Contributions to Game Theory and Management, XII, 325–341

Competitive Pricing for Cloud Information Resources

Pavel Zakharov

St. Petersburg State University,
7/9, Universitetskaya nab., St. Petersburg, 199034, Russia
E-mail: pavel.zakharov96@gmail.com

Abstract This article explores a pricing model for cloud resources, based on use of two different payment schemes - reservation and pay-as-you-go, each of which is controlled by its administrator. The process of prices determination has a form of a two-stage game. At the first stage, administrators set prices for their cloud resources, trying to maximize their revenue. At this stage, a static non-cooperative two-person game is solved, where administrators act as players; their strategies are the prices for resources; their utilities depend both on prices and on the number of resources sold. At the second stage, with prices values given, customers choose a scheme of payment. Making a choice they seek to minimize their expected costs, which consist of the financial component and the waiting costs. First Wardrop principle is used in order to describe user behaviour and optimality conditions in the second stage of the game. The analysis of the solutions obtained shows the economic efficiency of an additional payment scheme. The numerical examples show, that the utility of the reservation scheme administrator is higher than that of the pay-as-you-go scheme.

Keywords: pricing, cloud resources, two-stage non-cooperative game, Nash equilibrium.

1. Introduction

More and more scientific articles study economic aspects of cloud resources usage, including the issue of pricing for cloud resources (e.g., Xu and Li, 2013; Niu et al., 2012). One of the main types of cloud resources is «IaaS» - computing infrastructure (servers, data storages, networks, operating systems), provided to customers to deploy and run their own software solutions.

From technical point of view, «IaaS» is a remote set of servers and auxiliary equipment connected to a complex network; this equipment is provided to customers on a rental basis. Consequently, there is a specific characteristic associated with this approach - delay in provision of cloud resources. Queueing theory is a way to simulate such systems considering delay. This approach has been widely used within the last 10 years to study different aspects of the cloud (e.g., Anselmi et al., 2011; Ferreira, 2015). At the same time, if we consider the IaaS provider, we can distinguish a certain minimum package of cloud resources - for example, a1.medium universal instances from Amazon, by renting which the client receives 1 virtual processor and 2 gigabytes of memory per hour.

Large providers of cloud resources use different payment models. As concluded Al-Roomi et al. (2013), one of the most commonly implemented payment schemes is pay-as-you-go, in which customers pay for resources at the time and volume of their consumption. The disadvantage of this scheme is that the provider cannot plan the allocation of its resources, which may increase the delays in accessing the server. The alternative is to use a scheme where the payments are made in advance for some pre-specified amount of resources. This scheme (hereinafter referred to as reservation) allows better planning of load distribution, which leads to reduction of delays, as well as lowers prices for customers. At the same time, it is possible to combine these two schemes in order to increase revenue, plan the load on the system and reduce customer costs. Therefore, the interaction between customers and the provider considers a conflict, since the interests of the customers and the provider are different. The quality of service depends on resource allocation, prices and the load on the provider equipment.

The article studies the problem of pricing for cloud resources when introducing new payment scheme. Interests of administrators and customers are both in the scope. Scheme administrators select prices in order to maximize their own revenue. At the same time, the task of clients is to choose the payment scheme with the least possible expected costs. In the article, the two-stage model is considered. First stage is a static non-cooperative game (Osborne and Rubinstein, 1994) between the administrators for the opportunity to sell resources, where each administrator assigns the price in order to maximize his expected revenue. To simulate the reservation and pay-as-you-go schemes we use $M/M/\infty$ and $M/M/1$ (Sztrik, 2012) queues to take into account the correlation between response times and the flow rates of requests for the reservation and pay-as-you-go schemes, respectively. As a result, we derive sufficient conditions for the existence of a Nash equilibrium. In the second stage, competition among customers who wish to purchase cloud resources with minimal waiting and financial costs is studied. When the prices are set, we analyze clients choices of schemes. Here we find the Wardrop equilibrium, i.e. Nash User Equilibrium (Sheffi, 1985), achieved by clients when choosing a payment scheme.

At the end of the work, it is shown that implementation of the additional reservation scheme has a positive effect for the provider and the clients compared to a single pay-as-you-go-scheme. A numerical simulation of pricing is carried out for various values of parameters in order to determine the degree of influence of various factors on the equilibrium values of prices and utilities. However, the question of estimating the cost of the additional scheme implementation remains outside the scope. It is assumed in the paper that the provider can optimize the allocation of resources through reservation information; it is also assumed that this allows to level the costs of maintenance of the scheme.

The remainder of the article is organized as follows. Section 2 contains review of the subject area. Section 3 provides an overview of the scientific literature on cloud resources pricing. Section 4 includes a description of the pay-as-you-go and reservation scheme. In Section 5 we examine price competition among scheme administrators, as well as resource procurement competition between customers when choosing a payment scheme; a comparison of the case of one and two schemes in equilibrium is carried out, an analysis of the results of numerical simulation is given. Conclusions are formulated and possible areas for further research are indicated in Section 6.

2. Cloud Resources and Technologies

Since the inception of the cloud services market, these services have appeared in Microsoft (Azure cloud service), Amazon (AWS cloud service), Google (Google Cloud service), Yandex (Yandex.Cloud) and others. These services appeared because various companies have a need to process and store huge amounts of data. Hosting providers appeared due to the need to process, store and transfer data. These companies provide the ability to use their physical, system and software architecture for storing, processing and transmitting data. At the same time, there are two possible ways to provide access to the infrastructure - physical and cloud. The physical infrastructure assumes that the client rents a certain number of dedicated servers without the provider managing them. The virtual infrastructure uses a pool of integrated servers, controlled by the provider. With this approach, for customers it is easier to regulate their consumption, and for the provider it is easier to optimize the distribution of resources over time. The provider has full access to the information infrastructure, because of which the client can delegate the management of physical resources to the provider (Zhang et al., 2012).

