Dimitri *Journal of Cloud Computing: Advances, Systems and Applications* (2020) 9:14
https://doi.org/10.1186/s13677-020-00161-2

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

# Pricing cloud IaaS computing services


Nicola Dimitri

**Abstract**

The economics of cloud computing has recently attracted increasing attention. In particular, a topic which is still under debate is how prices charged to customers for cloud resources are formed, since alternative pricing rules could be considered. Based on three pricing schemes inspired by those used by Amazon EC2, the main global cloud service provider, in the paper we address two main issues. First we present a methodology for the relevant parameters of the pricing rules to be determined in an optimal way, that is to maximise the provider's revenue. Moreover, we discuss reasons for co-existence of three pricing rules, rather than fewer, to access the cloud. Our findings suggest that this may be due to a larger coverage of the potential demand, since customers applying for cloud services vary in their willingness to pay for the job, the time length of the service, the computational power requested etc. Furthermore, the pricing rule in the so-called, spot market, can provide the platform with useful information on the customers willingness to pay for cloud services. This is because in the spot market users offer a price for service, but pay less than that if their request is satisfied.

**Keywords:** Cloud computing services, Prices

## Introduction

One of the most remarkable phenomena in information technologies that took place over the last few years is cloud computing [1]. The possibility of profitable use of IT excess capacity by some providers, matched with the need by customers to save on buying IT infrastructures for software services, data storage etc. led to a flourishing market where potential users can successfully satisfy their demand for such requests. Its use is increasingly so wide spread that cloud computing is claimed to be the 5th utility, joining electricity, gas, telephony and water [2].

The main types of services, though not the only ones, typically available in the cloud are: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS), whose definition and features have been widely discussed in the literature [2–10]. In the paper we shall focus on pricing rules in IaaS. However, cloud computing is a continuously evolving market [11, 12] and more recently, to optimise the use of exisiting resources, the additional activity of Function as a Service (FaaS) is also taking place. Such service is developing under the so called *serverless computing*, where specific functions are performed without an explicit reference to infrastructure [13–19].

Because of its business potentials in recent years the economics of cloud computing began to attract major attention, from a number of different perspectives such as the providers' revenues and their market shares, the impact on employement and GDP, the pricing strategies [20–30]. As the range of services available in the cloud increases, a main issue which keeps being debated is the pricing policies adopted by the providers, which fundamentally affect their service profitability.

In this article we discuss the design of such price policies, focusing on pricing schemes that take inspiration from Amazon EC2 provision of IaaS; the reason is twofold. First, Amazon is the main global IaaS provider, with about 15.5*$bln* revenues and 48% share of the world market in 2018, against about 5*$bln* revenues and 15.5% market share by Microsoft [31], its main competitor. Moreover, although in the market there is a range of alternative proposals, its pricing schemes are akin to those used by other providers and the findings may therefore have a general interest.

Broadly speaking, the Amazon EC2 cloud platform adopts three main pricing rules. The first is the so called *pay-as-you-go* (PAYG) rule, where users are charged a monetary amount possibly based on two components: a

Correspondence: dimitri@unisi.it
Department of Political Economy and Statistics, University of Siena, Piazza San Francesco 7, 53100 Siena, Italy

fixed component plus a variable part depending on the length of time the cloud resource is used for.

The second pricing scheme is still based on a unit, hourly, rate however without the certainty that a request for service will be satisfied by the provider. For this reason is called *on demand* (OD), since users in this case decide not to pay a flat rate to make sure their need will be satisfied.

As a third possibility users could make a request for platform resources on the *spot market* (SM) [32, 33], where machine capacity currently idle is made available by the provider to potential customers. The spot market is characterised by the *spot price*, determined by the platform *considering* supply and demand of service [34]. As well as for the OD price scheme, there is a positive probability that the request will not be accepted if the price offered is below the spot price. Moreover, in SM there is also a positive probability that the job will be discontinued once it started. This suggests that both OD and SM may be undesirable for jobs which customers are unwilling to delay.

In the paper we investigate how users may select among the three pricing rules and, based on this, how the provider could determine the relevant parameters value of the price schemes. Our goal is mostly exploratory, as we confine ourselves to introducing a methodology and discuss few examples. The main formal constructs and quantities of the paper could then be estimated from the available data sets.

The paper is structured as follows. In Section 2 we discuss some related work. In Section 3 we introduce the three pricing schemes and start comparing them from the user's perspective. The discussion will represent the basis for the subsequent sections. In Section 4 we consider the provider perspective with both complete and incomplete information on the job value, and time length, and discuss how optimal pricing rules can be defined to take into account such uncertainty. Section 5 compares the provider's revenue generated by PAYG and OD with SM. Yet, depending upon the nature of the uncertainty on the spot price, as well as on considerations related to what we call SM *interruption costs* for the users, such comparison may not be straightforward. Section 6 contains some concluding remarks.

## Related work

Rates charged to cloud computing customers could be based on several elements, alone or in combination. Flat subscription, type of service, usage time, size of computing resources occupied, geographical location of resources, demand-supply interaction, system congestion. Moreover, prices could be kept constant over time or vary dynamically reacting to changing circumstances ([35–41]; Pal-Hui, 2013 [7, 42–59]).

Because of such variety of criteria it is natural to ask which pricing schemes could be more profitable, and under what conditions, for the providers. Moreover, why are providers using different price rules at the same time and how are the relevant parameters of pricing schemes determined. Intuitively such range of possibilities is offered to capture the different preferences and needs of the customers, as well as demand intensity, for which a variety of options may be preferable.

However, so far such questions have only received partial answers. More specifically, Ma and Huang [41] investigate the optimal users' submission strategy under such mixture of available pricing schemes. Huang et al. [60] interpreting the possibility of an interrupted job, as in the Amazon EC2 spot market, as a damaged service, discuss the advantages of hybrid markets, where fixed price and spot market cloud instances coexist. Hoy et al. [61] argue how co-existence of the spot market and reserve prices instances in Amazon EC2 could be due to the users' different risk attitudes.

Yet, no attempt seems to have been made to discuss how providers may proceed to determine such pricing rules in an optimal way. In particular, for a given price scheme how they should define its parameter values. Moreover, we further ask why could it be optimal for a provider to have different pricing rules coexisting at the same time. This is what we address in the rest of the paper where, with no major loss of generality, to keep the exposition simple we shall only consider time length of resource use as a criterion underlying the price charged.

## The three pricing schemes from the user perspective

In this section we introduce the three pricing schemes adopted by Amazon EC2 for IaaS, initially taking the user point of view, which will then form the basis for modelling the provider optimal determination of the pricing rules. To focus on the main message the discussion will concentrate on the essential features of such pricing schemes. Indeed, the model could then be amended to include further elements and additional details.

Consider a user who wishes to run a job, at a particular date in time, with economic value $v > 0$, which needs $\tau = 0, 1, 2, 3. .$ units of machine time in the cloud. More specifically, $v$ is the maximum monetary sum the user is willing to pay to run the job. To simplify the exposition, without losing much generality, in the paper we assume $\tau$ to be known by the user. In reality this may not be the case, since the customer may be uncertain about the time length needed for the job. One way to take such user uncertainty into account would be to introduce a probability distribution over the possible time durations of the job. A further source of uncertainty, which we do

not discuss in the paper, could be whether or not the machine capacity purchased by the user will be enough to run the job. In Amazon the issue could be coped with by choosing types of services with automatic scaling. The reason why $\tau = 0$ is also included will be clear below. Therefore, a job is defined by a pair $j = (v, \tau)$. Formally a job defines the user *type*, about which the provider may have different degrees of information.

