Information Systems 38 (2013) 495–508

Contents lists available at SciVerse ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Cost effectiveness of commercial computing clouds

Slaven Brumec [a], Neven Vrček [b,*]

[a] Koris d.o.o., Zagreb 10000, Croatia
[b] University of Zagreb, Faculty of Organization and Informatics, Varaždin 42000, Croatia

ARTICLE INFO

Article history:
Received 4 September 2012
Received in revised form
6 November 2012
Accepted 8 November 2012
Available online 28 November 2012

Keywords:
Cloud computing
Cost effectiveness
Fractional factorial design
Symmetric mediation plan

ABSTRACT

This paper presents the procedure for comparing costs of leasing IT resources in a commercial computing cloud against those incurred in using on-premise resources. The procedure starts with calculating the number of computers as depending on parameters that describe application's features and execution conditions. By measuring required execution time for different parameter values, we determined that this dependence is a second-order polynomial. Polynomial coefficients were calculated by processing the results of fractional factorial design. On that basis we calculated costs of computing and storage resources required for the application to run. The same calculation model can be applied to both a personal user and a cloud provider. The results will differ because of different hardware exploitation levels and the economy of scale effects. Such calculation enables cloud providers to determine marginal costs in their services' price, and allows users to calculate costs they would incur by executing the same application using their own resources.

Leasing in cloud establishes a business relationship: buyer wants to reduce costs, and cloud provider wants to generate profit. This relationship will be realized if the buyer and the provider agree on a mutually acceptable fair price that can be determined by the symmetric mediation plan.

All the steps in this procedure are integrated into CCCE method and represented as a process model.

1. Introduction

Although conceived as a technology a long time ago, it is over the last few years that cloud computing has established a more extensive commercial presence [3]. Most researchers [12,18,20], users and professional public [4,7,8,10] define commercial computer clouds as a set of network services intended for providing IT services, wherein their usage:

- can be characterised as being performed on the on-demand self-service basis,
- is available through standard network technologies and protocols,
- is based on resource virtualization,
- enables rapid elasticity, that is, scalability of provided services and available resources in accordance with the user's current business requirements, and
- is charged on the pay-as-you-go basis, like conventional utilities.

A number of authors have emphasised the reduction of capital costs as a key economic benefit of cloud computing in comparison with using conventional IT resources within private server farms [9,1,15,13,14]. The pay-as-you-go model, elasticity and scalability as well as the high level of virtualization of cloud IT resources enable the cloud user to avoid investing heavily into the acquisition of their own computing equipment in advance. Capital investment costs can be distributed into gradual and monitored operating costs of lease in the computing cloud. Furthermore, virtual

* Corresponding author. Tel.: +385 99 3390 827.
E-mail address: neven.vrcek@foi.hr (N. Vrček).

0306-4379/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2012.11.002

computers are leased in the cloud include systems and applications software are incorporated into a configured server farm. The cloud user puts the cloud provider in charge of maintaining the leased virtual computers and software as well as of configuring the server farm, thus freeing their IT staff from these jobs and enabling them to focus on the development of IT innovations and business applications support.

On their portals, commercial cloud providers offer a calculator for estimating the costs of leased cloud IT resources which, however, fail to address two crucial questions: what is the amount of IT resources that a particular application will use? and What is the cost of maintenance of those applications of on-premise computers within a conventional server farm?

A good indicator of cost effective usage of IT resources in a commercial cloud, both for the buyer and the provider, is the fair price of leasing those resources. Let C_P[$/h] be the cost of computing resources, and D_P[$/GB h] the cost of storage resources that the buyer would incur in using their privately-owned resources to run the application. Let C_S and D_S be adequate costs for the cloud provider. The price C and/or D at which the cloud provider leases IT resources needs to be sufficiently high to ensure profit (i.e., $C > C_S$ and $D > D_S$), but also sufficiently low to make leasing cloud resources more cost effective for the buyer than purchasing and using on-premise computing resources (i.e. $C < C_P$ and $D < D_P$ need to be fulfilled). Although cloud providers argue that for the buyer leasing is a more favourable option than purchasing, in this paper we show that this does not necessarily apply to all cases. A prerequisite for objective decision-making regarding the purchase and lease of IT resources was setting the quantitative criteria, which was achieved by developing adequate cost calculation methods and their integration into a uniform method presented in the following sections of this paper. The aforementioned issues are scientifically relevant as they indicate possible future avenues of development of cloud computing business models as well as the impact of the cloud computing paradigm on business subjects and their competitive edge.

It has to be noted that in this paper commercial clouds are explored, which excludes academic-only computing clouds. Moreover, in this paper the security and legal issues of cloud computing usage are not specifically addressed. Instead, it is assumed that users act within a legal framework that provides an adequate environment for cloud computing and that cloud-related security concerns have been resolved by cloud providers.

2. Research overview

The data for the analysis of cloud services price as viewed from the providers' perspective is hard to find in the literature since investors do not tend to reveal their calculations to the public. That makes the design of appropriate cloud service models more difficult for scientists. An influential paper in the field of cloud cost analysis was contributed by Xinhui et al. [22], whose authors originate from the IBM Research Lab in China. Their results correspond with those in the study by Sawyer [16], which deals with the cost of data centres in general, not only cloud-based ones.

From the investors' point of view, an important feature of cloud computing is the total cost of ownership (TCO) over a cloud data centre. Once the TCO has been calculated, it is possible to define general factors that determine operating costs and the terms of cloud cost effectiveness, e.g. the minimum price of a CPU hour or other calculation factors. The paper by Xinhui et al. [22] is significant in that respect as it is applicable to the TCO calculation method for any computing cloud.

Important contribution to the analysis of cloud services price from the users' perspective was given by Walker [19]. Detailed description of the procedure for comparing the price of cloud computing resources against the cost of acquisition and maintenance of on-premises hardware is provided in [19], which is of critical importance in lease versus purchase decision-making.

Unlike conventional financial models for analysis of purchase or lease of capital assets, Walker's model does not only include amortisation, but also takes into consideration the change in the CPU performance over time in accordance with Moore's law. Furthermore, Walker's model makes it possible to calculate the price of a CPU hour at the moment of purchasing computing resources, the price of a CPU hour for the purchased resources including the regular annual computing capacity update and the theoretical price of a CPU hour for leasing computing resources in a cloud.

Leasing cloud computing resources can generally be observed as leasing processor time expressed by means of the price of a CPU hour. With certain cloud providers (e.g. Google App Engine) that is the only resource that is explicitly leased. With other providers (e.g. Amazon or Windows Azure) CPU time is leased implicitly, through virtual computers use, wherein configurations with precisely defined features such as the number of cores in a processor, disk capacity, size of RAM and bandwidth are available to users.

Walker's work is therefore universally applicable when the comparison of the price of on-premise resources against the price of leasing resources in any commercial cloud is concerned. Walker's research also proves the intuitive fact that the exploitation of cloud IT resources is the crucial factor in the price of computing services. A higher level of average processor occupation implies that the cloud provider can offer cloud services to the user at a lower commercial price.

3. Method for calculating cloud computing cost effectiveness (CCCE)

The review of the recent literature has led the authors of this paper to conclude that a comprehensive theoretical model that would link the application features to leased cloud resources has so far not been devised. Within our research we therefore developed a method for calculating cloud computing cost effectiveness (CCCE method), which can be implemented in deciding objectively whether to purchase or lease computing resources. The method is