Major global hosting providers, such as Microsoft (Azure division), Amazon (AWS division), Alibaba (Ali Cloud division), Google (Google Cloud division) are already actively using a payment scheme whereby customers are given a discount on cloud resources if customers guarantee the consumption of a certain amount of resources specified in the contract for a certain period of time. These discounts range from contractual consumption and length of time and may vary from 25% to 75% of the regular price (Ben-Yehuda et al., 2011). In this regard, there is a need for a reasonable determination of the size of the discount in the contract for a long-term period.

2.1. The Evolution of Cloud Technologies

At the present time, tasks of various organizations are becoming increasingly large-scale and their implementation without use of significant amounts of computing resources is nearly impossible (Sun et al., 2015). Often, a number of programs are responsible for implementing different processes, coordinating between different departments, etc. The processes of transferring information between business units within the same company have become more complicated, and computing capacities are needed to implement business processes. Due to extensive usage of cloud information technologies in many areas let us describe it on a single example of logistics. The current stage of development of information technology in logistics is called "transition to managed hosting" (Lucas D. Introna, 1991). Logistics companies refuse to invest in the creation of their own computing infrastructure and the maintenance of specialized IT personnel. In this situation, the logistics company is a tenant of the information infrastructure of the provider and acts as a user of the software installed on the equipment of the provider. This interaction between the provider and the company is carried out at the expense of cloud technologies. At the same time, all work related to hardware and software falls on the provider. The provider is responsible for maintaining the infrastructure, managing it, installing the necessary software and monitoring its condition, as well as maintaining high performance and ensuring information security.

In the future, we will use the following definition of cloud technologies (cloud computing/cloud resources), given on the official Amazon Web Services (AWS) website. Cloud Computing is the provision of computing power, cloud storage for databases, applications and other IT resources via the Internet. All types of cloud technologies can be divided into several groups according to the type of organization of cloud architecture: Private Cloud, Public Cloud, Hybrid Cloud (Al-Roomi et al., 2013).

Public Cloud is a cloud infrastructure in which the organization of work is structured in such a way that many participants can use the infrastructure simultaneously. From a technical point of view, this way of organizing work in the cloud is the simplest.

Private Cloud is a cloud infrastructure in which the organization of work is built in such a way that the infrastructure can be used only within one organization. This way of organizing cloud infrastructure is more complicated, but it allows customizing the system for the tasks of a particular organization.

Hybrid Cloud is a way of cloud architecture organization, in which the provider uses a set of cloud solutions based on Public Cloud and Private Cloud, synchronized with each other. For example, a private computing cloud, a public cloud for data storage and a dedicated server are allocated. It also supports the interaction between these components. This concept is the most flexible and modern, and therefore is in high demand. Most often, large companies use Private Cloud because it is the best way to protect their data and keep it within the organization. For the organization of cloud architecture in the case of interaction between different companies (especially in the supply chain), Public Cloud technology is better suited. Public Cloud is cheaper than Private Cloud for all participants in the chain. At the same time, the Private Cloud technology concentrates on access to data within the system, since the Public Cloud is more open.

2.2. The Current Stage of Cloud Technology Development

The main advantage of using cloud technologies in comparison with organizing the computing structure within the company is the absence of needs for significant funds to organize and maintain the information system. Thus, the company has the opportunity to free up additional resources for the development of the organization.

Among the cloud technologies, three main types can be distinguished according to the degree of their penetration into the company: Infrastructure as a Service (IaaS segment), Platform as a Service (PaaS segment), Software as a Service (SaaS segment) (Li et al., 2014).

The Infrastructure as a Service segment is a distributed infrastructure without additional software pre-installed on it and is provided to customers on rental basis. This element is most often used in cloud technologies, since its organization does not require additional costs for the development of supporting applications and the development of platforms from the provider. The provider lends only hardware with operating system (optional), and the installation of applications rests with the client company.

The Platform as a Service segment provides a platform based on a virtual infrastructure, such as the provision of a database or an operating system. This element is based on IaaS, since the provider not only develops the infrastructure, but also is responsible for installing platforms on this infrastructure.

The Software as a Service segment provides, based on IaaS and PaaS (infrastructure and installed platform), a set of programs that meets the specific needs of the client. At the moment, it is the most advanced and deeply integrated solution for the organization of the cloud system.

3. Pricing for Cloud Resources in Scientific Literature

To begin with, we review some of the works written earlier on the issue of pricing for cloud resources and discuss promising approaches that use game theory to describe the competition process among customers and providers.

Urgaonkar et al. (2013) deal with pricing from the point of view of customers and cloud infrastructure providers; in this article, various types of organizations that have their own specificity in pricing are considered as clients and providers with different types of resources. Künsemöller and Karl (2012), using game theory, explore the pricing model for cloud resources based on client costs, depending on which client can either purchase services from a cloud provider or invest in the organization of his own computing infrastructure.

Hadji et al. (2011) study pricing, taking into account the geographical location of customers. The authors consider different cases - cases of equal and different prices for different clients and a case of equality of client utility coefficients (when clients equally value the utility per unit of resource acquired). Mazzucco and Dumas (2011) examine the issue of optimal planning of server operation, when provider uses two payment schemes - premium and basic. The premium model client is obliged to pay for the reservation of equipment for a certain period (for example, a year). At the same time, premium clients can use their resources at any time, paying for their consumption. The provider is forced to pay a penalty, if he is not able to allocate equipment for the needs of a premium client. The customers of the basic scheme do not make an advance payment for the reservation of equipment, but the price for cloud resources for them may be higher than that for premium customers. In order to reflect the possibility of denial of service, it appears that customers form a Poisson flow of service requests, and two cloud service schemes –premium and basic– are presented in the form of queues. However, this paper is not concentrated on pricing, but provides interesting concept of different types of users with service privileges.