The above definition of a job could be enriched in at least two directions. The first may be to add a third component $j = (v, \tau, \sigma)$, where $\sigma$ would stand for the size of occupied machine space. Such more general notion could allow discussing how time and size might, if at all, replace each other to execute a job. A second direction may be to define $v = v(\tau, \sigma)$, that is with the value as a (possibly increasing) function of needed time and space. However, since our main goal is to discuss a methodology, without much loss of generality we keep the notion of a job as simple as $j = (v, \tau)$. Moreover, still for simplicity we shall not deal with multiple sequential jobs, which could be introduced without much altering the main framework.

#### Pay-as-you-go (PAYG) only pricing scheme

According to PAYG customers can reserve the use of machine time by paying a flat rent for a given period $a > 0$, and then pay a time unit (hourly) fee $b > 0$ for use, so that a job with processing time length of $\tau = 0, 1, 2, 3,$ units would have a total cost given by

$$C_p(\tau) = a + b\tau$$

and its average cost per unit of time by

$$\frac{C_p(\tau)}{\tau} = \begin{cases} \frac{a}{\tau} + b \text{ for } \tau = 1, 2, \\ a \text{ for } \tau = 0, \end{cases}$$

Therefore with a flat, constant, rent component even if $\tau = 0$, that is no jobs executed, the customer will have to pay $a$. Yet, once the rent has been paid the longer $\tau$ the lower the cost component imputable to the fixed part. Indeed, for large enough $\tau$ the average cost will tend to coincide with the (marginal) unit cost $b$ of the job. Notice that a decreasing average cost implies increasing returns of scale, with respect to $\tau$, for the user.

Assuming risk neutrality, it follows that a customer payoff related to the job $j = (v, \tau)$ would be

$$\Pi_p(v, \tau) = v - (a + b\tau) \quad (1)$$

which is non negative for $\tau \leq \frac{v-a}{b}$. Therefore, PAYG can be considered by a user only if the job execution time is *short enough*, lower than an increasing linear function of $v$. Intuitively, the more valuable is the job the longer could be its time length to generate a positive profit.

Clearly, jobs such that $a > v$ would never find it profitable to apply for PAYG.

#### On-demand (OD) only pricing scheme

The second criterion bears some similarity with PAYG, as it is still a linear function of the job time length, however based on a user asking to run the job *on demand*, that is without paying a flat rate to reserve machine time. As a consequence, the request for resources has positive probability $1 > 1 - \theta > 0$ to be rejected. For simplicity, without major loss of generality, we assume $0 < \theta < 1$ to be constant though, more realistically, it may depend upon $\tau$. Moreover, $\theta$ could be strategically chosen by the provider, a possibility which we do not analyse in the paper. In this case, the cost for the customer will be a random variable defined as

$$C_d(\tau) = \begin{cases} c\tau \text{ with probability } \theta \\ 0 \text{ with probability } 1-\theta \end{cases}$$

where $c > 0$ is the unit cost, and his expected cost given by

$$EC_d(\tau) = \theta c\tau$$

Hence the customer expected payoff (profit) when operating on demand is

$$E\Pi_d(v, \tau) = \theta(v - c\tau) \quad (2)$$

which is instead non-negative for $\tau \leq \frac{v}{c}$.

The average (expected) cost will now be constant, rather than decreasing in $\tau$, and given by

$$\frac{EC_d(\tau)}{\tau} = \theta c \text{ for } \tau = 1, 2, 3, \ldots$$

#### PAYG and OD pricing schemes

When both PAYG and OD are available, as compared to pay as you go the customer would prefer to run job $j = (v, \tau)$ on demand if his payoff is non-negative and (2) is larger than (1). Namely if

$$c\tau \leq v < \frac{a}{1-\theta} - \frac{(\theta c - b)}{1-\theta}\tau \quad (3)$$

and therefore if $v$ is sufficiently small, though not too small. If $(\theta c - b) > 0$ then the right hand side of (3), as a function of $\tau$, is a negatively sloped line taking positive values over the time domain $0 \leq \tau \leq \frac{a}{(\theta c - b)}$. Alternatively, if

$(\theta c - b) \leq 0$ then the function $\frac{a}{1-\theta} - \frac{(\theta c - b)}{1-\theta}\tau$ is either constant or a positively sloped line in $\tau$ and, for this reason, always positive.

Likewise, if

$$Max\left[a + b\tau, \frac{a}{1-\theta} - \frac{(\theta c - b)}{1-\theta}\tau\right] \leq \nu \quad (4)$$

the user would prefer to submit for PAYG implying that there are jobs $j = (\nu, \tau)$, with low $\nu$ and large $\tau$, which the customer neither wants to run as PAYG nor as OD, since his expected profit could be negative in both cases.

Hence, broadly speaking, (3) suggests that OD is less attractive for jobs with either high value and/or high machine time. Intuitively, this is because customers who need to execute these jobs could not run the risk of having their request rejected when applying on demand.

#### The spot market (SM) only without interruption costs

As a third option the customer may buy machine time on the *spot* market where idle resources are offered by the provider at each date. As well as for PAYG and OD, SM is characterised by a variety of options and rules, notably on how to access the resources and terminate the job. However, for the purpose of this paper what follows is a sufficiently rich description of the main SM features, as originally proposed by Amazon. For this reason, as we shall see below, even with complete information at each date we consider a single, and the same, spot price for all the requesting users.

For a job of time length $\tau > 0$, acquisition of machine time and related payment work as follows. Suppose $s_t \geq 0$, with $t = 0, 1, \ldots, \tau - 1$, is the spot price at time $t$ for consuming a unit of machine time, from $t$ to $t + 1$, where at each date such price is determined by the provider considering demand-supply interaction, and possibly other elements. Then, at $t = 0$ the customer has to bid a unit price $q$ without knowing $s_t$. After having bid, the spot price becomes known to the user and if $q \geq s_t$ the customer obtains access to machine time, however paying $s_t$ rather than the submitted offer $q$, until inequality $q \geq s_t$ remains true for $t = 0, 1, \ldots, \tau - 1$. In reality, even if $q \geq s_t$, access could be denied or delayed if resources are not available when requested. However, with no major loss of generality, to simplify the exposition in the paper we assume that $q \geq s_t$ is a necessary and sufficient condition for accessing cloud resources via SM.

If at some date $t \leq \tau - 1$ it is $q < s_t$ then the job will be terminated and, again for simplicity, we could assume that machine time paid until then is lost and that the job will have to start again. An alternative, more realistic, version instead would be to assume that the job is discontinued, and data stored by the provider-user, until $q \geq s_t$ becomes true again, if ever. For example, suppose the job starts at $t = 0$ and that $q \geq s_t$ until $t = t' < \tau$. Then assume $q < s_t$ at dates $t$ such that $t' < t \leq \tau + t''$, with $t'' > 0$, and subsequently $q \geq s_t$ at times $\tau + t'' < t \leq 2\tau + t''' - t'$. This means that the job would be discontinued for $\tau + t'' - t'$ time units and could not be completed until date $2\tau + t''' - t'$.