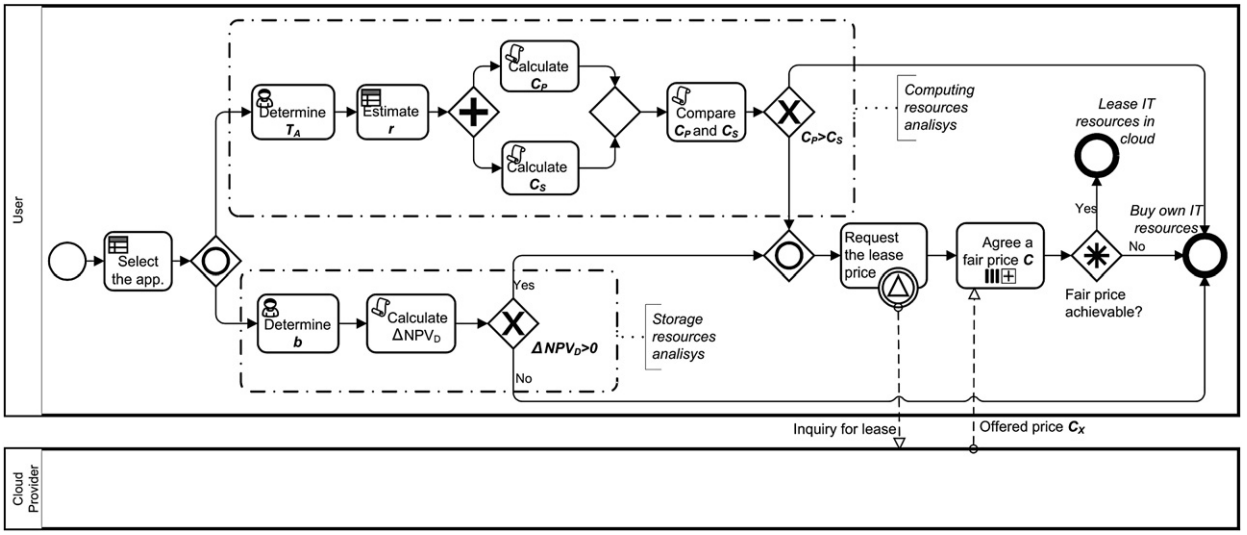

Fig. 1. CCCE method.

shown as process model in accordance with BPMN 2.0 in Fig. 1.

The method starts with determining T_A—the acceptable time required for the execution of a selected application. Next, the number of computers ***r***, upon which the selected application will be run so that T_A can be obtained, is selected. On the basis of the number of required computers ***r***, it is possible to further calculate the costs of on-premises computing services C_P and the costs of their lease in the cloud C_S (C_S is the cost price which does not include the provider's profit). Owing to higher equipment exploitation and the economy of scale effect, it is expected that $C_P > C_S$, which indicates that in this case the required resources should be obtained by leasing them in the cloud. If $C_P \leq C_S$ then purchasing on-premise resources is recommendable. When leasing cloud resources, quotations from several cloud providers should be requested. Because of the expected profit providers will quote the price C_X that is higher than the cost price C_S. Therefore the parties need to negotiate over a mutually acceptable price C. The decision to lease will ensue if the user has succeeded in agreeing on C with the provider, in other words, if $C_P > C > C_X \geq C_S$. This part of the procedure is represented in the upper section of Fig. 1.

A similar procedure is conducted for storage resources, starting with the estimation of database size ***b*** that the selected application works with. During the procedure, D_P—the costs of on-premise data storage and D_S—costs of storage resources that the cloud provider is expected to incur are calculated. It is also expected that leasing storage resources at price D will be realized if $D_P > D > D_X \geq D_S$ (where D_X is the price of data storage required by the service provider). This part of the CCCE method is shown in the lower section of Fig. 1.

Other details concerning the CCCE method are evident from Fig. 1. The content of particular activities and the explanation of the used labels are given in Table 1. The most important activities within the process model are based on the research conducted by the authors of this paper, which are described in the following section, while other activities are based on the studies analysed in the previous section.

4. CCCE method overview and implementation

Since the usability of a proposed method cannot only be analysed in theoretical terms, an initial case study was conducted that is based on its real-world application. This was further extended by other case studies which further verified the method and that are described at the end of this paper. In this section particular steps of the CCCE method are described following the order introduced in the upper part of Table 1. After a theoretical description of each step, its executability is verified by means of the case study. For simple steps, a short textual description containing relevant references is provided. The steps that incorporate algorithms that arose from the authors' targeted research are described in more detail.

4.1. Select application (1)

A business user will select an application that with regards to its business features and architecture is equally suitable for execution on on-premise IT resources and in the cloud. A sample application (SA), which in our case is used for the purpose of method applicability verification, is aimed at facial recognition. The way in which the application operates is as follows:

1. There is a reference database, adopted from the University of Essex [17], that contains portraits of approximately 400 persons.

2. The user sends the image of a person to be identified on the basis of similarity with one of the portraits in the database to SA. The user can be a person or automated software. A person will send the image via an executable programme or a web application, while automated software sends it via the SA web service.

Table 1
CCCE method activities and description of the used labels.

Activity number, name and description		
1	Select the application	Determine the application to be supported by IT resources
2	Determine T_A	Determine the acceptable application execution time
3	Estimate r	Determine the number of computers required for realizing T_A
4	Calculate C_P	Calculate the costs of on-premise computing resources
5	Calculate C_S	Calculate the cost price of provider's computing resources
6	Compare C_P and C_S	Compare the costs of on-premise computing resources with the costs in cloud
7	Determine b	Estimate the data volume and data increase rate
8	Calculate ΔNPV_D	Calculate the ratio between leasing and purchasing storage resources
9	Request the lease price	Analyse the costs of leasing IT resources in a commercial cloud
10	Agree a fair price	Agree a mutually acceptable price of leasing IT resources in a commercial cloud
Variable label and description		
T_A[s]		Application execution time
r [number]		Number of computers involved in running an application
C_P [$/h]		Costs of on-premise computing resources
C_S [$/h]		Cost price of cloud computing resources
C_X [$/h]		Requested price of leasing cloud computing resources
C [$/h]		Agreed price of leasing cloud computing resources
b [GB]		Size of database that the selected application works with
D_P [$/GB m]		Costs of on-premise storage resources
D_S [$/GB m]		Theoretical costs of cloud storage resources
D_X [$/GB m]		Requested price of cloud storage resources
D [$/GB m]		Agreed price of leasing cloud storage resources

3. Each instance of SA, run from another computer, reads images from the reference database and compares them with the image sent for recognition. The result of the operation can either be the selection of one image from the reference database (that most closely matches the image sent for recognition) or no image, in case the image sent for recognition differs considerably from each image in the reference database.

4.2. Determine acceptable execution time (2)

Application execution time T_A, which depends on the used computing power, can also be considered from the business perspective. Shorter execution time means that the user will obtain the required information on a certain state or a business event faster, and thus achieve a particular effect E for the organisation that can be measured in terms of profit or another value. In general, for the achieved effect $E=f_E\cdot(1/T)$ is applicable, wherein execution time is defined in the interval $T=[T_{MIN},T_{MAX}]$.

T_{MAX} is the longest acceptable application execution time since the purpose of using an application would diminish if it was executed for a longer period of time. T_{MIN} cannot be deliberately reduced since it is restricted by computing and communications resources on which the application is run. From the organisational perspective, a response obtained in an extremely short time is usually not necessary. It follows from the above that in planning the required computing resources, regardless of whether they are on-premise or leased in the cloud, the organisation needs to determine acceptable application execution time T_A for which $T_{MIN}\le T_A\le T_{MAX}$ applies. This time will be used in calculating required computing resources. In the case of SA is was estimated that the acceptable time is within the interval $60[s]\le T_A\le 180[s]$.

4.3. Estimate the number of computers (3)

It can be assumed that application execution time T_A, regardless of whether it is executed in a privately-owned data centre or in the cloud, is in a way functionally dependent on the number of used computers. The form of this functional dependence is unknown. It can be expected that T_A will also depend on other variables in addition to the number of used computers. On the basis of estimates of 8 independent experts who were interviewed by the authors of this paper it was assumed that application execution time T_A depends on the number of used computers as well as on five other independent variables, which can be described as

$$T_A=f_T(p,a,u,b,q,r) \tag{1}$$

Labels and descriptions of independent variables are shown in columns 1 and 2 in Table 2.

The analytical procedure whereby the assumed functional dependence f_T could be described is unknown. We hypothesised that the aforementioned functional dependence in each particular case could be determined by processing the data obtained by measuring T_A for different values of independent variables. If non-linear dependence of T_A on independent variables is assumed, then each measurement should be conducted at a minimum of three points for each possible combination of values of the 6 independent variables. This process should be repeated three times in order to determine the measurement error, in other words, $3\times 3^6=2187$ measurements should be conducted (6 variables, three measurement points, each measurement repeated three times). As a matter of fact, this procedure is impossible to conduct owing to its duration (356 h, with an average time of a single measurement amounting to 10 min), and the processing of the measured data would be extremely complex. The

Table 2
Examined variables for calculation of function $T_A=f_T(p,a,u,b,q,r)$.