Feng et al. (2014) consider pricing for cloud resources with price competition between providers serving a common pool of customers. Each client has a Poisson flow of requests with intensity $\overline{\lambda}$. Each of the providers is represented as an M/M/1 queue. At the same time, providers have different amounts of resources, expressed in the difference between the values of service rates. Cuong et al. (2016) explore the pricing for cloud resources in the presence of two different providers - a public provider and a cloud broker, who has the ability to purchase additional resources from other public providers. Both owners of cloud resources serve a common pool of potential customers, which generates a Poisson flow of requests that splits between the owners of the cloud infrastructure. In this case, the choice of a certain cloud service provider by its customers depends on the expected response time and on the price of cloud resources. The service model of a cloud broker is an $M/M/\infty$ queue due to the ability to manage the flow of requests for provision of equipment and redirect requests to other providers from whom the broker purchased resources. The public provider, in turn, is represented as an M/M/1 queue with the same output stream parameter as that of the broker. The price for cloud resources at a broker is higher than that of a public provider, but the average time that the service request spends in the system is less. The interaction of customers and suppliers is organized in the form of a two-stage game. At the first stage, the price interaction between the public provider and the broker is a non-cooperative static game in which both administrators choose a price, which maximizes their revenue. At the second stage, at given prices, customers choose from which of the two owners of cloud resources to buy, based on price and response time. As a result, numerical modeling of prices with different values of parameters showed the broker's advantage over the public provider in terms of revenue.

4. The Model of Competitive Pricing

The main goal of this article is to analyze the interaction between clients and cloud provider, when he can apply two different payment schemes – pay-as-you-go and reservation. The next step is to compare one payment scheme case with two schemes. Prices and response times are the main characteristics that affect the stream of clients. The interaction is held between clients and the provider, and among clients and administrators. The interaction has the following structure.

1. Both administrators apply their prices in order to maximize their expected revenue. The revenue of a scheme administrator depends on the number of clients that choose this scheme. The administrator of reservation scheme also determines the volume of reserved resources.

2. Clients choose payment scheme based on response time (delay) and price. They prefer the scheme that provides them the lesser expected total cost.

3. The response time (delay), experienced by clients of a scheme depends on the provider's equipment workload. Moreover, the workload depends on the intensity of request flow to this scheme.

4.1. The Problem Formulation

The provider obtains additional information and prepayments from the clients of the reservation scheme. It allows the provider to optimize his costs and resource allocation, so the response time decreases. Thus, provider is interested in this scheme, so he is ready to provide a discount for his cloud resources (Ben-Yehuda et al., 2011). Furthermore, clients wish to procure resources with lesser expected costs and choose this scheme.

Nevertheless, some clients prefer the simpler pay-as-you-go scheme. These clients are mostly non-commercial or small commercial organizations, that cannot analyze and plan resource consumption or do not wish to overpay for unused resources.

The provider implements the schemes by appointing an administrator in lead of each of them. These administrators serve the common pool of clients. Each client has a Poisson stream of service requests with intensity $\overline{\lambda}$. When he chooses a scheme, he joins the corresponding queue, formed by all clients of this scheme. The total flow of requests to administrator is Poisson and its intensity equals the sum of this scheme clients intensities (it is considered, that clients request flows are independent).

We assume that the average response time in reservation scheme is independent of the workload due to the effective scheduling. Therefore, the chosen model for the reservation scheme system is an $M/M/\infty$ queue. Average waiting time in this system does not depend on the intensity of request flow (Sztrik, 2012). The administrator of pay-as-you-go scheme cannot schedule the workload with the same efficiency. This system can be modeled as an $M/M/k$ queue or even more complex one; in order to simplify the formulas we use $M/M/1$ queueing model. Hence, the service rates of queueing systems represent resource capacities of administrators. Further, the interaction consists of two stages. At the first stage, both administrators compete by setting their prices to maximize their revenues. However, if price is too big, clients will deviate from this scheme to the other one. Therefore, both administrators should carefully choose their prices. At the second stage, when the prices are determined, clients choose a scheme. If too many clients choose a payment scheme, it may lead to performance degradation and increase of the response time. Therefore, part of clients will choose the alternative scheme. This process ends, when the expected costs of a client equals average costs among all clients.

4.2. The Provider Model

As said before, reservation scheme clients make payment at the beginning of the contract period; that allows provider to optimize resource allocation planning and provide more stable service for customers. For example, Calheiros et al. (2011) and Wang et al. (2015) study different ways of workload forecasting and infrastructure optimization for cloud providers. We assume, that reservation allows clients to have response time independent of the total request flow rate to this scheme. We suppose that the provider is able to serve the whole pool of clients using any scheme. This assumption is necessary for existence of the stationary regime in the queues and for providing analytical results for average response times (Sztrik, 2012).

Let us turn to the description of the model. There are N clients in total and each of them has his own Poisson stream of requests for service with rate $\overline{\lambda}$. Denote by λ_1 and λ_2 the rates of the total request flows to the reservation and pay-as-you-go schemes respectively, so $\lambda_1 + \lambda_2 = N\overline{\lambda}$. Here, the service rates of both queueing systems is μ . Denote by $n\,(0 \leq n \leq N)$ the number of clients choosing the reservation scheme. Thus, the number of pay-as-you-go scheme clients equals $N - n \geq 0$. If a client chooses the first scheme, he pays in advance for a certain amount of resources $\lambda_c \cdot t$, determined by the administrator, where time of contract $t = 1$ and is omitted further as we consider the interaction during one period. If client's consumption during the contract period exceeds λ_c , then the rest part of his requests is served by the pay-as-you-go scheme.

Client's costs consist of financial and waiting parts. Financial component C_f is the price of all cloud resources procured by a client. Waiting costs C_w represent financial equivalent of total time until a client is served . Then

$$C = C_f + C_w \ .$$

The expected number of requests from a client equals his flow rate. Consider the price p_1 set, the expected financial costs of a reservation scheme client are

$$C_f = p_1\lambda_c + I\left\{\lambda_c < \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right)p_2 \ .$$

where $I\left\{\lambda_c < \overline{\lambda}\right\}$ indicates, that expected consumption exceeds the contract size. Consider the price p_2 set, the expected financial costs of a pay-as-you-go scheme client are

$$C_f = p_2\overline{\lambda} \ .$$

The average time a request spends in system waiting for service and being served at the stationary regime of reservation scheme is

$$T_1 = \frac{1}{\mu}.$$

The average response time for a request in pay-as-you-go scheme depends on the incoming flow rate and equals

$$T_2 = \frac{1}{\mu - \lambda_2} \ .$$

Due to homogeneity of clients, the total request flow intensity for the first scheme

$$\lambda_1 = n\overline{\lambda}$$

and for the second scheme

$$\lambda_2 = I\left\{\lambda_c < \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right) n + (N - n)\,\overline{\lambda}$$

where the first summand shows the total over-consumption of the first scheme clients, and the second is the total consumption of the second scheme clients.

