda_i F_i(\mu v_i - p)\right) \tag{5}$$

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

If we now switch to the spot market in isolation, the expected payoff for a job in class i is $v_i - (cw(c; \mathbf{c}) + m(c))$ and by using (4), we get $v_i - \int_0^c w(t; \mathbf{c})dt$. Again, the agent will participate if the payoff is greater than zero and with our definition of a cutoff for each class we need a cutoff vector c^s that satisfies:

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

Theorem 1 now shows that there is a unique cutoff vector c^S satisfying (6) that characterizes the BNE for a spot market in isolation.

Theorem 1. *The following holds:*

(I) There is a unique solution c^S to the following system of equations in $x = (x_1, x_2, .., x_n)$

$$\int_0^{x_i} w(t; x)dt = v_i \tag{7}$$

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

Proof. We start with a new Lemma 7 that we will not proof.[2]

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

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

satisfies

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

for all $j \geq i$.

We use lemma 7 to proof *(I)* directly. This gives us c^S and the expected payoff from jobs with $c \leq c_i^S$ in the spot market is $v_i - \int_0^c w(t; c^S)dt$. Because c^S satisfies (6), we know that the payoff cannot be negative and so the job with waiting cost $c \leq c_i^S$ will participate in the spot market. This leaves us to proof that jobs with waiting cost $c > c_i^S$ do not participate in the spot market. As mentioned before, the job with the highest waiting cost, ergo highest priority, has no waiting time, thus the expected time in the system is $1/\mu$. Even if a job has waiting cost $t > c_1^S$, the expected time in the system remains $1/\mu$. Therefore, the payment for the job with waiting cost t is at least $m(c_1^S)$ and therefore the job misreports its waiting cost as $\hat{c} \in [0, c_1^S]$ to maximize the payoff. So the expected payoff for this job is

$$v_i - cw(\hat{c}, c^S) - m(\hat{c})$$

after a simple expanding we get

$$\begin{aligned} & v_i - c_i^S w(\hat{c}, c^S) - m(\hat{c}) - (c - c_i^S)w(\hat{c}, c^S) \\ \leq\ & v_i - c_i^S w(c_i^S, c^S) - m(c_i^S) - (c - c_i^S)w(\hat{c}, c^S) \\ =\ & v_i - \int_0^{c_i^S} w(t; c^S)dt - (c - c_i^S)w(\hat{c}, c^S) \\ =\ & -(c - c_i^S)w(\hat{c}, c^S) \\ <\ & 0 \end{aligned}$$

Using the incentive compatibility constraint (3) bounds the payoff as in the first inequality. This leads to a new term with a payoff from the spot market with waiting cost equal to the cutoff, hence 0. Then we are left with $-cw(\hat{c}, c^S) + c_i^S w(\hat{c}, c^S)$ which is negative since $c > c_i^S$. Therefore the payoff for a job with waiting cost $t > c_i^S$ is negative and the job will not participate in the spot market.

□

[2]The proof of Lemma 7 can be found in Appendix A in (Abhishek et al. 2012).

From now on, we use

$$m(c; c^S) = \int_0^{c_i^S} w(t; c^S)dt - cw(c; c^S)$$

since the payment is dependent on the waiting cost c and the cutoffs c^S. Then the revenue for the spot market in isolation is

$$R^s \triangleq \sum_i \lambda_i \int_0^{c_i^S} m(t; c^S) f_i(t)dt. \tag{8}$$

5 Revenue for Hybrid Market

Considering a market with both possibilities, PAYG or spot market (or bulking), we again need to find the cutoff vector c(p). The cutoff for each class is now dependent on the price p in the PAYG market. Then we can say that the job should participate in the spot market for $c < c_i(p)$, in the PAYG market for $c_i(p) < c$. Theorem 2 in (Abhishek et al. 2012) proofs the existence of such cutoff vector and further gives some structure to the possible result which leads in the next corollary.

Corollary 1. *All classes that participate in PAYG have the same cutoff, $c_i^*(p)$, and if the price is set higher than $\mu v_i - c_i^S$ then no class participates in PAYG and the outcome is the same as if only the spot market existed.*

So in other words, a price high enough, can eliminate the PAYG market and lead to an isolated spot market.

This characterization of the equilibrium in the hybrid market can be achieved by implementing the mechanism that higher waiting cost leads to higher priority and collect the payment according to (4). Then the hybrid market generates the following revenue

$$R^H \triangleq \sum_i \lambda_i \left(\frac{p}{\mu}[F_i(\mu v_i - p) - F_i(c_i(p))] + \int_0^{c_i(p)} m(t; c(p)) f_i(t)dt. \right) \tag{9}$$

6 Revenue Comparison

Since the isolated spot market is the same as the hybrid market with a price p high enough so that no job participates in the PAYG market, we can focus on the comparison between isolated PAYG and hybrid market.

In (Abhishek et al. 2012), there are two main experiments. First they generated the parameters like value or number of queues randomly and for each of those random comparisons, the PAYG market outperformed the hybrid market. This is even true for prices

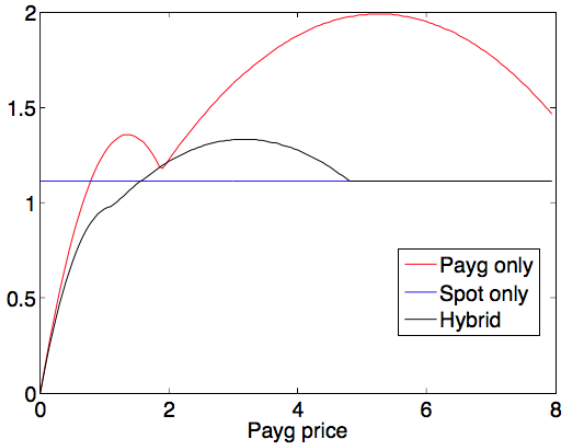

Figure 3: Revenues $R^{PAYG}(p)$, $R^s(p)$ and $R^h(p)$ as a function of PAYG price p.