Label (1)	Independent variable description (2)	Unit of measure in the real domain (3)	Real domain (4)	Independent variable variation range in the real domain (5)	Substitution for reduction of the real domain to [−1,+1] (6)	Measurement plan (7)
p	Power of used computers	integer	[1,4]	1: one core, 2 GB RAM 4: four cores, 8 GB RAM	$x_1=\frac{2}{3}\cdot p-\frac{5}{3}$	a
a	Application complexity	s/transaction	[1,10]	Number of repetitions of the same action in the programme	$x_2=\frac{2}{9}\cdot a-\frac{11}{9}$	b
u	Volume of data to be processed	KB	[5.4,540]	Size of data package to be processed	$x_3=\frac{2}{534.6}\cdot u-\frac{545.4}{534.6}$	c
b	Database size	MB	[1.2,5.8]	Size of the DB that the selected app. works with	$x_4=\frac{2}{4.6}\cdot b-\frac{7}{4.6}$	d
q	Number of CRUD actions per query	number/query	[10,200]	Number of instances of reading, writing, updating and deleting in the DB for each query	$x_5=\frac{2}{190}\cdot q-\frac{210}{190}$	e
r	Number of used computers	integer	[1,5]	Number of computers on which the app. can be run	$x_6=\frac{1}{2}\cdot r-\frac{3}{2}$	f

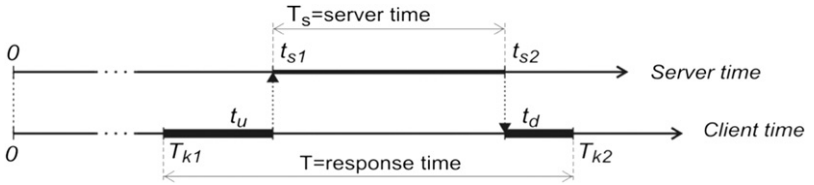

Fig. 2. Measuring SA execution time in the cloud.

obtained results would not be reliable since the effects of the examined variables over an extended measurement cannot be separated from the uncontrolled impacts of incidental variables.

For executable calculation of function (1) we adopted a fractional factorial design following the plan defined as 2^{k-r} (where k is the number of independent variables and 2^r replica size), as described in detail in [2]. Such a measurement plan is suitable for investigating phenomena dependent on more than three variables, wherein the type of functional dependence is unknown. If we assume that: (1) the examined phenomenon could be described with a second-order polynomial; (2) the impact of third-, fourth-, fifth-order variables is negligible and (3) it is possible to exclude certain two-factor interactions that can definitely not occur (e.g. those between variables ***p*** and ***u***), then on the basis of $2^{6-2}=16$ measurements it is possible to calculate the constant member of the polynomial, 6 coefficients that show the linear effect of independent variables on T_A and the coefficients for 9 expected two-factor interactions.

The executability of steps was verified by measurement on SA, which was structured so as to allow modification of the values of independent variables within the domain that is provided for each independent variable in columns 4 and 5 in Table 2. The measurement of execution time T_A was conducted on leased cloud resources, in accordance with the outline in Fig. 2.

Two aspects of application execution time were measured: computer T_R (i.e. the time of application execution by computing resources) and user T_K (i.e. the time of application execution from the user's perspective, from the moment at which the data are sent for processing to the moment when the user receives the results of processing). When an application is executed in the cloud, $T_K > T_R$ applies, owing to the time it takes for the data to be processed to be sent to the cloud and the time it takes for the user to receive the results of processing.

In accordance with the fractional factorial design defined as 2^{6-2} measurements are taken for two values of each independent variable and are repeated at the central point of the plan [2]. For easier processing of measurement results it is assumed that the lower margin value of each independent variable is -1 and the upper margin value is $+1$. Considering that lower margin d and upper margin g are real numbers different from -1 and $+1$, it is necessary to conduct substitution for each explored variable in accordance with (2). The results of such substitution are listed in column 6 in Table 2.

$$x_i=\frac{-2}{d-g}X_i+\frac{d+g}{d-g} \tag{2}$$

The results of 16 conducted measurements for T_A are shown in Table 3. Labels 1 ab, cd, abcd … abcdef refer to the schema according to which independent variables are set during measurement. The appearance of a certain letter during measurement means that this particular variable, to which the letter is attributed according to column 7 in Table 2, is set at the upper margin. If that particular letter does not appear in the measurement

Table 3
Measured T_A according to measurement plan 2^{6-2} and estimated regression coefficients.

Measurement plan (1)	Measured T_A [s] (2)	Coefficient value (3)	Coefficient significance (4)	Estimated effects (5)
1	24	155.6		
ab	119	38.3	**91.99**	**B**+3+3+5
cd	103	2.1	0.27	D+3+5+3
abcd	198	4.3	1.17	BD+AC+4+4
ace	232	66.7	**278.70**	**E**+5+3+3
bce	343	3.9	0.97	BE+4+AF+4
ade	134	−1.8	0.21	DE+4+4+CF
bde	295	6.4	2.60	BDE+3+3+3
acf	73	−25.4	**40.55**	**F**+5+3+3
bcf	124	−19.4	**23.68**	**BF**+4+AE+4
adf	18	0.6	0.02	DF+4+4+CE
bdf	52	−1.9	0.24	BDF+3+3+3
ef	149	−3.3	0.69	EF+6+AB
abef	164	−6.3	2.50	BEF+5+A+3
cdef	205	36.2	**82.07**	DEF+5+3+**C**
abcdef	256	0.2	0.00	BDEF+4+AD+BC
0	152	The mean value is given for all parameters.		
	161	Measurement is repeated 3 times, and the results		
	156	will be used for the assessment of the adequacy of the mathematical model.		

plan, the particular independent variable is set at the lower margin. For instance, the measurement plan labelled *cd* (third row in Table 3) means that T_A was measured for the case in which variables ***u*** and ***b*** were at the upper margin of the examined domain ($u=540$[KB] and $b=5.8$[MB]), whereas four other variables were at the lower margin. According to the same convention, the plan labelled 1 means that all the variables are at the lower margin, while the abcdf means that all the variables are at the upper margin. Such a plan is symmetric as each independent variable appears an equal number of times at the lower and upper margin of the examined domain.

Based on the measurement results (column 2 in Table 3) we calculated the coefficients of polynomial (3) that describes theoretical dependence of application execution time T_A on the investigated variables.

$$T_A = 155.6 + 38.3x_2 + 36.2x_3 + 66.7x_5 - 25.4x_6 - 19.4x_2x_6 \quad (3)$$

Polynomial coefficients listed in column 3 were calculated using the Yates procedure [2]. By using this procedure in accordance with the measurement plan specified in column 1, it is possible to estimate the values of effects of individual variables (***p***,***a***,***u***,***b***,***q***,***r***) and their interactions, as stated in column 5 (letters in uppercase express the coefficient of the effect of the corresponding variable labelled with the same letter in lowercase in the measurement plan). For instance, coefficient 38.3 that pertains to variable x_2 represents the effect of the application's relative complexity (according to Table 2, x_2 is substitution for ***a***), wherein the confounded effects of two three-factor and one five-factor interaction are included. However, according to Berger and Maurer [2], the effects of three-factor interactions of independent variables and higher-order interactions upon the dependent variable are negligible so it can be considered that the value 38.3 pertains only to coefficient D that represents the effect of variable x_2 on T_A. Similar considerations apply to other coefficients as well.

The significance of polynomial coefficients (3) was examined by using analysis of variance (ANOVA). Their significance coefficient is shown in column 4 in Table 3. **B,C,E** and **F** are significant coefficients, since they exceed the upper value $k_F=7.7086$ for F-distribution with $\alpha=0.05$ and degrees of freedom $df=(1.4)$. The coefficient that amounts to −19.4 is also significant and it estimates the confounded effects of interactions of variables bf and ae. However, considering that variable a refers to computer power (labelled ***p*** in Table 2) and variable e refers to the number of CRUD operations in the database (labelled ***q*** in Table 2), no synergy effect arising from their interaction is actually expected. Therefore it can be considered that the calculated value pertains only to coefficient BF that estimates the interaction of variables bf.

The significance of the mathematical model for T_A was tested by the Fisher's coefficient of model adequacy. The obtained $f_F=19.353$, which exceeds the upper value 15.23 for F-distribution with $\alpha=0.05$ and $df=(2.7)$.