In this article, we use the user urgency coefficient α to estimate the costs of waiting for service in monetary dimension. Average waiting costs at the reservation scheme are

$$C_w = I\left\{\lambda_c \geq \overline{\lambda}\right\}\left[\left(\overline{\lambda}t\right)\frac{\alpha}{\mu}\right] + \\ + I\left\{\lambda_c < \overline{\lambda}\right\}\left[(\lambda_c t)\,\frac{\alpha}{\mu} + \left(\overline{\lambda} - \lambda_c\right) t \frac{\alpha}{\mu - n\left(\overline{\lambda} - \lambda_c\right) - (N-n)\,\overline{\lambda}}\right],$$

where the first summand is non-zero if client does not exceed the reserved amount of resources and the second summand is non-zero in the other case; it consists of waiting costs during the contract and waiting costs of extra resources. Average waiting costs at the Pay-as-you-go scheme are

$$C_w = \left(\overline{\lambda}t\right)\frac{\alpha}{\mu - I\left\{\lambda_c \leq \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right) n - (N-n)\,\overline{\lambda}} \ .$$

Therefore, the first scheme client expected total costs are

$$C_1 = p_1 t \lambda_c + I\left\{\lambda_c < \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right) p_2 t + I\left\{\lambda_c > \overline{\lambda}\right\}\left[\left(\overline{\lambda}t\right)\frac{\alpha}{\mu}\right] + \\ + I\left\{\lambda_c \leq \overline{\lambda}\right\}\left[(\lambda_c t)\,\frac{\alpha}{\mu} + \left(\overline{\lambda} - \lambda_c\right) t \frac{\alpha}{\mu - n\left(\overline{\lambda} - \lambda_c\right) - (N-n)\,\overline{\lambda}}\right] \ . \qquad (1)$$

Similarly, the second scheme client expected costs are

$$C_2 = \left(\overline{\lambda}t\right)\frac{\alpha}{\mu - nI\left\{\lambda_c \leq \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right) - (N-n)\,\overline{\lambda}} + p_2\left(\overline{\lambda}t\right) \ . \qquad (2)$$

The revenue of the reservation scheme corresponds to the total revenue, obtained by pricing all clients of the scheme. Therefore, his utility function can be expressed as

$$U_1 = n\,(\lambda_c t)\,p_1 \ .$$

The revenue of the pay-as-you-go scheme consists of two components. The first component is the total amount of money paid by first scheme clients for the extra resources. The other component is obtained by pricing the second scheme clients. Thus, the utility function of this scheme can be written down as follows

$$U_2 = I\left\{\lambda_c \leq \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right) tnp_2 + (N-n)\left(\overline{\lambda}t\right) p_2 \ .$$

However, as we study situation during one contract period, we set t as one and skip the notion of time in formulas further.

5. Equilibrium Pricing

Two equilibria are to be obtained in this section.

1. Pair of equilibrium arrival rates $(\lambda_1^e, \lambda_2^e)$, formed by clients request flows to the first and the second scheme respectively.
2. Pair of equilibrium prices (p_1^e, p_2^e) , set by the scheme administrators.

In fact, in the next subsection we find the number n of the first scheme clients; the other $N-n$ clients choose the other scheme. Number n can be not natural, and then the ratio n/N shows the share of clients, that choose the reservation scheme.

5.1. Clients Equilibrium

With the values (p_1, p_2, λ_c) given, clients achieve the equilibrium flow rates $(\lambda_1^e, \lambda_2^e)$ by choosing the scheme. For the scheme choosing game there exist two conditions.

1. Each client individually minimizes his costs, expressed in (1) for the reservation scheme and in (2) for the pay-as-you-go scheme.
2. At equilibrium the average costs $C_1 = C_2$, are equal if there exist non-zero rate flows of requests to each scheme.

These conditions satisfy the first Wardrop principle (Wardrop, 1952). The definition of clients equilibrium for our problem can be given as follows:

Definition 1. A couple of arrival rates $(\lambda_1^e, \lambda_2^e)$ is a *Wardrop equilibrium*, if and only if there exists a constant $C > 0$ such that

$$C_i\left(\lambda_i^e\right) = C, \text{ if } \lambda_i^e > 0 \ ;$$

$$C_i\left(\lambda_i^e\right) > C, \text{ if } \lambda_i^e = 0, i = 1, 2 \ ;$$

$$\lambda_1^e + \lambda_2^e = \lambda \ .$$

Due to the connection between total flow rates, number of the first scheme clients n, client individual flow rate $\overline{\lambda}$ and the total number of clients N , Definition 1 can be reformulated.

Definition 2. Value n corresponds to *Wardrop equilibrium* if and only if there exists constant $C > 0$ such that

$$C_i(n) = C, \text{ if } N > n > 0 \ , i = 1, 2 \ ;$$

$$C_1(n) > C, \text{ if } n = 0 \ ;$$

$$C_2(n) > C, \text{ if } n = N \ ;$$

where $C_1(n)$, $C_2(n)$ are obtained from formulas (1) and (2) respectively.

At equilibrium, if $C_1 = C_2 = C$, then

$$p_1\lambda_c + I\left\{\lambda_c < \overline{\lambda}\right\} p_2\left(\overline{\lambda} - \lambda_c\right) + I\left\{\lambda_c > \overline{\lambda}\right\}\left[\overline{\lambda}\frac{\alpha}{\mu}\right] +$$
$$+ I\left\{\lambda_c \le \overline{\lambda}\right\}\left[\lambda_c\frac{\alpha}{\mu} + \left(\overline{\lambda} - \lambda_c\right)\frac{\alpha}{\mu - n\left(\overline{\lambda} - \lambda_c\right) - (N-n)\overline{\lambda}}\right] =$$
$$= \overline{\lambda}\frac{\alpha}{\mu - nI\left\{\lambda_c \le \overline{\lambda}\right\}\left(\overline{\lambda} - \lambda_c\right) - (N-n)\overline{\lambda}} + p_2\overline{\lambda}$$

There are two cases:
I: $\lambda_c \le \overline{\lambda}$ and II: $\lambda_c > \overline{\lambda}$.

