An Adaptive Real Time Mechanism For IaaS Cloud Provider Selection based on QoE aspects

Mohamed Souidi, Sami Souihi, Said Hocinei and Abdelhamid Mellouk
Image, Signal and Intelligent Systems Laboratory (LISSI) - Dept. of Networks and Telecoms, IUT C/V
University of Paris Est Creteil (UPEC), FRANCE
Email: ms.souidi@gmail.com, sami.souihi@u-pec.fr, hoceini@u-pec.fr and mellouk@u-pec.fr

Abstract—Traditionally, companies host their own services, platforms and infrastructures on their own servers. This policy results in high costs in terms of material and human resources. It may also be inadequate to the real needs of the company. In this context, one solution is to use cloud computing to outsource their services. The latter is defined by making available to the customer high-performance servers and high bandwidth. The cloud is also defined by renting software and hardware infrastructure to customers according to their needs. Cloud computing is made possible by the improvement of computer networks infrastructures. Indeed, broadband connections have reduced latency and thus enabled the use of remote resources. The success of cloud computing has led to a significant increase in the providers number offering many and varied cloud services. While the access to these services is made possible through a simple subscription, no technique is currently available to select the cloud provider that best fits their needs. Selecting a provider is an optimization problem that has been studied in several areas. Given the large number of parameters and actors in the cloud, this problem is known as NP-complete one.

In this work, we propose a new developed platform which plays the role of a broker between clients and cloud providers. Based on a set of benchmark tasks on provider services, it performs an adaptive cloud provider selection in accordance with the client needs. The experimental results show that the proposed approach gives benefits to subscribers in terms of QoE.

I. INTRODUCTION

The computer technology has evolved from central architecture (where terminals are connected to one single main frame), to the decentralized system (where each machine has its own computing processor). That kind of architecture was hard to maintain and expensive for companies. The new trend based on cloud computing aims to re-centralize this architecture.

The cloud computing means to move away from the traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the new OPEX model (use a shared cloud infrastructure and pay as one uses it). The cloud computing (CC) offers to enterprises a set of multiple services. These services can be categorized into 3 types: SaaS (Software as a Service), PaaS (Platform as a Service) and IaaS (Infrastructure as a Service): i) SaaS: applications installed, configured, ready to use and accessible remotely, for example “Gmail”. ii) PaaS: platforms hosted by the provider and accessible remotely such as a data base management system or a development platform. iii) IaaS: pre-configured hardware remotely controlled via Internet.

The number of cloud providers is significant which makes it difficult for customers to choose the right one. Although being of minor importance for personal use, the choice of the provider is an extremely important issue for companies. Indeed, nowadays, companies realize that the effective management of their purchases can be a competitive advantage. Supplier selection becomes a strategic decision that has a direct impact on the overall performance of the entire company. Selecting a cloud-provider is by nature a critical task. Moreover, the diversity of performance measurement criteria further compounds the difficulty of this task. In this article, we present an adaptive approach for cloud computing provider selection depending on the subscriber needs, a part of quality of experience (QoE) paradigm.

This paper is organized as follows: i) The next section shows the cloud computing key performance indicators. ii) Section 3 describes the developed benchmark platform that will be used to measure the cloud computing provider performance. iii) Then, section 4 explains the proposed adaptive approach for cloud computing selection depending on the measurement retrieving by the benchmark platform. iv) Finally, the last section summarizes the evaluation results and demonstrates the performance of the proposed adaptive approach.