By substituting variables x_2, x_3, x_5 and x_6 in (3) for the original values, and assuming average complexity algorithm ($a=4$) the final version of the polynomial is obtained in (4) and represents an adequate mathematical model of the dependence of sample application execution time on independent variables in the real domain.

$$T_A = 60.603 + 0.135u + 0.702q - 9.48r \quad (4)$$

From expression (4) it is evident that sample application execution time does not depend on all the six variables, as it was originally assumed. That time increases with the relative complexity of application ***a*** and the volume of data sent for processing as well as with the increase in the number of CRUD operations ***q***, while it is reduced by the using a larger number of computers ***r***.

The final member of the polynomial (interaction $\boldsymbol{a}\cdot\boldsymbol{r}$) indicates that the effect of the increase of the number of computers on the reduction of application execution time will be proportional to the complexity of application. By rearranging (4), expression (5) is obtained that states the number of required computers as depending on T_A and application requirements.

$$r=\frac{0.135u+0.702q-(T_A-60.603)}{9.48} \tag{5}$$

If SA is described by parameters $a=6$, $u=120$, $q=60$, then the number of required computers $\boldsymbol{r}$ can be calculated as depending on T_A, as shown in Fig. 3.

In this way it is possible to determine the required number of computers $\boldsymbol{r}$ for each application with known requirements. Significant variables and function $T=f_T(p,a,u,b,q,r)$ will differ across applications although the methodological procedure will remain unchanged. It has to be noted, however, that the established functional dependence applies within the domain that has been defined by variation limits of investigated effects.

This case study showed the use of the methodology and verified correctness of described mathematical expressions within given domain. Financial parameters were further elaborated in several case studies (named *App1*, *App2* and *App3*) described in subchapter Request the lease price. More relevant, production domain, case study which incorporates resources and calculation of costs is given at the end of this paper i.e. subchapter Production domain case study.

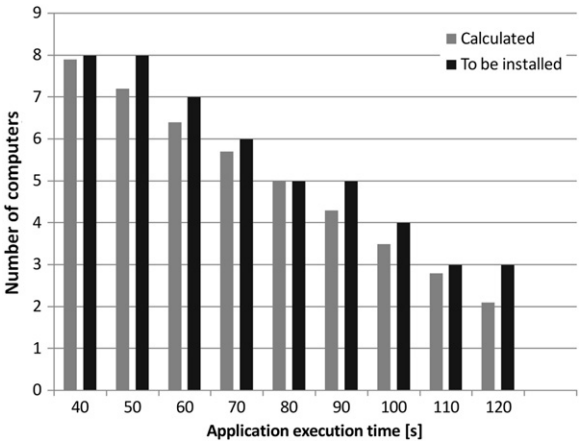

Fig. 3. Dependence of T_A on the number of computers for various application requirements levels.

4.4. Calculate the costs of on-premise computing resources (4)

The cost of on-premise computing resources, expressed as the hourly price of the operation a single CPU core can be calculated using Walker's model [19]. For the purchase option, the price of a CPU hour is calculated so that initial investment in servers and other equipment, to which annual maintenance costs are added, are divided by the exploitable technical capacity *TC* obtained by the purchased equipment. *TC* is calculated as the product of the number of utilised processors *TCPU*, their expected annual number of hours of operation *H* and server exploitation coefficient η:

$$TC=TCPU\cdot H\cdot\eta \tag{6}$$

All the expenses during the technological lifetime of the equipment are reduced to NPV. Total investment, according to Xinhui et al. [22], is comprised of initial investment (purchasing computers, software and networking and other equipment) and annual operating costs (costs of staff, operating and cooling energy and space renting for data centre hosting). Unlike conventional financial models for analysing the options of leasing versus purchasing particular capital assets, Walker's model does not only include computing equipment amortisation and the interest on investment, but also takes into consideration the relative change in the CPU performance over time in accordance with Moore's law. According to Walker [19], data centre ownership costs expressed as the price of a CPU hour [$/h] of a processor core, are calculated following the expression:

$$C_P=\frac{\left(1-\frac{1}{\sqrt{2}}\right)\sum_{T=0}^{Y-1}\frac{C_T}{(1+k)^T}}{\left(1-\left(\frac{1}{\sqrt{2}}\right)^Y\right)TC} \tag{7}$$

In the formula above, C_T refers to annual investment. For the first year initial investment and annual operating costs are included here, whereas for subsequent years

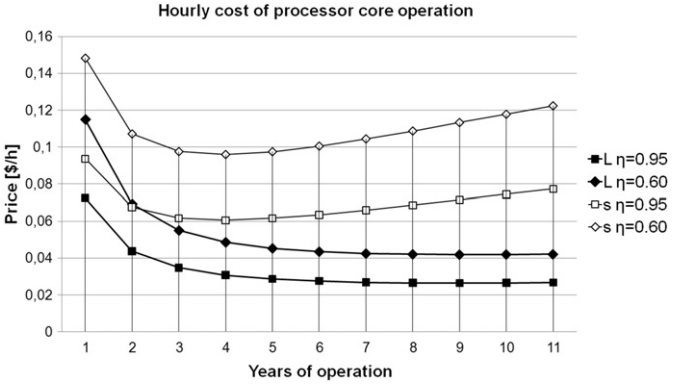

Fig. 4. Hourly cost of processor core operation in a data centre.

only the operating costs of running the data centre are included. The numerator in (7) refers to cumulative investment in running the data centre, discounted to NPV. The denominator is the technical capacity that diminishes over time in accordance with Moore's law.

The calculation above was verified for two virtual data centres: large (***L***) with 15,000 servers and annual operating cost $C_T=7\cdot10^6$[$], and small (***s***) with 1500 servers and annual operating cost $C_T=1\cdot10^6$[$]. Servers in each data centre contain 4 processor cores and work at two different degrees of exploitation $\eta=0.95$ and $\eta=0.60$. If the price of such a server with its corresponding equipment is 2000[$], the initial investment is $C_{0L}=30\cdot10^6$[$] and $C_{0S}=3\cdot10^6$[$]. If we assume that the annual price of capital $k=7$[%], formula (7) yields the result shown in the graph in Fig. 4.

In general, annual operating costs in comparison with initial investment are relatively higher in a smaller data centre owing to better exploitation of human resources and auxiliary equipment. Explanation for that is found in [5,6], where it is stated that in a small data centre one systems administrator maintains up to 100 computers, whereas in a large data centre up to 1000 computers are maintained by a single systems administrator. For this reason it is expected that the operating cost of a small number of computers for personal needs (expressed as the price of CPU operation) will be higher than in a large data centre. Moreover, when investment into on-premise computing resources is considered, meeting the peak demand also needs to be taken into account, which results in a lower average degree of their exploitation. By contrast, cloud providers can expect a permanently high degree of exploitation of their computing resources, staff and auxiliary equipment, which results in the lower price of CPU operation per hour. Such a difference in the price is an IT version of economy of scale.

Fig. 4 shows that the hourly cost of the operation of a processor core depends on exploitation coefficient η at which the data centre is working and on the technological lifetime of data centre exploitation (the length of the *x*-axis in Fig. 4 does not imply the technological lifetime of the data centre; the calculation covers a 11-year period so that the trend is more visible). A lower degree of exploitation means a higher price of a CPU hour. The price of a CPU hour will be minimal if the data centre is to be amortised within 5 years.

4.5. Calculate the costs of leasing cloud computing resources (5)

Costs of leasing can also be calculated using (7) if the change in the processor capacity over time is not included (assuming that the user always has access to new equipment) and considering that no initial investment exists and only annual operating costs of lease are incurred. For such a case, expression (8) is obtained for the price of a CPU hour that the user pays when leasing computing resources in the cloud:

$$C_S=\frac{\sum_{T=0}^{Y-1}\frac{C_T}{(1+k)^T}}{Y\cdot TC} \tag{8}$$

4.6. Compare the costs of on-premise computing resources and cloud resources (6)

In this step the costs of required on-premise computing resources C_P are compared with C_S—the costs the user would incur if leasing those resources in the cloud. The result of this calculation, in case the cloud resource provider requires the price C=0.08[$/h], is also shown in Fig. 4. Although it is considered that leasing computer resources is more favourable for the user than their purchase, it might not be generally true. From the data obtained in the calculation it is evident that leasing would not be a more cost effective option if the user could ensure a very high degree of exploitation of purchased resources.