This pricing framework is akin to second price Vickrey type-of-mechanisms ([62]; Hurwicz-Reiter, 2006 [63–67];) where *truth telling*, namely bidding one's value, is a weakly dominant strategy. Indeed, as we shall see below, this is because we assume the user to be uninformed (uncertain) about the spot price value and its formation, since he does not know how many requests the platform receives and how many cloud resources are available. Because of such uncertainty we suppose that strategic thinking could not take place, and the decision on which $q$ to offer based on the idea that the spot price $s_t$ is an exogenous random variable. It follows that, for a user, deciding which price to submit *is like* bidding in a second price auction against *nature* choosing $s_t$. In this case, for no unit price $q \neq \frac{\nu}{\tau}$ the user's payoff is strictly higher than when $q = \frac{\nu}{\tau}$, regardless of the level of $s_t$.

Therefore, we assume truthful bidding to be the case and the unit price submitted by the user equal to $q = \frac{\nu}{\tau}$.

However, as we said, when the user submits his own price he does not know what the current spot price is. Thus, since the customer is uncertain about the spot price, we assume he treats $s_t$ as a sequence of independent and identically distributed random variables, with distribution function $F(s)$, density function $f(s) = F'(s)$ and $s \geq 0$.

Suppose the customer pays no cost in case of job interruption. Then when such functions correctly represent the distribution of the spot price, for both the user and the provider, it follows that the customer's payoff $\Pi_s(\nu, \tau)$, for a job of time length $\tau$, when buying machine time on the spot market, is given by

$$\Pi_s(\nu, \tau) = \begin{cases} \nu - \sum_{t=0}^{\tau-1} s_t \text{ if } \frac{\nu}{\tau} \geq s_t \text{ for all } t = 0, 1, .., \tau - 1 \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

and the expected payoff by

$$E\Pi_s(\nu, \tau) = \int_0^{\frac{\nu}{\tau}} \ldots\ldots \int_0^{\frac{\nu}{\tau}} \left(\nu - \sum_{t=0}^{\tau-1} s_t\right) f(s_0) .. f(s_{\tau-1}) ds_0 .. ds_{\tau-1}$$

Then

$$\begin{aligned} E\Pi_s(\nu, \tau) &= \nu F\left(\frac{\nu}{\tau}\right)^{\tau} - \tau \frac{\nu}{\tau} F\left(\frac{\nu}{\tau}\right)^{\tau} \\ &\quad + \tau F\left(\frac{\nu}{\tau}\right)^{\tau-1} \int_0^{\frac{\nu}{\tau}} F(s) ds \\ &= \tau F\left(\frac{\nu}{\tau}\right)^{\tau-1} \int_0^{\frac{\nu}{\tau}} F(s) ds \geq 0 \quad (6) \end{aligned}$$

precisely because the user bids the job value. Notice that the expected cost for the customer will be

$$Ec_s(\nu,\tau)=\tau F\left(\frac{\nu}{\tau}\right)^{\tau-1}\left[\left(\frac{\nu}{\tau}\right)F\left(\frac{\nu}{\tau}\right)-\int_0^{\frac{\nu}{\tau}}F(s)ds\right]$$

which, unlike PAYG and OD pricing rules, as well as depending on $\tau$ depends also on $\nu$, since the price offered to access resources on the cloud via SM is related to the job value.

It is important to point out again that definition (5) disregards interruption costs, in the sense that if the job is discountinued before its completion then (5) assumes that monetary resources invested until then is not completely lost. Perhaps this is because the job will be finalised later, or because money spent until then is considered as sunk, and so not part of the expected payoff computation. Below we shall also discuss how the presence of such costs could affect the main findings.

#### PAYG, OD and SM pricing schemes

Define $E\Pi_{p,\ d,\ s}(\nu,\tau)$ to be a customer expected payoff when the three pricing rules coexist. Then the following statement summarises the previous considerations.

**Proposition 1** *Suppose* $F(0)=0$ *and* $F(s)>0$ *for all* $s>0$; *then, without interruption costs for any given job* $j=(\nu,\tau)$ *it is* $\mathrm{E}\Pi_{p,\ d,\ s}(\nu,\tau)\geq 0$.

*Proof* Immediate. Indeed, though expressions (3) and (4) imply that there could be jobs for which the user has negative payoffs in both PAYG and OD, expression (6) implies that those jobs can obtain non-negative expected profits by relying on SM.

The proposition could be rephrased as the following characterisation of SM

**Corollary 1** *SM, without interruption costs, complements PAYG and OD criteria by guaranteeing that any job* $j=(\nu,\tau)$ *can obtain non-negative expected profits.*

Therefore, with no interruption costs SM may be interpreted as an institutional mechanism attempting to attract those jobs that otherwise would not apply to the platform, if PAYG and OD only were available. Finally (1), (2) and (6) suggest that the user profits are all decreasing in $\tau$, with the probability of a customer being served playing a role in (2) and (6) only.

#### The spot market (SM) only, with interruption costs

The above definition of the user costs in SM does not consider the possibility that if a job is interrupted, before completion, the whole amount paid by the user until then will be lost. In this paragraph we briefly see how interruption costs may affect previous findings.

Consider a generic date $t$, with $0<t\leq\tau-1$, such that $s_i<\frac{\nu}{\tau}$ for $i=0,\ 1,\ \ldots,\ t-1$ and $s_t>\frac{\nu}{\tau}$. Then, until $t-1$ the customer pays $\sum_{t=0}^{t-1}s_i$ however without completing the job. Therefore, in this case the user's expected cost $\mathrm{E}c_{s(i)}(\nu,\tau,t)$ is given by

$$\mathrm{E}c_{s(i)}(\nu,\tau,t)=\left[1-F\left(\frac{\nu}{\tau}\right)\right]tF\left(\frac{\nu}{\tau}\right)^{t-1}\left[\left(\frac{\nu}{\tau}\right)F\left(\frac{\nu}{\tau}\right)-\int_0^{\frac{\nu}{\tau}}F(s)ds\right]$$

however without enjoying any profit since the job would be discontinued before its completion. Full consideration of all such sunk costs would lead to total expected costs as defined by

$$\begin{aligned}\mathrm{E}c_{s(i)}(\nu,\tau)&=\sum_{t=1}^{\tau-1}\mathrm{E}c_{s(i)}(\nu,\tau,t)\\&\quad+\tau F\left(\frac{\nu}{\tau}\right)^{\tau-1}\left[\left(\frac{\nu}{\tau}\right)F\left(\frac{\nu}{\tau}\right)-\int_0^{\frac{\nu}{\tau}}F(s)ds\right]\\&=\frac{\left[1-F\left(\frac{\nu}{\tau}\right)^{\tau}\right]}{\left[1-F\left(\frac{\nu}{\tau}\right)\right]}\left[\left(\frac{\nu}{\tau}\right)F\left(\frac{\nu}{\tau}\right)-\int_0^{\frac{\nu}{\tau}}F(s)ds\right]\end{aligned}$$

Therefore, the customer expected profit is now

$$\mathrm{E}\Pi_{s(i)}(\nu,\tau)=\nu F\left(\frac{\nu}{\tau}\right)^{\tau}-\frac{\left[1-F\left(\frac{\nu}{\tau}\right)^{\tau}\right]}{\left[1-F\left(\frac{\nu}{\tau}\right)\right]}\left[\left(\frac{\nu}{\tau}\right)F\left(\frac{\nu}{\tau}\right)-\int_0^{\frac{\nu}{\tau}}F(s)ds\right]\quad(7)$$

Since $\frac{\left[1-F\left(\frac{\nu}{\tau}\right)^{\tau}\right]}{\left[1-F\left(\frac{\nu}{\tau}\right)\right]}>\tau F\left(\frac{\nu}{\tau}\right)^{\tau-1}$ there is no guarantee that (7) would be positive for all $\nu$. Indeed, with interruption costs we can now state the following

**Proposition 2** *In a SM with interruption costs, for a given* $\tau$ *and some* $\nu$ *it is possible that* $\mathrm{E}\Pi_{s(i)}(\nu,\tau)<0$.