II. CLOUD COMPUTING KPI

Cloud computing has helped enterprises to overcome the problem of accessing to machines with high performance, while neglecting the hard maintenance of the hardware [1]. In fact, cloud computing provides the enterprises with a ready-to-use platform, while making abstraction of the hardware used. It’s reduces the difficulties due to the hardware development and maintenance. A cloud provider uses the virtualization of hardware to insure the decoupling of software and thus create an abstraction level to enterprises. The direct manipulation of the system installed upon a physical machine is not used anymore. This virtualization offers too many benefits to services or computer users. These benefits are: i) Fault tolerance: In case of breaking of a server, the virtual machine (VM) migrates to another server in seconds. ii) Security: Each virtual server has a specific task and the ports unused for each server will be locked. iii) Quick maintenance and migration from one server to another: decoupling the software from the hardware. All these technologies contributed to the success of cloud computing and led to the emergence of several cloud providers. Next section will explain the technologies behind the cloud and will define the metrics to evaluate it.
A. Definition of cloud computing

As defined by NIST, cloud computing is a set of software and hardware resources remotely controlled via the Web. As explained in the introduction, the cloud is subdivided into three classes. The main class used by enterprises is the IaaS which is the basic layer that offers hardware to the cloud subscribers [2]. Cloud computing is based on the principle of virtualization of machines. That explains the total separation and decoupling of the hardware from the software. This technique will guarantee availability, scalability and reliability.

The availability is the most important factor in cloud computing. In fact, the quality of service is directly related to the availability [3]. This latter is the engagement of the provider to ensure a rate of availability in relation to the usage time. This means that, when a server crashes or breaks, it takes only a few seconds to merge the virtual machine VM to another physical server. The scalability in cloud computing is the possibility of changing characteristics of virtual machines as a function of the real use of the service. In fact, the scalability is the ability of platform evolution in relation to the number of users. This factor highlights the performance of the service. In a practical way, whenever the client asks for more performance, the provider adds more VM to the cluster. The reliability of cloud computing is also the engagement of providing a reliable service that tolerates the failure and overcomes the abuse of the service. Due to heterogeneous hardware, software, equipment and services, the reliability becomes a hard task and a key factor in installing a cloud service [4].

The heterogeneity of hardware is the fact that the cloud providers are not restricted to one supplier or one technology while preparing their solutions. In fact, in one data center, there are different kinds of equipments. So, the provider should be able to connect all these equipments together in order to see it as one logical solution. The heterogeneity of operating systems is represented in the systems and applications used on these heterogeneous equipments. Providers use different operating systems (OS) on top of virtual machines and need to communicate them together.

The final feature is networking. This feature is the most sensitive one because the quality of service (QoS) offered by the cloud providers depends on the quality of service of the network operators. The operators do pick up the right places for their data centers in order to offer the best quality of service QoS [5].

B. Cloud computing KPI

Through the literature study and the study of the cloud computing specifications, we specify the cloud computing key performance indicators, namely the availability, the security, the reliability, the network performance, the scalability, the instance performance and the cost. These are the major factors affecting the service performances. Because of the distant aspect of the cloud services, the network will have a great impact on the cloud service performance [6]. This present work will focus on network performance, reliability and instance performance.

The major factor that impacts the services is the network performance. Providers set up their data centers in places where they are spending less (the furthest the data centres are from cities, the lowest the costs spent for their set-up). As a matter of fact, the distance will have an impact on network performance and quality of services. The cloud clients are from all over the globe and thus the network performance will have an impact on quality of service [7]. The proposed work will focus on the availability, the instance performance, the network performance and the cost of VM.

The provider selection depends on two factors: the provider performance and the end-user perception [8]. As previously explained, the provider performance depends on three major parameters, namely localization, time and VM performance. The end-user perception, or quality of experience (QoE), is a subjective measure of the adequacy of a service compared to the user expectation. This latter is explicitated as function of different KPI chooses by users.

Fig. 1 outlines the metrics to measure cloud computing providers. It shows the metrics related to the provider and the metrics related to the client as well as the relation between the client metrics and the KPI of cloud computing provider. The cost means the cost of using the service, which is measured in hour utilization. It is expressed in number of virtual machines or processor using rate. The availability is a rate that indicates the availability of the service in a fixed period of time. The reliability will be measured by a series of benchmark executed in the virtual machines. This benchmark task executes tests to measure the processing power, the I/O to data speed, speed accessing to memory. Network performance is represented by measuring bandwidth and latency. These two parameters are measured from an end-user perspective. In order to have an end-user perspective, the network measures are executed on client machines to take into account localisation and time parameters.