In the case I, we have a trivial equilibrium. Due to the restriction $p_1 \le p_2$ both the waiting C_w and financial C_f costs for the reservation scheme clients are less than for pay-as-you-go scheme clients. Therefore, all clients choose the first scheme; this corresponds to the situation, when $n = N$.

Let us take a closer look at the case II.

Value n can be expressed as a function of λ_c, p_1, p_2 :

$$n = N - \left[\mu + \frac{\alpha}{p_2 - p_1\frac{\lambda_c}{\overline{\lambda}} - \frac{\alpha}{\mu}}\right]\frac{1}{\overline{\lambda}} \ . \tag{3}$$

Consider the inequality $0 < n < N$, we obtain the following restriction for prices values:

$$p_1\frac{\lambda_c}{\overline{\lambda}} > p_2 > p_1\frac{\lambda_c}{\overline{\lambda}} + \frac{\alpha}{\mu} - \frac{\alpha}{\mu - N\overline{\lambda}} \ .$$

5.2. Equilibrium in the Scheme Competition Model

We formalize the interaction between the administrators as a two person non-cooperative static game (Osborne and Rubinstein, 1994). The first and the second players are the reservation scheme and the pay-as-you-go scheme administrators respectively. Each player strategy is the price p_1 or p_2 respectively, and they choose them in order to maximize their utilities.

Each player can choose the strategy that maximizes his utility function when the strategy of his opponents is known. Denote by (p_1^e, p_2^e) a situation, when no player has an incentive to change his strategy unilaterally. Therefore, the point (p_1^e, p_2^e) can be obtained by best responses that are the best strategies for each player, when the other player strategy is known

$$BR_1(p_2) = \arg\max_{p_2 > p_1 > 0} U_1(p_1, p_2),$$

$$BR_2(p_1) = \arg\max_{p_1\frac{\lambda_c}{\overline{\lambda}} > p_2 > p_1\frac{\lambda_c}{\overline{\lambda}} + \frac{\alpha}{\mu} - \frac{\alpha}{\mu - N\overline{\lambda}}} U_2(p_1, p_2) \ .$$

Then Nash equilibrium for our problem can be defined as follows:

Definition 3. Situation (p_1^e, p_2^e) is a Nash equilibrium if and only if $p_1^e \in BR_1(p_2^e)$, $p_2^e \in BR_2(p_1^e)$.

According to the second order condition (Boyd and Vandenberghe, 2004), the convexity of the utility functions $U_1(p_1, p_2)$ and $U_2(p_1, p_2)$ can be characterized as shown in the following lemma.

Lemma 1. *For a given price $p_1 > 0$ the function $U_2(p_1, p_2)$ is strictly concave with respect to $p_2 \in \left[0, p_1 \frac{\lambda_c}{\overline{\lambda}} + \frac{\alpha}{\mu}\right)$. For a given price p_2 , if $p_2 < \frac{\alpha}{\mu}$, then the function $U_1(p_1, p_2)$ is strictly concave; otherwise, it is strictly concave if $p_1 \in \left[0, \frac{\overline{\lambda}}{\lambda_c}\left(p_2 - \frac{\alpha}{\mu}\right)\right]$ and convex if $p_1 > \frac{\overline{\lambda}}{\lambda_c}\left(p_2 - \frac{\alpha}{\mu}\right)$.*

Proof. The proof follows strictly from the second order conditions

$$\frac{\partial^2 U_1}{\partial p_1^2} < 0, \ \frac{\partial^2 U_2}{\partial p_2^2} < 0,$$

where

$$\frac{\partial^2 U_1}{\partial p_1^2} = \left(\frac{\lambda_c}{\overline{\lambda}}\right)^2 \frac{2\alpha\left(\frac{\alpha}{\mu} - p_2\right)}{\left[p_2 - p_1\frac{\lambda_c}{\overline{\lambda}} - \frac{\alpha}{\mu}\right]^3}, \ \frac{\partial^2 U_2}{\partial p_2^2} = 2\frac{\left(p_1\frac{\lambda_c}{\overline{\lambda}} + \frac{\alpha}{\mu}\right)\alpha}{\left[p_2 - p_1\frac{\lambda_c}{\overline{\lambda}} - \frac{\alpha}{\mu}\right]^3}.$$

Due to the lemma, to find the intersection point of two reaction curves it is necessary to solve simultaneously two maximization problems as follows:

$$\arg \max_{p_2 > p_1 > 0} U_1(p_1, p_2) \ ,$$

$$\arg \max_{p_1\frac{\lambda_c}{\overline{\lambda}} > p_2 > p_1\frac{\lambda_c}{\overline{\lambda}} + \frac{\alpha}{\mu} - \frac{\alpha}{\mu - N\overline{\lambda}}} U_2(p_1, p_2) \ ,$$

where the utility functions $U_1(p_1, p_2)$, $U_2(p_1, p_2)$ with n defined by Wardrop equilibrium as (3) are:

$$U_1(p_1, p_2) = n\lambda_c p_1 = \left(N\lambda_c - \frac{\mu}{\overline{\lambda}}\lambda_c\right)p_1 - \frac{\frac{\lambda_c}{\overline{\lambda}}\alpha p_1}{p_2 - p_1\frac{\lambda_c}{\overline{\lambda}} - \frac{\alpha}{\mu}} \ , \tag{4}$$

$$U_2(p_1, p_2) = (N - n)\,\overline{\lambda} p_2 = \mu p_2 + \frac{\alpha p_2}{p_2 - p_1\frac{\lambda_c}{\overline{\lambda}} - \frac{\alpha}{\mu}} \ . \tag{5}$$