4.7. Estimate the data volume and data increase rate (7)

The volume of data ***b*** that a particular application works with depends on a company's data pool. Therefore business analysts and software engineers should work together to determine the current needs and the annual increase of storage resources, based on the data pool volume and type.

4.8. Calculate the ratio between leasing and purchasing storage resources (8)

The more favourable option (lease versus purchase) for obtaining the required disk capacity can be estimated following the procedure in [21]. In that paper, the authors calculated the net present value of profit NPV_P for the case in which purchased disks are utilised by the user for number of years *T* and the net present value of profit NPV_L for the case in which leased disks are used over the same number of years. If the difference $\Delta NPV_D=NPV_P-NPV_L$ is positive ($\Delta NPV_D\geq 0$), then purchase is a more favourable option for the buyer (as higher profit remains with the user), whereas if $\Delta NPV_D\leq 0$, leasing presents a more favourable alternative.

A detailed study in [21] shows that ΔNPV_D can be expressed as:

$$C_P=\frac{\left(1-\frac{1}{\sqrt{2}}\right)\sum_{T=0}^{Y-1}\frac{C_T}{(1+k)^T}}{\left(1-\left(\frac{1}{\sqrt{2}}\right)^Y\right)TC} \tag{9}$$

where

$$\Delta NPV=\sum_{T=0}^{N}\frac{C_T-E_T+L_T}{(1+k)^T}+\frac{S}{(1+k)^N}-C \tag{10}$$

$$E_T=(1{,}03\lceil V_T\rceil_\Omega-\lceil V_{T-1}\rceil_\Omega)\Omega Ke^{-0{,}438T} \tag{11}$$

$$C_T=-\rho H_T-8760\delta(P_C+P_D)\lceil V_T\rceil_\Omega) \tag{12}$$

The explanations of labels in the expressions above are provided in first three columns in Table 4. Expression (10) represents the remaining value of disks after the number of years of use T. Expression (11) yields the costs of the capital invested in purchasing disks of nominal capacity Ω, which includes the required capacity enlarged by 3% owing to possible breakdowns and the decrease in the

Table 4
Parameters of the model for calculating ΔNPV_D value.

Label (1)	Parameter description (2)	Unit of measure (3)	Sample calculation data		
			Example 1 (4)	Example 2 (5)	Example 3 (6)
δ	Price of electricity	*[$/kWh]*	0.05	0.05	0.05
Ω	Initial disk capacity	*[GB]*	500	3000	5000
ρ	Difference between the costs of using on-premise and leased disks	*[relative number]*	0	0.5	0.5
γ	Decrease in the disks market value	*[relative number]*	0.2	0.2	0.2
C	Price of a disk control unit	*[$]*	500	1000	2000
H_T	Annual staff costs	*[$]*	–	50000	75000
k	Interest rate on financial assets	*[$/$]*	0.08	0.08	0.08
K	Price of disk capacity for T=0	*[$/GB]*	0.30	0.30	0.30
L_T	Annual cost of disk capacity lease	*[$/GB]*	1.8	1.8	1.8
P_C	Disk control unit power	*[kW]*	0.50	0.70	1.00
P_D	Power required by a disk unit	*[kW]*	0.01	0.01	0.01
V_T	Required annual additional disk capacity	*[GB]*	100	1000	10000
	Operator for calculating the number of disks with capacity Ω required for sufficing the required capacity V_T				
	$\Delta NPV_D = NPV_P - NPV_L$ (after 3 years)	−456[$]	−18023[$]	50198[$]	

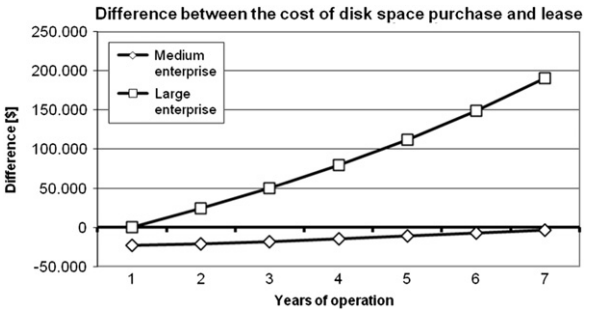

Fig. 5. Comparison between disk purchasing and leasing for small/medium and heavy users.

price of disks over the years. Expression (12) is used to calculate operating energy and staff costs.

The calculation of ΔNPV_D was verified in three business case studies, the data for which are found in last three columns in Table 4.

The first case refers to an individual user whose initial requirements are 500[GB] and are annually increased by 100[GB]. The second case concerns a small enterprise whose initial requirements are 3[TB] and are annually increased by 1[TB], whereas the third case refers to a heavy user whose initial requirements are 5[TB] and are annually increased by 10[TB]. Results of the calculation, given in the last row in Table 4 and showed in Fig. 5, lead us to conclude that:

- Leasing is a more favourable option for individual users or fairly small enterprises, provided it does not exceed the technological lifetime of the equipment.
- For small and medium-sized enterprises leasing disk space is always financially more favourable than investing into privately-owned disk capacity.
- For large enterprises and other organisations with huge disk storage requirements purchasing the equipment is by far a more favourable option than leasing.

Obviously, the calculation of ΔNPV_D depends on a number of parameters, so the obtained results will apply as long as the value of parameters in Table 4 remains unchanged. Based on the presented method, every organisation needs to conduct own calculation and decide on the more favourable option for ensuring the required disk capacity.

4.9. Request the lease price (9)

In the previous steps of the method the number of computers ***r***, required for execution of an application during T_A was obtained. In general, it is evident from (1) that ***r*** depends on T_A and 5 variables that describe the application. Although cloud service providers calculate their services in various ways, parameters for calculating the quoted price almost regularly include variables ***p***,***u***,***b***,***q*** and ***r***. Variable ***a*** is indirectly included since it influences the number of computers to be leased. Furthermore, the price of service is affected by some other specific parameters (e.g. type of OS used by virtual computers).

The number and type of parameters for lease calculation vary between providers. Therefore, to enable comparison between quoted prices, we developed formulas for calculation of the total price of PaaS for Amazon, Microsoft and Google. Parameters used by particular providers and unit prices at which lease services are calculated are shown in Table 5.

Total prices are expressed as C_S, where S refers to the provider, so that $P \in \{A,M,G\}$ applies to Amazon, Microsoft and Google, respectively. Parts of the total price are expressed as $P_X \cdot y$, where P_X is the unit price of x^{th} service provided by provider P, and y refers to the number of leased unit services. For indices x and y the following applies: $x,y \in \{r,p,u,d,b,q,rb,os,e\}$.

The lease of computing and storage resources integrated in the formulas can be expressed as follows:

$$C_A = A_r r p A_{os} + A_u u + A_d d + A_q q + A_b b + A_{rb} rb \quad (13)$$

$$C_M = M_r r p + M_u u + M_d d + M_q q + M_b b + M_{rb} rb \quad (14)$$

$$C_G = G_r r + G_u u + G_d d + G_b b + G_e e \quad (15)$$

Table 5
Price calculation parameters.

	Description	Domain and unit of measurement	Amazon (A)	Microsoft (M)	Google (G)
r	Number of used computers	{1,2, …}[unit]	A_r=0.085[$/h]	M_r=0.12[$/h]	G_r=0.10[$/h]
p	Power of leased computers	Typical configurations []	*p*=1 (1.7 GHz; 4 GB RAM) for A, M and G		
u	Volume of sent data	{1,2, …}[GB]	A_u=0.10[$/GB]	M_u=0.10[$/GB]	G_u=0.10[$/GB]
d	Volume of received data	{1,2, …}[GB]	A_d=0.15[$/GB]	M_d=0.15[$/GB]	G_d=0.12[$/GB]
b	Amount of data in the DB	{1,2, …}[GB m]	A_b=0.25[$/GB m]	M_b=0.15[$/GB m]	G_b=0.15[$/GB m]
q	Number of r/w transactions	{10,200}[action/query]	A_q=0[a]	M_q=0.01[$/10000]	
rb	Amount of data in RDB	{1,2, …}[h] or [GB m]	A_{rb}={0.105, 0.155}[$/h]	M_{rb}=(complex calc.)	
os	Operating system	{Linux, Win, …}[]	A_{os}={1, 1.41}		
e	Number of sent messages	{1,2, …}[]			G_e=0.01[$/100]

[a] A_q=0 only if simple DB is used, as presumed in this example.