To see this, as an example, suppose the user is completely uncertain about $0\leq s\leq 1$ which, for this reason, he models as a uniform random variable with density $f(s)=1$, and distribution function given by $F(s)=s$. Then, with $\tau=1,\ 2$, from (7) we obtain

$$\begin{aligned}\mathrm{E}\Pi_{s(i)}(\nu,\tau=1)&=\frac{\nu^2}{2}>0\ \textit{for all}\ \nu;\mathrm{E}\Pi_{s(i)}(\nu,\tau=2)\\&=\frac{\nu^2}{4}\left(\frac{3\nu}{4}-\frac{1}{2}\right)>0\ \textit{for}\ \nu>\frac{2}{3}\end{aligned}$$

which illustrates the statement. The example seems to suggest that with interruption costs negative expected profits are more likely to emerge with longer, low value, jobs.

#### Waiting time costs

Before going to the next section it is worth pointing out that costs, for users, could also be due to how long they have to wait before job completion, or to a delay beyond $\tau$. For example, with PAYG the overall cost could be formulated as

$$c_p(\tau) = a + b\tau + w(\tau)$$

where $w(\tau)$, an increasing function of $\tau$, is the user's cost for waiting $\tau$ units of time for job execution. However, explicit consideration of such costs would not basically alter the main findings of the analysis, which is why we decided to keep costs expressions in their simplest form.

## Optimal PAYG and OD pricing rules without SM

Based on the previous section we now proceed gradually, supposing first that the spot market is not available, to discuss the optimal choice of the parameters by the provider, in the PAYG and OD expressions. We do so by comparing the pricing schemes under alternative information on $v$ and $\tau$, assuming first complete information by the platform on these two quantities. Though informationally very demanding for the provider, and for this reason *highly unrealistic*, complete information represents an interesting benchmark in the analysis.

Excluding SM initially will allow us to begin with a simpler framework, as well as to better discuss the implications of the presence of SM on the provider's payoff.

### Complete information on $v$ and $\tau$

If for, each single job, $v$ and $\tau$ were known by the provider then with PAYG his revenue would be

$$R_p(v, \tau) = (a + b\tau)$$

and so the optimal value of the parameters $a$ and $b$ is found simply by solving the following problem

$$\begin{aligned} max_{a,b} &= R_p(v, \tau) = (a + b\tau) \; such \; that \; \Pi_p(v, \tau) \\ &= v - (a + b\tau) \geq 0 \; and \; a, b \geq 0 \end{aligned} \quad (8)$$

from which the optimal $a$ and $b$ satisfy

$$v - (a + b\tau) = 0$$

That is parameter values should be *incentive compatible*, granting the user a non negative payoff. It follows that any pair $a$ and $b$ satisfying

$$a = v - b\tau$$

is optimal. Therefore, because of complete information the provider would extract full rent from the customer obtaining as revenue the whole value $v$. Hence, in this case his revenue would not depend upon the job time length but on the job value only.

As a consequence, in this extreme case, the platform would adjust the parameters to different customers depending upon $v$ and $\tau$. For example, consider the job $j = (v = 10;\ \tau = 5)$; then any combination of parameters $a$, $b \geq 0$ satisfying

$$a = 10 - 5b$$

is optimal for the job. For this reason, although each single job would induce such combination, there could be parameters that might be optimal for more than one job.

Indeed, consider job $j' = (v = 10;\ \tau = 3)$; then parameters $a = 10$ and $b = 0$ are optimal for both $j$ and $j'$.

With OD the provider's expected revenue would be

$$ER_d(v, \tau) = \theta c \tau$$

and, in analogy with (8), maximised by choosing $c$ to satisfy

$$E\Pi_d(\tau) = \theta(v - c\tau) = 0$$

hence $c = \frac{v}{\tau}$ leading to $ER_d(v, \tau) = \theta v < v = R_p(v, \tau)$ which implies that PAYG would be *ex-ante*, before the user submits a request, preferable for the provider. However *ex-post*, *conditionally* upon the customer having been served, both PAYG and OD would generate the same revenue $v$.

### Incomplete information on $v$ and complete information on $\tau$

Suppose now the provider knows $\tau$ but he's uncertain about $v$, a rather more realistic assumption since users would typically hide their willigness to pay to the provider. Such uncertainty can be modelled assuming $v$ to be a random variable for the platform, with distribution function $G(v)$, density function $G'(v) = g(v)$ and $v \geq 0$. For this reason, in this section we shall determine the parameter levels as depending on $\tau$ only, and not on $v$.

In so doing we are not explicitly considering *incentive compatibility* constraints [68, 69] for each $v$. This is because our main goal is to discuss a methodology and, for the purpose of this paper, the chosen approach would allow us to keep the analysis at a tractable level without losing much content, along lines akin to those recently advocated by [70]and Hartline [71]

From (3) and (4), for given $v$ and $\tau$, it follows that the customer would prefer PAYG to OD when $v$ is large enough. Therefore, if $c > b$ the provider's expected revenue $ER_{p,\,d}(\tau)$ will be

$$ER_{p,d}(\tau) = \begin{cases} \int_{c\tau}^{\frac{a-\tau(\theta c-b)}{1-\theta}} c\tau g(v)dv + \int_{\frac{a-\tau(\theta c-b)}{1-\theta}}^{\infty} (a + b\tau)g(v)dv \; if \tau \leq \dfrac{a}{(c-b)} \\ \int_{(a+b\tau)}^{\infty} (a + b\tau)g(v)dv \;\; if \;\; \tau > \dfrac{a}{(c-b)} \end{cases}$$

that is

$$ER_{p,d}(\tau) = \begin{cases} c\tau\left[G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right) - G(c\tau)\right] + (a + b\tau)\left[1 - G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)\right] if \; \tau \leq \frac{a}{(c-b)} \quad (9a) \\ (a + b\tau)[1 - G(a + b\tau)] \; if \; \tau > \frac{a}{(c-b)} \quad (9b) \end{cases}$$

Given $\tau$, the provider must decide whether the parameters $a$, $b$, $c$ should be fixed in such a way that $\tau \leq \frac{a}{(c-b)}$

or $\tau > \frac{a}{(c-b)}$. Below we discuss conditions under which the former inequality is preferable for the platform

To see this first differentiate (9b), with respect to $a + b\tau$, to obtain the following first order condition

$$a + b\tau = \frac{1-G(a+b\tau)}{g(a+b\tau)} = \frac{1}{\gamma(a+b\tau)} \quad (10)$$

where $\gamma(v) = \frac{g(v)}{1-G(v)}$ is the *hazard rate* of the random variable $v$. Then the following result holds

**Proposition 3** *Suppose (9b) is maximised by (10). Then, for any given τ, it is optimal for the provider to choose parameters a, b, c in such a way that* $\tau \leq \frac{a}{(c-b)}$.