The main contribution of this paper is to measure network capability from an end-user perspective. The network capability will change from a country to another and the quality of cloud service will mostly depend on that fact. In order to get a concrete idea of the real measures of the network performance from an end user perspective, we thought of using a system that will offer the possibility of making different tests from
different geographical area. To achieve this goal, we use a planet-lab [9]; it consists of a set of servers that are set-up around the globe. The purpose is to set a series of benchmarks to a different cloud provider from the servers of planet-lab.

When opting for cloud computing, the major expenses of the company are linked to the computing and network performances. These expensive features define the quality of the offered service. They also have different priority levels inside the enterprise depending on the enterprise needs. As a matter of fact, the parameters that will be constantly measured for the purpose of evaluating the cloud services are the CPU, the bandwidth and the latency.

III. CLOUD COMPUTING PROVIDER SELECTION

Previous sections defined the cloud computing and the criteria used to measure its providers performance. In order to achieve the selection of cloud computing Provider (CCP), this section will explain the problematic of choosing the provider. It will also explain the selection algorithms. The correlated KPI of CC makes the problem of selection in the case of CCP much more complicated. Indeed, the traditional selection algorithm does not fulfill the CCP selection task. Next section will propose a new approach for CCP selection. That approach is a real time mechanism based on real time performance measures.

A. Selection problem

The selection can be discussed in all fields; whenever there is a choice to make, the problem of selection is raised. Different algorithms are found in the literature to address the selection problem [10]. These algorithms can be subdivided into two categories: the classifications models, and the deduction models. i) The first one are used to classify the sets of inputs from their scores. They are mainly used when the input is important in order to reduce the number of sets. In practice, these categories are used in case of intolerable criteria. For instance, if one of the inputs does not meet the criteria, it will be eliminated from the set of inputs [11]. ii) On the other hand, the deduction models are a bench of mathematical algorithms, artificial algorithms and statistical algorithms. The mathematical models are used in case of reduced dimensions. Linear programming is one of these mathematical models. This algorithm gives weights to inputs depending on their importance, then optimizes a subjective function in order to obtain the optimal solution. This second category includes also artificial intelligence programming such as neural networks and expert systems. These algorithms are used in case of big data set [12].

The cloud computing selection operator takes part of the multi-criteria selection process. It needs to be evaluated from an end-user perspective. The problematic of CCP selection is much more complex than a simple service provider selection. This complexity is due to the diversity of criteria and the variation of the results. The KPI used as input to the algorithm are not on the same set. They are divided into objective criteria (data center localization, time, VM performance), subjective criteria (flexibility of the agreement between the client and provider, the diversity of services price). The example of billing in cloud providers shows the complexity of comparing the services of different providers, Google count the price per number of CPU, Amazon count per number virtual machines. This example focuses on the computing part and demonstrates the difficulty of comparing between these providers. Additional parts include the correlation of the services and the client needs. However, it is complicated to take all these parts into account in one single algorithm while trying to systematically select the appropriate CCP [13].

B. Selection algorithms

The selection of CCP depends on the provider quality of services, but it also depends on end-user parameters such as users needs, localization and specifications. Indeed, the proposed algorithm is based on a cost function optimization. This function takes as input the parameters measured by the benchmark and defines the user expectation by weighting them. For example, if the client need more real time capability, he will increase the latency parameter weight. By default, the algorithm give the same weight to all considered parameters.

The current work proposes two algorithms: i) the first one selects service provider according to the benchmark test and client specifications. This algorithm consists in a simple cloud provider selection (SCPS), ii) The second algorithm, called adaptive cloud provider selection (ACPS), compares the cloud provider performance to client specifications and selects the best provider at each time.