Table 6
Data and results of TCO calculation for a 3-year period.

				App1	App2	App3
1	Complexity of algorithm a[]		**Required features**	1	6	10
2	Volume of sent data per month u[]		**of virtual applications**	10	100	500
3	Number of CRUD actions per query q[]			10	90	200
4	Number of queries per month x[ ]			10000	100000	500000
5	Annual data storage requirements b[GB]			100	1000	5000
6	Organisation time [s]			20	90	120
7	Number of used computers r[]		**Calculated value of**	2	7	11
8	Power of used computers p[]		**virtual applications**	1.7 GHz; 4 GB RAM		
9	Three-year costs of purchased computers $\sum C_{P(3)}$			**12201**	**42704**	**67107**
10	Three-year costs in case only computers are leased	Amazon $\sum C_{S,A(3)}$[$]		5385	18846	29616
		Microsoft $\sum C_{S,M(3)}$[$]		5391	18870	29652
		Google $\sum C_{S,G(3)}$[$]		4494	15726	24711
11	Three-year costs in case all the resources required for executing an application are leased	Amazon $C_{A(3)}$[$]		6375	28746	79116
		Microsoft $C_{M(3)}$[$]		6024	25206	61332
		Google $C_{G(3)}$[$]		5112	24711	55671

An example of calculation of the price of IT resources lease for three application types entitled App1, App2 and App3 is given in Table 6. Their features are defined by parameters in the first 6 rows of Table 6.

App1 is a low complexity application that would correspond to simple warehousing operations in a small business. App2 is more complex, comparable to executing an MRP algorithm in a medium-sized business, whereas App3 is the most complex and corresponds to executing an MRP algorithm on a complex structure of finished products in a medium-sized business.

The calculation results based on which it is possible to make a lease versus purchase decision concerning the required computing resources were obtained following the method steps 3–8 and are shown in rows 7–12 of the Table 6. The number of computers for each application type was calculated in accordance with (5).

Calculation results indicate that the cost of computer lease in all the three providers for all the three application types is 2.3–2.7 times lower than the cost incurred by using on-premise computers. The costs vary (row 10), wherein Google is the cheapest since it operates on the simplest platform. The computer component of the Amazon cloud for Linux is cheaper than that provided by Microsoft, while the prices quoted for Amazon Windows are fairly identical to Microsoft. Amazon's data storage is most expensive and it somewhat exceeds the features of its competitors (e.g. Hadoop support).

The costs of leasing computer time, which are comprised by coefficients indexed by ***r*** in formulas (13)–(15), are not the only costs incurred by the user in executing real applications. To these costs other components need to be added such as data transfer and storage, number of connexions, number of r/w operations, etc. The total costs that the user would incur over a three-year period lead us to conclude that: for application types App1 and App2 lease costs are lower than the costs of using on-premise resources although that difference tends to diminish (with the ratio amounting to 1.5–2.3). Whereas lease costs for the execution of a very complex application App3 in the Microsoft and Google cloud are almost identical to the costs of running the application on on-premise resources, leasing the required resources from Amazon would be more expensive than using on-premise computers.

The following general conclusion to be drawn from the aforementioned are: although leasing computer time in the cloud is on the whole a more favourable option for less demanding applications, total costs of executing very complex applications can be lower in case of using on-premise resources. The costs should therefore be carefully calculated for each particular case. We believe that the CCCE method can provide guidelines for cost calculation

to be appropriately conducted, thus enabling the user to select a more appropriate option.

4.10. Agree a fair price (10)

It is highly probable that a heavy user of IT resources will consider the possibility of using cloud. By implementing the CCCE method a user will calculate the amount of required computing and storage resources as well as the cost of using such resources on site or, alternatively, of leasing them in a cloud.

A heavy user will probably attempt to negotiate a lower price of using cloud resources with commercial cloud providers by requesting a quantity discount. The questions that arise in this context are: What is a fair price when cloud computing is concerned? What are the concessions that cloud providers can make in terms of reducing the price of IT resources while still making a decent profit? In discussing these questions the classical problem of honesty in trade, covered in detail in [11], needs to be brought up.

That problem belongs to the game theory as a part of economics. There are two sides in a trade relationship: the buyer (potential cloud computing user) and the vendor (commercial cloud provider). For each side's position there's a strong/weak value:

- A strong buyer wants to purchase goods at a low price and can afford to bargain in order to reduce the price. A weak buyer aims to purchase goods, even expensively.
- A strong vendor wants to sell goods at a high price and can afford to reject buyers that do not offer a sufficient price. A weak vendor aims to sell goods, even cheaply.

In cloud computing, the buyer's weak position is expressed by the difference between the calculated price of on-premise resources and the calculated theoretical price of leasing resources in a cloud: the buyer is not willing to pay the price of cloud resources that exceeds the price of equivalent on-site resources. At the same time, by using the CCCE method to calculate the cost price of leasing resources in a cloud the buyer can estimate their strong position, that is, determine the amount of reduction of the price of computing services that they can insist on in negotiating with the commercial cloud provider.

The vendor's (cloud provider's) weak position is primarily defined by the exploitation of computers in the data centre. The vendor's strong position is limited by the price of IT equipment available in the market that the buyer can independently configure into a privately-owned data centre.

The buyer tends not to reveal their position, in other words, to disclose the calculated costs of the required on-premise IT resources C_P and D_P prior to negotiating with the commercial cloud provider. By using the CCCE method the buyer can estimate the cost price of IT resources in a cloud C_S and D_S and start to negotiate the lease if $C_S < C_P$ and $D_S < D_P$ applies.

The vendor also tends not to reveal their weak position, although informations regarding their strong position are publicly known: these are the announced prices of services, available on official websites and cost calculators. The vendor's reductions cannot go below the cost price of services in a cloud C_S and D_S that represent their weak position. The vendor will therefore request prices C_X and D_X so that $C_X > C_S$ and $D_X > D_S$ applies.

A business relationship between the buyer of computing resources and the cloud provider will be established by a lease at prices C and D only if $C_S < C < C_P$ and $D_S < D < D_P$ applies.

If the aforementioned relations are placed in the context of Fig. 4 (with price approximations for clearer representation), which shows the cost of a CPU hour over the years discounted to NPV, we can argue that:

- The buyer's weak position is represented by curve ***s*** for $\eta = 0.60$, that is, by the price of CPU operation in the first year amounting to 0.14[$/h].
- The buyer's strong position is represented by curve ***s*** for $\eta = 0.95$, that is, by the price of CPU operation in the first year amounting to 0.10[$/h].
- The vendor's weak position is represented by curve ***L*** for $\eta = 0.95$, that is, by the price of CPU operation in the first year amounting to 0.08[$/h]. This is the lower margin at which profit can be generated, assuming that the cloud centre exploitation is relatively high.
- The vendor's strong position is not represented by curve ***L*** for $\eta = 0.6$. Instead, it is expressed by the commercial price of cloud services published on the cloud provider's website. The vendor cannot go beyond that price for the risk of losing business credibility. We shall assume that the vendor's strong position amounts to 0.12[$/h], which is in accordance with commercial prices of basic computer configurations in Microsoft and Amazon clouds.

A systematic overview of these relationships is given in Table 7, made for the purpose of analysing a symmetric mediation plan according to Myerson [11]. It is based on the assumption of equal probabilities that the buyer and the vendor will be found in their strong/weak position.

The probability of buying and selling cloud services is labelled ***q***. For ***q*** the probability constraint applies: $0 \le q \le 1$.

Table 7
Symmetric mediation plan for trading IT resources

			Buyer (cloud computing user)	
			Strong (0.5)	Weak (0.5)
			$C_P = 0.10$[$/h]	$C_P = 0.14$[$/h]
Vendor (cloud provider)	Strong (0.5)	$C_S = 0.12$[$/h]	0, no transaction	q, 0.14–y[$/h]
	Weak (0.5)	$C^S = 0.08$[$/h]	q, 0.08 + y[$/h]	1, 0.11[$/h]

The amount to be added to the cloud service price for the cloud provider to obtain certain profit is labelled ***y***. For ***y*** the participation constraint applies, as expressed in Table 7: $y \le 0.02$[$/h].