*Proof* Consider (9a). Since $\tau \leq \frac{a}{(c-b)}$, hence $c\tau \leq a + b\tau$, it follows that

$$c\tau\left[G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)-G(c\tau)\right] + (a+b\tau) \times \left[1-G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)\right] \geq$$

$$\geq c\tau\left[G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)-G(c\tau)\right] + c\tau\left[1-G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)\right] \quad (11)$$

and so that

$$c\tau\left[G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)-G(c\tau)\right] + (a+b\tau) \times \left[1-G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)\right] \geq c\tau(1-G(c\tau)) \quad (12)$$

In analogy with (9b), this implies that the right hand side of the above inequality $c\tau(1 - G(c\tau))$ is maximised by $c\tau$ solving the first order condition

$$c\tau = \frac{1-G(c\tau)}{g(c\tau)} = \frac{1}{\gamma(c\tau)} \quad (13)$$

Therefore, since (10) maximises (9b), by setting $c\tau$ as in (13), and $a$, $b$ such that $c\tau \leq a + b\tau$, the provider can obtain at least as much as the maximum level of (9b), which proves the result.

As an example, consider again $v$ to be uniformly distributed on the unit interval. It follows that (9a) and (9b) become

$$ER_{p,d}(\tau) = \begin{cases} c\tau\left[\frac{a-\tau(\theta c-b)}{(1-\theta)}-c\tau\right] + (a+b\tau)\left[1-\frac{a-\tau(\theta c-b)}{(1-\theta)}\right] \text{ if } \tau \leq \frac{a}{(c-b)} & (14a) \\ (a+b\tau)[1-(a+b\tau)] \text{ if } \tau > \frac{a}{(c-b)} & (14b) \end{cases}$$

Consider first (14b) and notice that it is maximised by

$$a + b\tau = \frac{1}{2} \quad (15)$$

providing an expected revenue for the platform equal to $\frac{1}{4}$. That is, for any $\tau$, if $a$, $b$, $c$ are set in such a way that equality (15) and inequality $\tau > \frac{a}{(c-b)}$ are satisfied, then the provider obtains as expected revenue $\frac{1}{4}$.

Inequality (12) in this case becomes

$$c\tau\left[\frac{a-\tau(\theta c-b)}{(1-\theta)}-c\tau\right] + (a+b\tau) \times \left[1-\frac{a-\tau(\theta c-b)}{(1-\theta)}\right] \geq c\tau(1-c\tau) \quad (16)$$

Since the right-hand side of (16) takes as maximum value $\frac{1}{4}$, at $c\tau = \frac{1}{2}$, inequality (16) implies that for any $\tau$, fixing $c = \frac{1}{2\tau}$, $a$ and $b$ such that $\frac{1}{2} = c\tau \leq a + b\tau$, it is always possible for the provider to obtain an expected revenue at least as large as $\frac{1}{4}$.

To find the optimal (expected revenue maximising) values of $a$, $b$ and $c$ we partially differentiate (14a) with respect to to $a+b\tau$ and $c$, under the constraint $b < c \leq \frac{\frac{a}{\tau}+b}{\theta}$.

Now, for given $a$, $b$ and $\tau$, the first order condition related to (14a) with respect to $c$ gives

$$\left[\frac{(a+b\tau)-\tau\theta c}{(1-\theta)}-c\tau\right]-\frac{c\tau}{(1-\theta)}+(a+b\tau)\frac{\theta}{(1-\theta)} = 0 \quad (17)$$

which solved entails

$$c\tau = \frac{(a+b\tau)(1+\theta)}{2} \quad (18)$$

Differentiating (14a) with respect to $a + b\tau$, and taking the related first order condition, we obtain

$$\frac{c\tau}{(1-\theta)} + \left[\frac{1-\theta-(a+b\tau)+\tau\theta c}{(1-\theta)}\right]-\frac{(a+b\tau)}{(1-\theta)} = 0 \quad (19)$$

leading to

$$(a+b\tau) = \frac{c\tau(1+\theta)}{2} + \frac{(1-\theta)}{2} \quad (20)$$

Replacing (18) into (20), and solving for $(a + b\tau)$ and $c\tau$, provides

$$(a+b\tau) = \frac{2}{(3+\theta)}; c\tau = \frac{(1+\theta)}{(3+\theta)} \quad (21)$$

It is immediate to check that the second order condition for a maximum of (14a) is satisfied and so (21) is a solution of the problem.

Finally, notice that $(a + b\tau)$ is decreasing in $\theta$ while $c\tau$ is increasing, and that both tend to $\frac{1}{2}$ as $\theta$ tends to 1. Moreover, inserting (21) into (14a) provides an expected revenue for the platform equal to $ER_{p,d}(\tau) = \frac{1}{3+\theta} > \frac{1}{4}$, for all $\theta < 1$.

The example suggests that, for any $\tau$, it would be optimal for the provider to set the parameter values in a way that such parameters should depend upon $\tau$, hence *job specific*. For instance, consider $c$; if $\tau = 10$ and $\theta = \frac{1}{2}$ then $c = \frac{3}{70}$ while if $\tau = 20$ then $c = \frac{3}{140}$. Finally notice that, in this case, it is also best for the platform to exclude from PAYG and OD jobs with values $v < \frac{1}{3}$.

To conclude, if instead is $b \geq c$ then the provider expected revenue is

$$ER_{p,d}(\tau) = \int_{c\tau}^{\frac{a-\tau(\theta c-b)}{1-\theta}} c\tau g(v)dv + \int_{\frac{a-\tau(\theta c-b)}{1-\theta}}^{\infty} (a+b\tau)g(v)dv$$

regardless of the value of $\tau$.

#### Incomplete information on $v$ and $\tau$

In case the provider would neither know $v$ nor the time length $\tau$, that a submitted job will take, then $\tau$ too could be treated as a random variable. With no major loss of generality, to simplify the exposition, we assume $\tau$ to be a continuous random variable, independent of $v$, with distribution function $H(\tau)$ and density function $H'(\tau) = h(\tau)$.

Hence, if $c > b$ the provider expected revenue $ER_{p,\ d}$ could now be expressed as the sum of two components

$$ER_{p,d} = ER_{(1)p,d} + ER_{(2)p,d} \ (22)$$

where

$$ER_{(1)p,d} = \int_0^{\frac{a}{c-b}} c\tau\left[G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right) - G(c\tau)\right]h(\tau)d\tau + \int_0^{\frac{a}{c-b}} (a+b\tau)\left[1-G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)\right]h(\tau)d\tau$$

and

$$ER_{(2)p,d} = \int_{\frac{a}{c-b}}^{\infty} (a+b\tau)[1-G(a+b\tau)]h(\tau)d\tau$$

while if $b \geq c$ then

$$ER_{p,d} = \int_0^{\infty} c\tau\left[G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right) - G(c\tau)\right]h(\tau)d\tau + \int_0^{\infty} (a+b\tau) \times \left[1-G\left(\frac{a-\tau(\theta c-b)}{1-\theta}\right)\right]h(\tau)d\tau \ (23)$$

Parameters $a$, $b$ and $c$ maximising (22) and (23) now will no longer be job specific, that is will be the same for all pairs $j = (v, \tau)$. This is what in fact Amazon EC2 proposes to customers since parameters, in its pricing schemes, do not change with $v$ nor with $\tau$. Obtaining a closed form solution for the optimal values of $a$, $b$, $c$ may be cumbersome even for manageable density functions $h(\tau)$, such as the exponential. Yet, in general, they would depend on $\theta$.