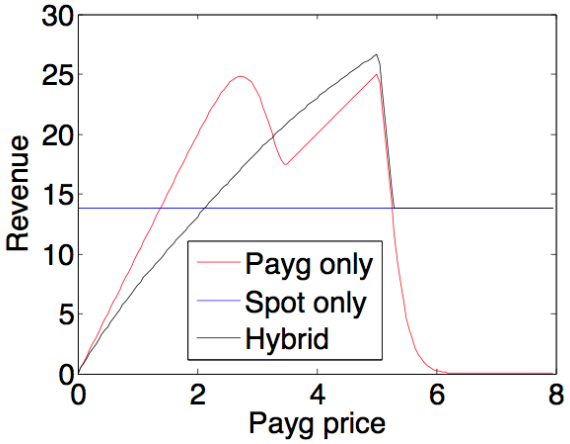

Figure 4: Optimal $R^h(p)$ can outperform $R^{PAYG}(p)$ with $v_1 = 10$ and $v_2 = 3.5$.

$p > \mu v_2 - c_2^S$ as seen in figure 3. A low price p means that every job participates in the PAYG market, for isolated PAYG and hybrid. Increasing p leads to some jobs switching from PAYG to spot market in the hybrid version, hence isolated PAYG outperforms hybrid. This holds until price p reaches the point where some jobs in the isolated PAYG market start bulking. The only point where hybrid outperforms PAYG is the exact point of p, where all jobs from class 2 bulk in the isolated PAYG market. When increasing p from that point on, the revenue from the isolated PAYG increases faster, so that hybrid clearly generates lower revenue for the provider compared to the isolated PAYG market.

In the second experiment it was all about trying to get the hybrid market to outperform the isolated PAYG. The basic idea is to generate two classes, one with high value and high waiting costs and one with low value and low waiting cost, so that in hybrid the jobs with high value participate in PAYG and the low value jobs in spot market while in the isolated PAYG the low value jobs bulk. Unfortunately, in hybrid there are

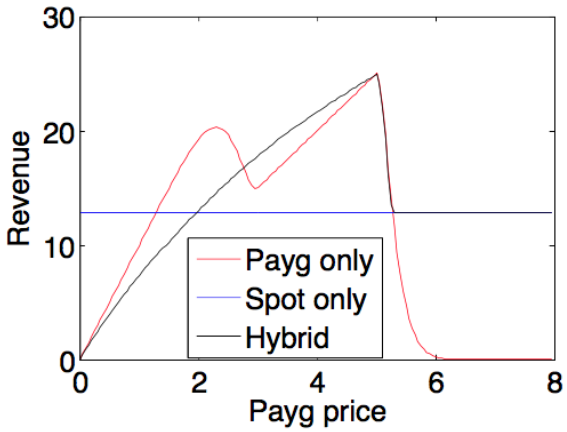

Figure 5: Same as figure 4 but with $v_2 = 3$ and highest optimal revenue changed to $R^{PAYG}(p)$

always some high value jobs participating in spot market and some low value jobs that do not participate. In other words, it is not possible to perfectly separate the two classes by price alone. (Abhishek et al. 2012) nevertheless decided to try to find the optimal parameters to get the best separation in order to strengthen the hybrid market. In figure 4, we see that with the separated waiting cost the hybrid can outperform the isolated PAYG market. But this was an Attempt to get the best out of the hybrid market and it still does not utterly outperform the isolated PAYG market. Furthermore, by changing the valuation from the second low value class just from 3.5 to 3, the hybrid market already performs only for some price p better and has a lower optimal revenue as seen in figure 5.

7 Conclusion

(Abhishek et al. 2012) has shown that regardless of the auction mechanism for the spot market, the isolated PAYG market performs better in regard to revenue. At the beginning we said, it is intuitively not clear whether the advantage of the price discrimination to serve new customers with low value achieves more revenue than the loss of PAYG customer to the spot market. Now, we must say that the loss of revenue when customer choose the spot market instead of the PAYG market is higher than the gain from new customers in the spot market.

One assumption that (Abhishek et al. 2012) made, could potentially change the result in favour of the hybrid market. The assumption that the only cost in the spot market is the waiting cost, leads to the results from above. Let us consider that there is also a cost for preemption or rewriting applications to handle preemptions. That could lead to jobs sticking to the PAYG market longer because of the higher cost in the spot market and with still some jobs that bulk in the isolated PAYG but use the spot market in the hybrid version, the revenue could then potentially increase in hybrid to a point where it

outperforms the isolated PAYG market.

The result from the experiment begs the question why does Amazon provide a spot market. While the answer to that question can only be given by Amazon itself, we can think of a couple of reasons. First of, it is possible that one or more assumptions make this result not applicable in their setting. On the other hand, it may have reasons other than revenue. Amazon could for example value the insights and information gained from a spot market higher than the actual revenue or they are happy to provide a cheaper solution for low scale customer so that they do not use a competitor's service. Which is consistent with the observation that Amazon makes it difficult to use the spot market in a large scale.

References

Abhishek, Vineet, Ian A. Kash, and Peter Key (2012). "Fixed and Market Pricing for Cloud Services". In: *CoRR* abs/1201.5621. URL: `http://arxiv.org/abs/1201.5621`.