If the weak vendor acts honestly during quotation, their expected profit, according to the expected value calculation formula, will amount to:

$E(\text{profit\_i}) = 0.5 \cdot q \cdot (0.08 + y - 0.08) + 0.5 \cdot 1 \cdot (0.11 - 0.08)$, that is, $E(\text{profit\_i}) = 0.5 \cdot q \cdot y + 0.015$.

If the weak vendor is dishonest, the expected profit amounts to:

$E(\text{profit\_n}) = 0.5 \cdot 0 + 0.5 \cdot q \cdot (0.14 - y - 0.08)$, that is, $E(\text{profit\_n}) = -0.5q \cdot y + 0.03 \cdot q$.

A weak vendor will actually be honest if that is profitable for them, in other words, if the information constraint condition applies:

$E(\text{profit\_i}) \ge E(\text{profit\_n})$, that is, if inequation $0.5 \cdot q \cdot y + 0.015 \ge -0.5 \cdot q \cdot y + 0.03 \cdot q$ applies.

By rearranging the inequation we obtain:

$q \le \frac{0.015}{0.03 - y}$, that is $y \ge \frac{0.03 \cdot q - 0.015}{q}$.

The highest probability that buying and selling will occur is when profit $y \in [0.015, 0.02]$. Both margin sizes satisfy the condition $C_S < C < C_P$:

- When the buyer is weak and the vendor is strong, then $C_R \in [0.12, 0.125]$[$/h], $C_{RS} = 0.12$[$/h] and $C_{RP} = 0.14$[$/h].
- When the vendor is strong and the seller is weak, then $C_R \in [0.95, 0.10]$[$/h], $C_{RS} = 0.08$[$/h] and $C_{RP} = 0.10$[$/h].

If we observe the entire procedure from the weak buyer's perspective, we obtain their expected costs:

$E(\text{cost\_i}) = -0.5qy - 0.015$

$E(\text{cost\_n}) = 0.5qy - 0.03q$

It follows that the weak buyer will be honest if the information constraint condition applies:

$E(\text{cost\_i}) \le E(\text{cost\_n})$

$q \le \frac{0.015}{0.03 - y}$, that is, $y \ge \frac{0.03q - 0.015}{q}$.

The same terms of honest negotiating during buying and selling apply to the weak buyer and the weak vendor. This means that a business relationship between the provider and the user in this case will certainly ($q = 1$) be established at the lease cost $0.095 \le C \le 0.125$[$/h].

A symmetric mediation plan is an adequate way of calculating the fair price of cloud services since it allows for the margins of cloud service prices that both the provider and the user would agree upon to be estimated. In other words, it enables the vendor to engage in business and make a profit while keeping the service price sufficiently low for the user.

5. Production domain case study

Expression (5) for estimating the number of computers can be confirmed in production domain case studies other than described sample applications. One such case study is Croatian Information System of Higher Education Institutions (ISVU; www.isvu.hr). ISVU is complex information system which supports operations of several Croatian universities. Approximately 7000 University teachers and clerks, and more than 130,000 students on 95 faculties and high schools use ISVU on daily basis. There is on average 20,000 daily logins to ISVU, and up to 60,000 during peak demand periods (in July, September and October).

ISVU is run from servers at the University of Zagreb Computing Centre (SRCE; http://www.srce.unizg.hr/), main computing provider for academic community in Croatia. SRCE's data centre is highly virtualized and used in many computing, storage and network jobs (including HPC) for both academic community and private sector (mostly telecoms). For those reasons, SRCE's data centre has all the main features of small commercial computing cloud.

Resources on which ISVU is run have quite clear boundaries inside SRCE's cloud so it is relatively easy to analyse them. Present computing resources (***r***) engaged to run ISVU in SRCE's cloud are shown in Table 8

Amount of ISVU data transactions (CRUD) in year 2011 was a little bit more than 16 million. Monthly average is about 530,000, and monthly maximum during peak demand periods in July, September and October is about 2.5 million.

Overall monthly data traffic of ISVU is about 10 TB, but overwhelming majority of it, at least 90%, is backup transfer (for both databases and virtual computers). The remainder is, judging by ISVU client programs' features, divided equally between upload and download. From those data we can estimate upload ***u*** during peak demand periods as:

$$u = \frac{\textit{monthly data upload}[\textit{bytes}]}{\textit{maximum monthly transactions}} = \frac{0.5 \cdot 10^{12}[\textit{bytes}]}{2.5 \cdot 10^{6}} = 200[kB]$$

Table 8
ISVU computing resources (**r**).

Computers (processor & OS)	Type (physical F/ virtual V)	Amount	Number of CPU cores (r)	CPU cores usage (η)	Purpose
SPARC/Solaris	F	1	4	~10%	Informix DB server
x86/Linux	V	9	1	~40%	Apache web server ISVU app.
x86/Windows	V	3	1	~40%	Active Directory MS SQL

This value can also be checked by measuring data traffic per one user login:

$$u' = \frac{monthly\ data\ upload\,[bytes]}{number\ of\ \ monthly\ logins} = \frac{0.5\cdot 10^{12}\,[bytes]}{20\cdot 10^{3}\cdot 25} = 1000\,[kB]$$

Because $u'=5u$, during one session user makes about 5 transactions. That is consistent with average way of ISVU usage.

From those parameters we can calculate required computing power ($\boldsymbol{r}$) for ISVU according to (5) as follows:

- Response time is very low (as required for front office work), standing at $T=5$[s]
- Average data upload per usage case is relatively high standing at $\boldsymbol{u}=200$[kB]
- Average number of CRUD per transaction is about $\mathbf{q}=20$

So, by using expression (5) we get $\boldsymbol{r}=10.2$ that is reasonably close to the real number of CPU cores on which ISVU is actually run i.e. 16. The difference comes from relatively low CPU usage of ISVU's virtual computers that is much lower than in fully commercial computing clouds giving certain amount of redundancy in the system. If that usage was higher, number of CPU cores needed to run ISVU would certainly be lower and even closer to the calculated value. In the case of ISVU calculated values might be undertaken if, for example, the option of running it from some another cloud than SRCE's ever becomes interesting.

Another interesting consideration is how much ISVU would cost if it was run from fully commercial cloud. Next part of this case study is calculation of costs for leasing ISVU-like amount of IT services in Amazon cloud. As first, we will supplement ISVU parameters from first part of this case study with yearly data storage needs that are shown in Table 9.

Data storage requirements for items 1, 2 and 3 in Table 9 can only be met by "large" relational database management system (RDBMS) such as Oracle Database or Microsoft SQL Server (MSSQL). Such services are provided by Amazon RDS in which one can choose between those two RDBMS (and MySQL). We consider Oracle DB as more adequate replacement for ISVU's Informix database than MSSQL because of OS-independency.

Regarding computing resources, we'll calculate with real number of ISVU CPU cores ($r=16$), two on Windows, and the others on Linux. Of those 16 CPU cores, 5 are used for running RDBMS, and others for computing i.e. running the application. We'll assume that all leased resources are of classical, on-demand type, situated in Irish data centre (and with "bring-your-own-licence" model for leased RDBMS). Other assumptions are as follows:

Table 9
ISVU's yearly data storage needs.

	Purpose & type	Capacity [GB]
1.	Transactional (OLTP) database—Informix	105
2.	Data warehouse (OLAP)—MSSQL	25
3.	Database logs	150
4.	Backup (databases and virtual computers)	50000

- 9 computing cores are provided by 9 Amazon EC2 Large Instances, 8 on Linux and 1 on Windows, for hourly price of $0.34 and $0.46, respectively.
- 4 cores needed for running OLTP database are provided by one Amazon RDS Double Extra Large DB instance running Oracle Database, for hourly price of $1.315.
- 1 core for running OLAP database is provided by one Amazon RDS Large DB instance running MSSQL, for hourly price of $0.568.

Amazon charges only outgoing network traffic from both EC2 and RDS (and most of other of its services) at rate of $0.12 per GB (for monthly network traffic between 1 [GB/month] and 10 [TB/month]). So, outgoing network traffic for OLTP and OLAP is 500 [GB/month] for EC2 and also 500 [GB/month] for RDS.

Situation is interesting when one considers backup traffic which accounts for 9 [TB/month]. We don't have the analysis of that traffic, but it can be safely assumed that overwhelming majority of it is outgoing network traffic toward backup points, also charged at the rate of $0.12 per GB. It should be noted that moving virtual machines requires usage of Amazon EBS or similar services, but during negligible time periods and therefore with negligible cost.