Solving simultaneously the first order conditions $\partial U_1/\partial p_1 = 0$ and $\partial U_2/\partial p_2 = 0$ we obtain:

$$\begin{cases} p_1 = \frac{\overline{\lambda}}{\lambda_c}\left[\sqrt{\left(\frac{\alpha}{\mu} - p_2\right)\frac{\alpha}{\mu - N\overline{\lambda}}} - \left(\frac{\alpha}{\mu} - p_2\right)\right] \\ p_2 = \left(\frac{\lambda_c}{\overline{\lambda}}p_1 + \frac{\alpha}{\mu}\right) - \sqrt{\frac{\alpha}{\mu}\left(p_1\frac{\lambda_c}{\overline{\lambda}} + \frac{\alpha}{\mu}\right)} \ . \end{cases} \tag{6}$$

Let us define

$$a = \frac{\alpha}{\mu}, \ b = \frac{\alpha}{\mu - N\overline{\lambda}}, \ l = \frac{\lambda_c}{\overline{\lambda}} \ .$$

The prices (6) in the new notation take the following form:

$$\begin{cases} p_1 = \frac{1}{l}\left[\sqrt{(a - p_2)\,b} - (a - p_2)\right] \\ p_2 = (lp_1 + a) - \sqrt{a\,(lp_1 + a)} \ . \end{cases}$$

By solving the system we obtain

$$\begin{cases} p_1^e = \frac{1}{l}\left[\sqrt{b\left(a - \frac{ab(a+2b)-\sqrt{a^4b(5b+4a)}}{2(a+b)^2}\right)} + \frac{ab(a+2b)-\sqrt{a^4b(5b+4a)}}{2(a+b)^2} - a\right] \\ p_2^e = \frac{ab(a+2b)-\sqrt{a^4b(5b+4a)}}{2(a+b)^2}\ . \end{cases} \tag{7}$$

Combining first-stage and second-stage equilibrium conditions, we formulate the following definition of Nash interior equilibrium prices, suitable for the model.

Definition 4. If a pair of Nash equilibrium prices (Petrosian et al., 2012) (p_1^*, p_2^*) satisfies

$$p_2^* > p_1^* > 0\ , \tag{8}$$

$$N > n = N - \left[\mu + \frac{\alpha}{p_2^* - p_1^* \frac{\lambda_c}{\overline{\lambda}} - \frac{\alpha}{\mu}}\right] \frac{1}{\overline{\lambda}} > 0\ , \tag{9}$$

then (p_1^*, p_2^*) is *an interior Nash equilibrium.*

Theorem 1. *If* (p_1^e, p_2^e)*, defined by (7), satisfies*

$$p_2^e > p_1^e\ ,$$

then (p_1^e, p_2^e) *is an interior Nash equilibrium.*

Proof. We need to show that conditions of Definition 4 are satisfied. Condition (8) is obviously satisfied due to the theorem formulation. Condition (9) is equivalent to condition $p_1^e l > p_2^e > p_1^e l + a - b$ and is guaranteed by the theorem. We now prove that (p_1^e, p_2^e) is a Nash equilibrium.

Since we have $p_2^e < a$, by using Lemma 1 function $U_1(p_1, p_2^e)$ is strictly concave with respect to $p_1 > 0$. We can find its maximum by solving the first order condition. Since p_1^e is the root of $\frac{\partial U_1}{\partial p_1} = 0$ it maximizes $U_1(p_1, p_2^e)$. Since $p_1^e < p_2^e$ by the theorem formulation, these prices are in the feasible region.

It follows from Lemma 1 that function $U_2(p_1^e, p_2)$ is strictly concave with respect to $p_2 \in [0, p_1^e l + a]$ and $0 < p_2^e < a$; therefore, its maximum can be found as the root of $\frac{\partial U_2}{\partial p_2} = 0$, which is p_2^e .

Then, the pair of prices (p_1^e, p_2^e) satisfies all conditions in Definition 4. Therefore, the proof is complete.

However, it is important to investigate the impact of the value $l = \frac{\lambda_c}{\overline{\lambda}}$ on equilibrium prices. Since the Theorem 1 formulation, the condition $p_2^e > p_1^e$ is equivalent to $l > \frac{\sqrt{(a-p_2^e)b} + p_2^e - a}{p_2^e}$, where p_2^e, p_1^e satisfy (7). Since the value $\overline{\lambda}$ is given, the contract size of consumption λ_c needs to satisfy the following inequality

$$\lambda_c > \lambda_{bottom} = \frac{\sqrt{(a - p_2^e)\, b} + p_2^e - a}{p_2^e} \overline{\lambda}$$

5.3. Economic Effect

In this subsection, we look at the economic effect of additional scheme implementation. We check, is additional scheme profitable for provider and for clients.

Let us denote

$$U_1 = U_1\left(p_1^e, p_2^e\right) = n\lambda_c p_1^e \ , \tag{10}$$

$$U_2 = U_2\left(p_1^e, p_2^e\right) = (N-n)\,\overline{\lambda} p_2^e \ . \tag{11}$$

Utilities U_1 and U_2 in (10), (11) correspond to the revenue from the first and the second schemes respectively at prices set by formulas (7). Then, in the first case, the total revenue of the provider is

$$U = U_1 + U_2 \ . \tag{12}$$

We define by U_0 the total revenue in case of single pay-as-you-go scheme, when the whole flow of requests is served according this scheme:

$$U_0 = N\overline{\lambda} p_2^e \ . \tag{13}$$

Then we formulate the difference between the revenues in both cases as follows:

Theorem 2. *The total revenue U in the first case and the total revenue U_0 in the second case satisfy the following inequality:*

$$U > U_0 \ . \tag{14}$$

Proof. Since (p_1^e, p_2^e) satisfy conditions (7), they also fulfill (6). Then,

$$U = \alpha - \sqrt{\left(a - p_2^e\right) b}\mu + \sqrt{\left(a - p_2^e\right) b} N\overline{\lambda} + N\overline{\lambda} p_2^e - \frac{\alpha a}{\sqrt{\left(a - p_2^e\right) b}} + a\mu - aN\overline{\lambda} \ .$$