The first proposed algorithm follows the reasoning below, starting from the measures that depend on the provider to the client needs [14]. The algorithm runs as follows: It uses the user localization to define the measures to use in order to define the values that match the user country. For example if the end-user is from the USA, the algorithm will use the measures that come from USA servers.

The algorithm 1 represents the SCPS approach. It’s a linear programming model that maximize the cost function used to select the best CCP. This latter takes as input the defined criteria and then aggregates them. In order to provide the end-user with the best provider who matches his needs, we set different function weights on the basis of the user preferences. In case the client does not specify his needs or specifications, the cost function will be predefined (same weight for all parameters).

Algorithm 1 SCPS algorithm

\[
loc(U, P) = \min_{i=0 \to n} d(ULoc - PLoc_i)
\]

\[
bw = \frac{\bar{Bw} + \text{std}(Bw)}{\|P\|_x}
\]

\[
l_t = \frac{\|Lt\|_x}{\|P\|_x + \text{std}(Lt)}
\]

\[
p = \frac{\|P\|_x}{\|P\|_x + \text{std}(P)}
\]

\[
pe = \frac{\bar{Pc} + \text{std}(Pc)}{\|P\|_x}
\]

\[
\text{sel}(U, P) \rightarrow \max_{i=0 \to n} \left(x_1 bw_i + x_2 l_t_i + x_3 p_i + x_4 pe_i\right)
\]

This equation displays the solution of the cloud provider selection. The \(x_i\) terms in the formula are the weights taken from the user specifications. For example, if the user is
searching for a service where he uses the most file transfer, he will give a big importance to the bandwidth. On the other hand, if the user is dealing with real time communications, he will give a big importance to the latency. We let the algorithm be open to all user specifications in order to respond to his real needs. The idea of the algorithm is to select the cloud provider that gives the highest score. These scores are measured by aggregating the key performance indicator that defines the bandwidth (bw), the latency (lt), the cost (p) and the virtual machines local performance (pe). Indeed, the local performance of the virtual machines is a series of benchmark tests that measure the processing power, the input/output speed and the process creation speed. The algorithm uses an aggregation of the average and standard deviation of each value in order to smooth the values.

Once the provider is chosen for the particular end user, the proposed approach continues the measurements in order to provide the end-user with the best choice. This continuous measurement builds a knowledge plane that will recognize the system in different time slots and places. To give the user the best choice of the cloud provider that best responds to his needs, we use an algorithm that detects the decrease of the quality of the offered services. If there is another cloud provider that gives more benefits to the subscriber, the algorithm will launch a notification to start a new selection depending on the new data. This algorithm is based on Page Hinkley approach [15]. It analyses the data changing in real time. This sequential analysis technique allows for an efficient detection of changes in the normal behaviour of a statistical process (break point) without any prior hypothesis. This test estimates a cumulative variable $X_t$ defined as the cumulative difference between the observed values and their average up to the current time.

The sequence of the algorithm 2 is based on measuring the current average of score. The score is calculated by the adaptive cloud provider selection (ACPS) algorithm. The real need for a real time average is to have a constant way of the network variation. It often compares the new score with the current average. The algorithm of Page hinkley will detect if there is a break point between the new score and the current average. The sigma represents the tolerated difference; if the algorithm detects a score under sigma, it will launch the selection process.

### Algorithm 2 ACPS algorithm

```
algorithm : allowed change value
while $t \geq 0$ do
    $X_t \leftarrow \sum_{t=1}^{T} (v_t - \bar{v}_t)$
    $d_t \leftarrow \sum_{t=1}^{T} d_t$
    $m_t \leftarrow \min(X_t, t = 1..T)$
    $dPHT \leftarrow X_t - m_t$
    if $dPHT \geq \sigma$ then
        launch selection procedure (SCPS algorithm)
    end if
end while
```

As previously explained, the algorithm gives an adaptive mechanism to select the cloud provider. The algorithm adapts its features according to the client preferences. It takes dynamic decisions depending on the network, the instance performance and the subscription cost. The continuous aspect of the platform through the measures gives the data plane a real time aspect. Since the the proposed algorithm takes decisions that depend on data plane, it gives the client an adaptive real time mechanism for cloud computing selection.