Requirements for item 4 in Table 9—backup storage can be met by much simpler storage system than RDBMS, preferably one specialized for archiving. Amazon Glacier is one such system with price of $0.011 [GB · month].

So, final calculation of yearly costs for ISVU-like information system in Amazon cloud is:

$$\begin{aligned} C_{ISVU} = {} & 8760\,[h]\left(9\cdot 0.34\left[\frac{\$}{h}\right]+1\cdot 0.46\left[\frac{\$}{h}\right]\right. \\ & \left.+1\cdot 1.315\left[\frac{\$}{h}\right]+1\sqrt{2}0.568\left[\frac{\$}{h}\right)\right) \\ & +12\,[month]\,0.12\left[\frac{\$}{GB}\right]\left(1\cdot 500\left[\frac{GB}{month}\right]\right. \\ & \left.+1\cdot 500\left[\frac{GB}{month}\right]+9000\left[\frac{GB}{month}\right]\right) \\ & +12\,[month]\,0.011\left[\frac{\$}{GB\cdot month}\right]50000[GB] \end{aligned}$$

C=$68,330.28

Total yearly cost of ISVU-like system run on Amazon would be (as calculated from CCCE) about $68,330.28. Real yearly cost of ISVU with included value-added tax of 25% is about $96,000, so without VAT it would be about $77,000. It is somewhat more expensive than running in large commercial cloud would be, but there are additional security and legal benefits in having it physically under domestic jurisdiction. These aspects are not covered by CCCE and it is up to every user to analyse them for a given type of application.

6. Conclusion

In this paper we analysed the criteria and defined the procedures that a company management can use to make a quantified decision regarding the technical sustainability and economic cost effectiveness of purchasing the

privately-owned computing infrastructure as opposed to leasing cloud computing resources. In doing so, it is assumed that all the security and legal aspects of using cloud computing have been considered.

Although all cloud providers offer online calculators for estimating the costs of their services, a potential cloud computing user first needs to calculate the amount of the required cloud resources, in other words, what they should enter into the cloud services cost calculator. The number of computers $\boldsymbol{r}$ required for executing an application in acceptable time T_A definitely represents an important item among the aforementioned resources. In this paper we showed how measuring application execution time in the cloud can help to determine the number of required computers in a more exact way. The number of computers $\boldsymbol{r}$ is the function of several variables, whose effect was determined by processing the results of the fractional factorial design applied in measuring sample application execution time. Processing the measurement results is fairly complex and needs to be conducted separately for each application that is suitable for execution within the cloud.

Drawing on our previous research, we proposed a resolution of another major dilemma in deciding whether to use cloud computing: how to determine with certainty whether leasing cloud IT resources is cheaper than purchasing an adequate amount of these resources and using them in a privately owned data centre? This dilemma is resolved by comparing the sum of costs of on-premise computing and storage resources against those in a commercial computing cloud. Considering that different providers use a variety of calculation plans for their services, in this paper we devised formulas for comprehensive calculation of usage costs based on the number of computers $\boldsymbol{r}$ and parameters that describe the applications. Formula validity was verified on an example of the execution of three virtual applications of various complexities with three different cloud providers. It was shown that the total costs of executing highly complex applications in the cloud can be higher than those incurred by executing them in a private data centre, while in the case of less complex application leasing proved to be more favourable.

The decision to buy or lease the required resources will surely be affected by the perception of the fairness of the cloud service lease price. We demonstrated the manner in which a symmetric mediation plan can be designed that would result in the highest probability of computing services lease at the price that would meet the expectations of both the cloud provider (vendor) and the user (buyer). The implementation of such an approach can greatly contribute to cloud computer usage and its further development.

All the described procedures of calculating the cost effectiveness of lease are integrated in the unique CCCE method, which will enable cloud computing users to repeatedly apply this procedure in accordance with their needs. The validity and executability of particular steps of the method was verified on several business cases and three sample applications. Although the conclusions arising from each method application can differ from those presented in this paper, the methodological procedure will remain unchanged.

It should also be mentioned that (in addition to described case studies of computer number estimation) CCCE method has been checked with few experts from telecom industry. They consider CCCE method described in this paper as complete, although their company uses much shortened version, calculating only with hardware, software and data centre accommodation costs in the process of forming prices for their commercial and public cloud services.

References

[1] C. Babcock, Management Strategies for the Cloud Revolution: How Cloud Computing is Transforming Business and Why You Can't Afford to be Left Behind, McGraw-Hill, Columbus, Ohio, 2010.

[2] P. Berger, R. Maurer, Experimental Design with Applications in Management, Engineering and the Science, Thompson Learning, London, 2002.

[3] J. Caceres, L. Vaquero, L. Rodero-Merino, A. Polo, J. Hierro, Service scalability over the cloud, in: B. Furht, A. Escalante (Eds.), Handbook of Cloud Computing, Springer, New York, 2010., pp. 357–377.

[4] J. Carolan, S. Gaede, Introduction to cloud computing architecture, white paper, Sun Microsystems, 2009.

[5] A. Greenberg, J. Hamilton, D.A. Maltz, P. Patel, The cost of cloud: research problems and data center networks, white paper, Microsoft Research, Redmont, Washington, 2009.

[6] R. Harms, M. Yamartino, The economics of the cloud, white paper, Microsoft Research, Redmont, Washington, 2010.

[7] ISACA, Cloud computing: business benefits with security, governance and assurance perspectives, white paper, ISACA, Rolling Meadows, Illinois, 2009.

[8] R. Jennings, Cloud Computing with the Windows Azure Platform, Wiley Publishing, Indianapolis, Indiana, 2009.

[9] D.S. Linthicum, Cloud Computing and SOA Convergence in Your Business, Addison-Wesley Professional, Reading, Massachusetts, 2009.

[10] P. Mell, T. Grance, Effectively and securely using the cloud computing paradigm, NIST, 2009. 〈http://www.nist.gov/itl/cloud/〉. (accessed 13 02 11).

[11] R.B. Myerson, Perspectives on mechanism design in economic theory, Nobel Prise Lecture, 2007. 〈http://home.uchicago.edu/rmyerson/research/nobel.pdf〉. (accessed 09 04 12.

[12] M. Naghshineh, R. Ratnaparkhi, D. Dillenberger, J.R. Doran, C. Dorai, L. Anderson, G. Pacifici, J.L. Snowdown, A. Azagury, M. Venderwiele, Y. Wolfstahl, IBM Journal of Research and Development 53 (4) (2009) 1.1–1.10. IBM Research Division Cloud Computing Initiative.

[13] G. Reese, Cloud Application Architectures: Building Applications and Infrastructure in the Cloud, O'Reilly Media, Sebastopol, California, 2009.

[14] J. Rhoton, Cloud Computing Explained, Recursive Press, Tunbridge Wells, 2010.

[15] J.W. Rittinghouse, J.F. Ransome, Cloud Computing-Implementation, Management and Security, CRC Press, Taylor&Francis Group, New York, 2010.

[16] V. Sawyer, Calculating total power requirements for data centers, APC, white paper 3, 1–16, 2004. 〈http://dcmms.com/whitepapers/2.pdf〉, (accessed 13 02 13).

[17] L. Spacek, Description of the collection of facial images, 2008. 〈http://cswww.essex.ac.uk/mv/allfaces/index.html〉. (accessed 13 02 11).

[18] M. Vouk, A. Rindos, S. Averitt, J. Bass, M. Bugaev, A. PEeler, H. Schaffer, E. Sills, S. Stein, J. Thompson, M. Valenzisi, Using VCL technology to implement distributed reconfigurable data centers and computational services for educational institutions, IBM Journal of Research and Development 53 (4) (2009) 2.1–2.18.

[19] E. Walker, The real cost of a CPU hour, IEEE Computer (2009) 35–41.

[20] L. Wang, G. Von Laszewski, Cloud Computing—A Perspective Study, Rochester Institute of Technology, Rochester, New York, 2008.

[21] E. Walker, W. Brisken, J. Romney, To lease or not to lease from storage clouds, IEEE Computer 43 (4) (2010) 44–50.

[22] L. Xinhui, L. Ying, L. Tiancheng, Q. Jie, W. Fengchun, The method and tool of cost analysis for cloud computing, in: Proceedings of the IEEE International Conference on Cloud Computing (CLOUD), 2009, Bangalore, India, pp. 93–100.