Let us denote $\varDelta U = U - U_0$. Now we show, that $\varDelta U > 0$. Indeed:

$$\varDelta U = \alpha - \sqrt{\left(a - p_2^e\right) b}\mu + \sqrt{\left(a - p_2^e\right) b} N\overline{\lambda} - \frac{\alpha a}{\sqrt{\left(a - p_2^e\right) b}} + a\mu - aN\overline{\lambda} \ ,$$

Then, after transformation we obtain

$$\varDelta U = \left(a - \sqrt{\left(a - p_2^e\right) b}\right)\left[\mu - N\overline{\lambda} - \frac{\alpha}{\sqrt{\left(a - p_2^e\right) b}}\right] \ .$$

Since $a - \sqrt{\left(a - p_2^e\right) b} < 0$, then $\varDelta U > 0$ if and only if $\mu - N\overline{\lambda} - \frac{\alpha}{\sqrt{\left(a - p_2^e\right) b}} < 0$. Let us notice, that the following two inequalities are equivalent:

$$\mu - N\overline{\lambda} - \frac{\alpha}{\sqrt{\left(a - p_2^e\right) b}} < 0, \ \sqrt{\left(a - p_2^e\right) b} < \frac{\alpha}{\mu - N\overline{\lambda}} = b \ .$$

Since $\left(a - p_2^e\right) < b$, the inequalities are verified. Therefore, $\varDelta U > 0$. The proof is complete.

As the next step, we calculate the difference between the expected costs in both cases. Let us denote by C^1 the expected costs in the single scheme case, and by C^2 the expected costs in the two schemes case. Then we have

$$C^1 = p_2^e\overline{\lambda} + \frac{\alpha}{\mu - N\overline{\lambda}}\overline{\lambda} \ ,$$

$$C^2 = p_2^e\overline{\lambda} + \frac{\alpha}{\mu - (N-n)\,\overline{\lambda}}\overline{\lambda} \ ,$$

where n is taken from (3) with respect to (7). Then the difference between client's expected costs is

$$\Delta C = C^1 - C^2 = \frac{\alpha n \overline{\lambda}^2}{\left(\mu - N\overline{\lambda}\right)\left(\mu - (N-n)\,\overline{\lambda}\right)} > 0 \ .$$

Therefore, the expected costs for clients are less in the case of two schemes; the revenue for the provider is bigger in this case. This proves the efficiency of the additional scheme implementation for both the provider and clients.

5.4. Numerical Examples

In this subsection, we calculate and analyze the numerical results of price competition modeling with different values of parameters. This allows analyzing the effect of parameters on the equilibrium prices, flow rates and administrator utilities.

Firstly, consider the impact of the service rate μ as Nash equilibrium prices highly depend on the administrators resource capacities μ .

Table 1. Utilities, prices and first scheme client share at service rate μ , $\alpha = 0.5$, $\overline{\lambda} = 2$, $N = 5$, $\lambda_c = 1.001\lambda_{bottom}$.

μ	U_1	U_2	n/N	p_1^e	p_2^e	$costC$	λ_{bottom}
30	0.0303	0.0085	0.6533	0.0024597	0.0024622	0.0426	3.7726
35	0.0216	0.0060	0.6554	0.0017291	0.0017308	0.0352	3.8069
40	0.0161	0.0044	0.6569	0.0012817	0.0012831	0.0299	3.8323
45	0.0125	0.0034	0.6580	0.0009880	0.0009890	0.0260	3.8518
50	0.0100	0.0027	0.6589	0.0007849	0.0007857	0.0230	3.8674

The results of numerical modeling of equilibrium prices, utilities and client shares at different values of service rate μ are shown in Table 1. We observe that the utility of the reservation scheme administrator is higher than the utility of the other one for each value of μ in the table. The values of equilibrium prices and expected costs C and utilities decrease, when the service rate grows, but the share of clients stays almost the same. The lower bound for the contract size of consumption λ_{bottom} grows with increase of the service rate. The resource capacity affects the equilibrium prices more, than the costs of clients.

Table 2 contains results of numerical modeling of equilibrium prices, utilities and client shares at different values of client pool size N . As expected, the equilibrium prices, utilities and expected costs increase with the growth of the pool of clients; the lower bound for contract size goes down at the same time. Nevertheless, the utilities grow faster, than clients costs. The client share of the first scheme slightly decreases, when N grows.

Table 2. Utilities, prices and first scheme client share at clients pool size N; $\alpha = 0.5$, $\overline{\lambda} = 2$, $\mu = 30$, $\lambda_c = 1.001\lambda_{bottom}$.

N	U_1	U_2	n/N	p_1^e	p_2^e	$costC$	λ_{bottom}
5	0.0303	0.0085	0.6533	0.0024597	0.0024623	0.0426	3.7726
7	0.0654	0.0194	0.6473	0.0039266	0.0039304	0.0478	3.6745
9	0.1201	0.0377	0.6409	0.0058291	0.0058350	0.0542	3.5726
11	0.2019	0.0673	0.6339	0.0083531	0.0083614	0.0623	3.4665
13	0.3222	0.1147	0.6263	0.0117942	0.0118060	0.0729	3.3557

Finally, we analyze the correlation between desired values and clients urgency α . As it is shown in Table 3, the variation of this parameter does not affect the client share and the lower bound of contract size. The other values vary in direct proportion to the change in the coefficient α .

Table 3. Utilities, prices and first scheme client share at urgency α; $N = 5$, $\overline{\lambda} = 2$, $\mu = 30$, $\lambda_c = 1.001\lambda_{bottom}$.

α	U_1	U_2	n/N	p_1^e	p_2^e	$costC$	λ_{bottom}
0.5	0.0303	0.0085	0.6533	0.0024597	0.0024623	0.0426	3.7726
1	0.0606	0.0171	0.6533	0.0049196	0.0049245	0.0852	3.7726
1.5	0.0909	0.0256	0.6533	0.0073794	0.0073868	0.1278	3.7726
2	0.1213	0.0341	0.6533	0.0098392	0.0098490	0.1705	3.7726
2.5	0.1516	0.0427	0.6533	0.0122990	0.0123113	0.2131	3.7726

The numerical examples show that the size of the clients pool has the biggest impact on values of the equilibrium prices. At the same moment, the increase in resource capacity leads to decrease in values of the equilibrium prices, and an increase in the size of client pool has the opposite effect.