## Optimal PAYG and OD pricing rules with SM

In this chapter we add SM to PAYG and OD, to see how it may affect the provider's expected revenue, when the three pricing systems are all in place.

Under complete information and no interruption costs it is straightforward to compare the customer profit under PAYG or OD with profit $\tau F(\frac{v}{\tau})^{\tau-1}\int_0^{\frac{v}{\tau}} F(s)ds$, obtainable in SM. Indeed, from Section 4.1 it follows that with PAYG and OD the provider will extract the maximum willingness to pay of the user who would obtain zero profit. For this reason, customers should prefer executing the job on the spot market since their expected profit will always be positive. Therefore, the presence of SM in this case would typically lower the provider's payoff, with respect to when PAYG and OD only are available.

However, if interruption costs are taken into account the customer choice may depend upon $v$, since in SM his expected profit may be negative.

With incomplete information on $v$, and complete information on $\tau$ the analysis becomes more articulate than without SM.

#### SM with no interruption costs

Before stating a main finding of this section consider again the user's expected payoff given by

$$E\Pi_s(v,\tau) = vF\left(\frac{v}{\tau}\right)^{\tau} - \tau F\left(\frac{v}{\tau}\right)^{\tau-1}\left[\left(\frac{v}{\tau}\right)F\left(\frac{v}{\tau}\right) - \int_0^{\frac{v}{\tau}} F(s)ds\right] \ (24)$$

where $vF(\frac{v}{\tau})^{\tau}$ and $\tau F(\frac{v}{\tau})^{\tau-1}[(\frac{v}{\tau})F(\frac{v}{\tau}) - \int_0^{\frac{v}{\tau}} F(s)ds]$ are, respectively, the user's expected revenue and expected cost.

In what follows it is useful to reformulate (24) adding and subtracting $v$ to obtain

$$E\Pi_s(v,\tau) = v - v\left[1-F\left(\frac{v}{\tau}\right)^{\tau}\right] - \tau F\left(\frac{v}{\tau}\right)^{\tau-1}\left[\left(\frac{v}{\tau}\right)F\left(\frac{v}{\tau}\right) - \int_0^{\frac{v}{\tau}} F(s)ds\right] \ (25)$$

Define

$$Ec_{as}(v,\tau) = v\left[1-F\left(\frac{v}{\tau}\right)^{\tau}\right] + \tau F\left(\frac{v}{\tau}\right)^{\tau-1}\left[\left(\frac{v}{\tau}\right)F\left(\frac{v}{\tau}\right) - \int_0^{\frac{v}{\tau}} F(s)ds\right] \ (26)$$

the user's *adjusted* expected cost, because the factor $v[1 - F(\frac{v}{\tau})^{\tau}]$ is added to the expected cost to keep the expected payoff unaltered when the expected revenue is $v$ rather than $vF(\frac{v}{\tau})^{\tau}$. Below, expressions (25) and (26) will be used to facilitate comparison with PAYG and OD.

When $v$ is uncertain, expression (26) becomes

$$Ec_{as}(\tau) = \int_0^{\infty} \left( v\left(1 - F\left(\frac{v}{\tau}\right)^{\tau}\right) + \tau F\left(\frac{v}{\tau}\right)^{\tau-1} \left[ \left(\frac{v}{\tau}\right) F\left(\frac{v}{\tau}\right) - \int_0^{\frac{v}{\tau}} F(s)ds \right] \right) g(v)dv$$

However, we assume that the provider's expected revenue $ER_s(v, \tau)$ in SM coincides with the user's expected cost and is given by

$$\begin{aligned} ER_s(v,\tau) &= \int_0^{\frac{v}{\tau}} ..... \int_0^{\frac{v}{\tau}} \left( \sum_{t=0}^{\tau-1} s_t \right) f(s_0)..f(s_\tau)ds_0..ds_{\tau-1} \\ &= \tau F\left(\frac{v}{\tau}\right)^{\tau-1} \left[ \left(\frac{v}{\tau}\right) F\left(\frac{v}{\tau}\right) - \int_0^{\frac{v}{\tau}} F(s)ds \right] \leq Ec_{as}(v,\tau) \end{aligned}$$

which when $v$ is uncertain becomes

$$ER_s(\tau) = \int_0^{\infty} \tau F\left(\frac{v}{\tau}\right)^{\tau-1} \left[ \left(\frac{v}{\tau}\right) F\left(\frac{v}{\tau}\right) - \int_0^{\frac{v}{\tau}} F(s)ds \right] g(v)dv \leq Ec_{as}(\tau)$$

Likewise, consider $E\Pi_d(v, \tau) = \theta(v - c\tau)$ to which again we add and subtract $v$ to obtain

$$E\Pi_d(v,\tau) = v - (1-\theta)v - \theta c\tau \quad (27)$$

Then, analogously, define the adjusted expected cost in this case as

$$EC_{ad}(v,\tau) = (1-\theta)v + \theta c\tau \quad (28)$$

The introduction of the adjusted expected costs allows us a direct comparison of (1), (25) and (27) since, for each $v$, all them are expressed as $v$ minus a quantity

Based on the above considerations we can now formulate the following result.

**Proposition 4** *Take PAYG, OD and SM, with no interruption costs*; *then* $ER_{p,\ d,\ s}(\tau) \leq Ec_{as}(\tau)$.

*Proof* For any given $v$, the user now has three options, with SM always inducing positive profits. Therefore, his payoff would be

$$max\left[ (v - (a + b\tau)), \theta(v - c\tau), E\Pi_s(v,\tau) = \tau F\left(\frac{v}{\tau}\right)^{\tau-1} \int_0^{\frac{v}{\tau}} F(s)ds \geq 0 \right] \geq E\Pi_s(v,\tau)$$

Hence, the provider's revenue is given by

$$ER_{p,d,s}(v,\tau) = \begin{cases} ER_s(v,\tau) \text{ if } Ec_{as}(v,\tau) \leq \ min((a+b\tau), (1-\theta)v + \theta c\tau) \\ min((a+b\tau), c\tau) \text{ if } Ec_{as}(v,\tau) > \ min((a+b\tau), (1-\theta)v + \theta c\tau) \end{cases}$$

and taking the expectation of $ER_{p,\ d,\ s}(v, \tau)$, with respect to $v$, we obtain the provider's expected revenue as given by

$$ER_{p,d,s}(\tau) = \int_0^{\infty} ER_{p,d,s}(v,\tau) g(v)dv \leq Ec_{as}(\tau)$$

which proves the proposition.

Because SM introduces an upper bound on the platform revenues, without SM the provider may obtain a higher payoff. As a consequence, under the basic assumptions adopted, the model without interruption costs conveys a somewhat counterintutive conclusion. Indeed, adding SM to PAYG and OD may lower the provider's revenue. However, if interruption costs are taken into consideration things may change.