Let us note that the equilibrium price for cloud resources at the first scheme is inversely proportional to the contract size. Therefore, the administrator can sell additional amount of unused cloud resources by increasing the contract size.

6. Conclusion

In this paper, the two-stage pricing model for cloud resources has been studied. At the first stage we have modelled the price competition between two administrators as a non-cooperative static game. Then, the equilibrium prices have been derived and the sufficient conditions for their existence provided. At the second stage we have found the client shares using Wardrop's user equilibrium principle. It has been shown, that implementation of the addition scheme has a positive effect on the expected costs of clients and the provider's revenue. The numerical modeling results with varying parameters show, that the client pool size and the service rate have strong influence on equilibrium prices. At the same time, at the equilibrium the utility of the reservation scheme administrator is always bigger, than the utility of the pay-as-you-go administrator.

The operating costs, which are a function of resource capacity μ provide great interest and potential for future research. Especially, the analyses of more complex M/M/k queues with priorities is another interesting way of future work. This leads to another additional problems and mechanisms, such as resource allocation between two schemes and different pricing models for different types of customers. Another interesting generalization of the current model is the case of heterogeneous clients (e.g. when their request flow rates differ).

Acknowlegments. I would like to express my great appreciation to Professor Nikolay Zenkevich for the help with working on this article.

References

Al-Roomi, M., S. Al-Ebrahim, S. Buqrais and I. Ahmad (2013). *Cloud Computing Pricing Models: A Survey.* International Journal of Grid and Distributed Computing, **6**, 93–106.

Anselmi, U., S. Ayesta and A. Wierman (November 2011). *Competition yields efficiency in load balancing games.* Performance Evaluation, **68(11)**, 986–1001.

Ben-Yehuda, O. A., M. Ben-Yehuda, A. Schuster and D. Tsafrir (2011). *Deconstructing Amazon EC2 Spot Instance Pricing.* IEEE Third International Conference on Cloud Computing Technology and Science, Athens, 304–311.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*, Cambridge University Press.

Calheiros, R. N., R. Ranjan and R. Buyya (2011). *Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments.* International Conference on Parallel Processing, Taipei City, 295–304.

Cuong, T., H. Nguyen, E.-N. Huh, C. S. Hong, D. Niyato and Z. Han (2016). *Dynamics of service selection and provider pricing game in heterogeneous cloud market.* Journal of Network and Computer Applications, **69**, 152–165.

Feng, Y., B. Li and B. Li (Jan. 2014). *Price Competition in an Oligopoly Market with Multiple IaaS Cloud Providers.* IEEE Transactions on Computers, **63(1)**, 59–73.

Ferreira, M. A. M. (2015). *Networks of Queues Models with Several Classes of Customers and Exponential Service Times.* Applied Mathematical Sciences, **9**, 3789–3796.

Hadji, M., W. Louati and D. Zeghlache (2011). *Constrained Pricing for Cloud Resource Allocation.* IEEE 10th International Symposium on Network Computing and Applications, Cambridge, MA, 359–365.

Introna, D. L. (1991). *The Impact of Information Technology on Logistics.* International Journal of Physical Distribution & Logistics Management, **21**, 32–37.

Künsemöller, J. and H. Karl (2012). *A Game-Theoretical Approach to the Benefits of Cloud Computing.* In: Economics of Grids, Clouds, Systems, and Services. GECON 2011. Lecture Notes in Computer Science (Vanmechelen K., Altmann J., Rana O.F., eds), Vol. 7150, pp. 148–160. Springer, Berlin, Heidelberg.

Li, C., X. Zhang and L. Li (2014). *Research on Comparative Analysis of Regional Logistics Information Platform Operation Mode Based on Cloud Computing.* International Journal of Future Generation Communication and Networking, **7(2)**, 73–80.

Mazzucco, M. and M. Dumas (2011). *Reserved or On-Demand Instances? A Revenue Maximization Model for Cloud Providers.* IEEE 4th International Conference on Cloud Computing, Washington, DC, 428–435.

Niu, D., C. Feng and B. Li (2012). *Pricing cloud bandwidth reservations under demand uncertainty.* Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems (SIGMETRICS '12). ACM, New York, NY, USA, 151–162.

Osborne, M. J. and A. Rubinstein (1994). *A Course in Game Theory*, MIT Press, Cambridge, Mass.

Petrosian, L. A., N. A. Zenkevich and E. V. Shevkoplyas (2012). *Game Theory.* Saint-Petersburg: BHV-Petersburg (in russian).

Sun, G., X.-Y. Wang, H. Wang and J. Zhao (2015). *Construction of Regional Logistics Information Platform Based on Cloud Computing* .International Conference on Computational Science and Engineering, Atlantis Press.

Sheffi, Y. (1985). *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Inc., Englewood Cliffs, N.J. 07632.

Sztrik, J. (2012). *Basic Queueing Theory*, University of Debrecen Faculty of Informatics.

Urgaonkar, B., G. Kesidis, U. V. Shanbhag and C. Wang (2013). *Pricing of service in clouds: optimal response and strategic interactions*. SIGMETRICS Performance Evaluation Review, **41(3)**, 28–30

Wang, W., D. Niu, B. Liang and B. Li (2015). *Dynamic Cloud Instance Acquisition via IaaS Cloud Brokerage*. IEEE Transactions on Parallel and Distributed Systems, **26(6)**, 1580–1593

Wardrop, J. G. (1952). *Road paper. Some theoretical aspects of road traffic research*. Proceedings of the Institution of Civil Engineers, **1(3)**, 325–362 Part 1.

Xu, H. and B. Li (2013). *A study of pricing for cloud resources*. ACM SIGMETRICS Performance Evaluation Review, **40(4)**, 3–12.

Zhang, S., H. Yan and X. Chen (2012). *Research on Key Technologies of Cloud Computing*. Physics Procedia **33**, 1791–1797.