### SM with interruption costs

To see that with interruption costs Proposition 3 may fail to hold consider again the example following Proposition 2, with $E\Pi_{s(i)}(v, \tau = 3) = \frac{5v^4}{162} - \frac{v^2}{18} - \frac{v^3}{54} < 0$, which would make SM unattractive for customers.

To summarise, the model suggests that coexistence of PAYG, OD and SM could be justified by bidders incorporating interruptions costs in their expected payoff computations, which may favour the use of PAYG or OD rather than of SM

## Discussion and conclusions

In the paper we compared three main pricing schemes for cloud services, taking inspiration from those adopted by Amazon for its EC2 cloud platform, providing IaaS services. Though related to Amazon, the analysis can be of more general interest, for analogous pricing rules used by other cloud services providers. We compared PAYG, OD and SM pricing rules and discussed how the relevant parameters could be optimally determined by the provider. We analysed when it is preferable for the provider to propose both the PAYG and OD systems. We then discussed how coexistence of the three pricing rules could be explained by assuming that some bidders take into explicit consideration the costs to be paid when in SM a job is not completed, and the monetary resources invested until then lost by the customer. Indeed, without considering interruption costs the model suggests that adding SM to PAYG and OD may generate lower expected revenues to the provider, which would make it difficult to justify the presence of SM.

Moreover SM can also be important for the provider to try evincing the distribution of job values in the market. Indeed, since in the model SM may induce truthful bidding by the customers, that is offers based on their true job value, price offers received by the platfom in the spot market could be useful to estimate and reconstruct the value distribution of jobs executed in the cloud. Such estimate could then be used to appropriately calibrate the parameters in PAYG and OD. To summarise, in our view the main contribution of the paper, for the platforms providing IaaS, is the proposed methods on how to determine the revenue maximising parameter values of the pricing rules. Despite the potential difficulties in estimating some of the relevant elements of the analysis, such as the distribution functions $F$, $G$ and $H$, we believe the model may provide useful insights to cloud computing operators when deciding how to price resources.

To better focus on the methodology, in the paper we did not discuss in detail the role of possible resource

constraints for the provider. Yet, some insights are already embodied in the model. Indeed, the constant term in the PAYG scheme allows users to hedge against the risk that machine time may be rejected when requested. If resources were surely available, at any time, such term may not be justified.

The framework we considered for price setting is static, but it provides suggestions on its dynamic extension. Indeed, if pricing rules parameters reflect the characteristics of users' demand for running jobs they would change over time according to how the jobs nature and distribution, in the population of users, would evolve.

We conclude with some considerations on energy consumption and cloud computing. This is an issue which is currently raising a major concern [72, 73] because of the increasing energy use that ICT, hence cloud computing, are making. In our analyses we did not explicitly considered energy costs which, if included, would define the provider objective function as revenues minus such, and possibly others, costs. Likely, energy costs would be increasing in the execution time of a job and machine resources occupied.

**Acknowledgements**
I would like to thank the Editor and two anonymus referees for constructing comments. I'm also indebted to Ian Kash and Peter Key for having brought the issue of pricing in cloud computing services to my attention. Moreover, I wish to thank participants to seminars in Liverpool and Mannheim for their observations. Finally I'm grateful to Corvers Services for having supported an earlier version of this paper.

**Authors contribution**
I'm the sole author. The author read and approved the final manuscript.

**Funding**
No research funding has been received for this publication

**Availability of data and materials**
"Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

**Competing interests**
No competing interest to declare

Received: 15 June 2019 Accepted: 11 February 2020
Published online: 03 March 2020

**References**

1. Varghese B (2019) A history of the cloud. ITNOW 61:46–48
2. Buyya E, Yeo C, Venugopal S, Broberg J, Brandic I (2009) Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Generation Comput Syst 25:599–616
3. Ambrust M, Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M (2009) Above the clouds: a Berkeley view of cloud computing, *Technical report n° UCB/EECS-2009-28*. University of California Berkely
4. Botta A, De Donato W, Persico V, Pescapé A (2016) Integration of cloud computing and internet of things: a survey. Futur Gener Comput Syst 56: 684–700
5. Garg S, Versteeg S, Buyya R (2013) A framework for ranking of cloud computing services. Future Generation Computer Systems 29:1012–1023
6. Gorelik E (2013) Cloud computing model, *CISL Working Paper n°1*. Sloan School, MIT
7. Hsu P-F, Ray S, Li-Hsieh Y-Y (2014a) Examining cloud computing adoption intention, pricing mechanism, and deployment model. Int J Inf Manag 34: 474–488
8. Mell P, Grance T (2011) The NIST definition of cloud computing. Natl Inst Stand Technol, US Dept Commerce, Spec Publication:800–145
9. Yoo C (2011) Cloud computing: architectural and policy implications. Rev Ind Organ 38:405–431
10. Zhang Q, Cheng L, Boutaba R (2010) Cloud computing: state of the art and research challenges. J Internet Serv Appl 1:7–8
11. Buyya R, Narayana SS, Casale G, Calheiros R, Simmhan Y, Varghese B, Gelenbe E, Javadi B, Vaquero LM, Netto M, Toosi AN, Rodriguez MA, Llorente I, De Capitani di Vimercati S, Samarati P, Milojicic D, Varela C, Bahsoon R, Dias De Asuncao M, Rana O, Zhou W, Jin H, Gentzsch W, Zomaya A, Shen H (2018) A manifesto for future generation cloud computing: research directions for the next decade, *ACM Computing Surveys*, 51. Art 105:1–38
12. Varghese B, Buyya R (2018) Next generation cloud computing: new trends and research. Futur Gener Comput Syst 79:849–861
13. Fox G., Ishakian V., Muthusamy V., Vinod A., (2017) Report from workshop and panel on the status of serverless computing and function-as-a-service (FaaS) in industry and research, *First International Workshop on Serverless Computing* (WOSC) 2017
14. Hellerstein J., Faleiro J., Gonzalez J., Schleir-Smith J., Sreekanti V., Tumanov A., Wu C., (2018) Serverless computing: one step forward, two steps back, arXiv: 03651v1 [cs. DC]
15. Lloyd W., Ramesh S., Chinthalapati S., Ly L., Pallickara S., (2018) Serverless computing: an investigation of factors influencing microservice performance, *2018 IEEE International Conference on Cloud Engineering*
16. Lynn T., Rosati P., Lejune A., Emehakaroha V., (2017), A preliminary review of enterprise serverless cloud computing (function-as-a-service) platforms, *2017 IEEE 9thInternational Conference on Cloud Computing Technology and Science*
17. McGrath G., Brenner P., (2017) Serverless computing: design, implementation, performance, *2017 IEEE 37thInternational Conference on Distributed Computing Systems Workshops*
18. Savage N., (2018) Going serverless, *Communications of the ACM*,81-15-16
19. Spillner J., (2019) Quantitative analysis of cloud function evolution in the AWS serverless application repository, arXiv:1905.04800 [cs. DC]
20. Baryak E, Conley J, Wilkie S (2011) The economics of cloud computing, *Working paper n° 11, W18*. Department of Economics, Vanderbilt University
21. Boja C, Pocatilu P, Toma C (2013) The economics of cloud computing on educational services. Procedia- Soc Behav Sci 93:1050–1054
22. Etro F (2009) The economic impact of cloud computing on business creation, employment and output in Europe. Review of Business and Economics 54:179–208
23. Fershtman C, Gandal N (2012) Migration of the cloud ecosystem: ushering a new generation of platform competition. Digiword Economic Journal 85: 109–123
24. Harms R, Yamartino M (2010) The economics of the cloud. Technical report, Microsoft
25. Keskin T, Taskin N (2015) Strategic pricing of horinzotally differentiated services with switching costs: a pricing model for cloud computing. Int J Electron Commer 19:34–53
26. Liebenau J., Karrberg P., Grous A., Castro D., (2012) Modelling the cloud, *Report LSE Enterprise*
27. Marston S, Li Z, Bandyopadhyay S, Zhang J, Ghalsasi A (2011) Cloud computing- the business perspective. Decis Support Syst 51:176–189
28. Pal R, Hui P (2013) Economic models for cloud service markets: pricing and capacity planning. Theor Comput Sci 496:113–124
29. Tak B., Urgaonkar B., Sivasubramaniam A., (2011) To move or not to move: the economics of cloud computing, *Computer Systems Laboratory, Pennsylvania State University Technical Report CSE-11-002*
30. Zissis D, Lekkas D (2012) Addressing cloud computing security issues. Futur Gener Comput Syst 28:583–592
31. Gartner (2019) Market share analysis. IaaS and IUS, worldwide 2018
32. Keller R, Hafner L, Sachs T, Fridgen G (2020) Scheduling flexible demand in cloud computing spot markets. Bus Inf Syst Eng 62:25–39
33. Portella G., Rodrigues G., Nakano E., Melo A.,(2018) Statistical analysis of amazon EC2 cloud pricing models, Concurrency Comput Pract Exp, https://doi.org/https://doi.org/10.1002/cpe.4451
34. Mishra A, Yadav D (2017) Analysis and prediction of Amazon E2C spot instance prices. Int J Appl Eng Res 12:11205–11212

35. Al-Roomi M, Al-Ebrahim S, Buqrais S, Ahmad I (2013) Cloud computing pricing models: a survey. Int J Grid Distributed Comput 6:93–106
36. Burgess M, Wiedenbeck B (2010) Strategic bidding on Amazon E2C. Unpublished manuscript University of Michigan
37. Durkee D (2010) Why cloud computing will never be free. Commun ACM 53:62–69
38. Javadi B, Thulariasm R, Buyya R (2013) Characterizing spot price dynamics in public cloud markets. Future Generations Computer Systems 29:988–999
39. Kamra V, Sonawanc K, Alappanavar P (2012) Cloud computing and its pricing schemes. International Journal of Computer Science and Engineering 4:577–581
40. Li C-F (2011) Cloud computing system management under flat rate pricing. J Netw Syst Manag 19:305–318
41. Ma D., Huang J., (2012) The pricing model of cloud computing services, *2012 International Conference of Electronic Commerce*
42. Abhishek V., Kash I., Key P., (2017) Fixed and market pricing for cloud services, arxiv 1201.5621v2 [cs, GT]
43. Aldossary M, Diemame K, Alzamil I, Kostopoulos A, Dimakis A, Agiatzidou E (2019) Energy-aware cost prediction and pricing of virtual machines in cloud computing environments. Future Generations Computer Systems 93: 442–459
44. Chen S, Lee H, Moinzadeh K (2019) Pricing schemes in cloud computing: utilization-based vs reservation-based. Prod Oper Manag 28:82–102
45. Chun S-H, Choi B-S (2014) Service models and pricing schemes for cloud computing. Clust Comput 17:529–535
46. Gohad A., Narendra N., Ramachandran P., (2013) Cloud pricing models: a survey and position paper. *2013 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*
47. Jain T, Hazra J (2019) On demand pricing and capacity management in cloud computing. Journal of Revenue and Pricing Management 18:228–246
48. Javed B, Bloodsworth P, Rasool R, Munir K, Rana O (2016) Cloud market maker: an automated dynamic pricing marketplace. Future Generations Computer Systems 54:52–67
49. Kash I, Key P (2016) Pricing the cloud. IEEE Internet Comput 20:36–43
50. Lancon J., Kunwar Y., Stroud D., McGee M., Slater R., (2019) AWS EC2 instance spot price forecasting using LSTM netwowrks, *SMU Data Science Review*, 2, article 108
51. Lee I (2019) Pricing schemes and profit-maximizing pricing for cloud service. J Revenue Pricing Manage 18:112–122
52. Li W., Svard P., Tordsson J., Elmroth E., (2013) Cost-optimal cloud service under dynamic pricing schemes, *2013 IEEE/ACM International Conference on Utility and Cloud Computing*
53. Mazrekaj A, Shabani I, Sejdiu B (2016) Pricing schemes in cloud computing: an overview. Int J Adv Comput Sci Appl 7:80–86
54. Wu C, Buyya R, Ramamohanarao K (2019) Could pricing models: taxonomy, survey and interdisciplinary challenges, *ACM Computing Surveys (CSUR)*,102. Article 108
55. Xu H, Li B (2013) A study of pricing for cloud resources. Perform Eval Rev 40:3–12
56. Atar R, Cidon I, Shifrin M (2014) MDP based optimal pricing for a cloud computing queueing model. Perform Eval 78:1–6
57. Block G., Bachrach Y., Key P., (2014) The shared assignment game and applications to pricing in Cloud computing, *in Proceedings of the 13th AAMAS conference, May 5–9 2014 Paris*, (by Lomuscio-Scerri-Bazzan-Huns eds)
58. Hsu P-F, Ray S, Li-Hsieh Y-Y (2014b) Examining cloud computing adoption intention, pricing mechanism, and development model. Int J Inf Manag 34: 474–488
59. Wu C, Buyya R, Ramamohanarao K (2019b) Value-based could price modeling for segmented business to business market. Future Generations Comput Syst 101:501–523
60. Huang J, Kauffman R, Ma D (2015) Pricing strategy for cloud computing: a damaged services perspective. Decis Support Syst 78:80–92
61. Hoy D., Immorlica N., Lucier B., (2016) On demand or spot? Selling the cloud to risk averse customers, *arXiv, 1612.04367v1 [cs, GT]*
62. Vickery W (1961) Counterspecultation, auctions and competitive selaed tenders. J Finance 16:8–37
63. Borgers T., (2015) *An introduction to mechanism design*, Oxford University Press
64. Krishna V (2010) Auction theory (2nd ed). Academic Press
65. Roughgarden T (2016) Algorithmic game theory. Cambridge University Press
66. Shoham Y, Leyton-Brown K (2009) Multiagent systems. Cambridge University Press
67. Vohra R (2011) Mechanism design. Cambridge University Press
68. Bolton P, Dewatripont M (2005) Contract theory. MIT Press
69. Hurwicz L, Reiter S (2008) Designing economic mechanisms. Cambridge University Press
70. Hartline J, Lucier B (2015) Non-optimal mechanism design. Am Econ Rev 105:3102–3124
71. Hartline J., (2017) Mechanism design and approximation, Manuscript http://jasonhartline.com/MDnA/
72. Awada U, Keqiu L, Yanming S (2014) Energy consumption in cloud computing data centers. Int J Cloud Comput Serv Sci 3:145–162
73. Kuribayashi S (2012) Reducing Total power consumption method in cloud computing environments. Int J Comput Netw Commun 4:84


## Publisher's Note