### IV. Experimental results

The first experience is a series of network benchmark. We developed a javascript code that measures the network performance from a client perspective. We developed also a code in cloud computing side. Indeed, the client side script measures performance from the client side. It sends http request to cloud count (to the script installed on cloud computing side). Then, the network performance is measured by the client side script using the reply of the cloud computing provider.

To achieve our purpose, we used a set of two hundred servers distributed around the globe from Japan, Australia to America through Europe. We deployed a piece of script on planet-lab servers. This scripts measures latency, bandwidth and localization from a client perspective (which are the servers in our case). These scripts start to measure network capabilities each 10 minutes, for 24 hours. These measures define a knowledge plane that will give a global idea of the network behaviour. While analysing these values, we tried to define a trend in the network evolution, to see when the bandwidth is the more consumed, when the latency is the less, and in which countries the traffic is the more intense.

In order to measure the cloud provider performance, we create instances on several providers: Amazon, HP, Microsoft Azure and Numergy. These providers are the most effective providers at the current market place. From planet-lab servers, we launch the series of tests on these virtual machines of cloud providers. In this section, we will explain the experimental results retrieved from the built platform. The experience took 24h and the platform made use of 200 servers from the planet-lab system.

#### A. Benchmark

A series of benchmark task was performed in order to evaluate the cloud provider performance. The KPI of CC was explained in the previous section. Now, we will explain the platform developed to measure the CCP performance.

We developed a central module that represents the data base and the head quarter. The javascript code deployed on planet-lab servers is a script in cloud computing data centres. These scripts on plant-lab servers will execute a series of requests entered to the scripts deployed on cloud computing servers. The obtained results will be sent to the main application for processing purposes. The data base will have values of CCP performance from a client side. These values are latency, bandwidth, localization and some other values like the client operator and operating system. Another script was developed and deployed in cloud provider instances. These scripts measure the instance performance and send the result to the centralized data base.

#### B. Results

The first results that we obtained are the network performance measures. The fig.2 explains the latency variations
(shown in milliseconds) and the fig.3 shows the bandwidth variation (shown in kbits/s). However, given the big density of countries, the figure is split into two parts.

These figures depict an average of twenty four hour of measuring. The results vary as a function of slots of time. In this work, the variance is represented by the standard deviation which shows the real extreme values measured throughout the day. These graphs prove that no conclusion can be drawn about the provider performance on latency or bandwidth because of the dynamical aspect of the network. For that reason, the proposed algorithm gives a practical solution that provides an adaptive real time answer to the client request.

The cost of one instance changes from a provider to another; each one of them gives different specifications in order to improve and respond to its client needs. The proposed platform makes a series of benchmark tasks. These tasks measure the network performance through the latency and the bandwidth and also measure the instance performance.

A script has been deployed into the VM’s on cloud. The current script executes a series of benchmarks to measure the instance performance before sending them to a central data base. The deployed script measures the I/O to database rate, the processing power rate, the time to launch a new VM and the storage speed. This script gives a score to each provider and that score does not change with time as it has almost the same rate.

The table I displays the results obtained from the benchmark task executed in instances where we subscribed. It shows the cloud providers and their respective calculated scores. In order to achieve these series of tests, we made use of virtual machines from the following some; Amazon, HP, Google and

<table>
<thead>
<tr>
<th>Provider</th>
<th>Amazon</th>
<th>HP</th>
<th>Microsoft</th>
<th>Numergy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark score</td>
<td>71.5</td>
<td>1398.7</td>
<td>762.0</td>
<td>1724.4</td>
</tr>
</tbody>
</table>

**Fig. 3. Bandwidth average and standard deviation per country**

**TABLE I. COMPUTING PERFORMANCE**

![Bandwidth average and standard deviation per country](image)

![Latency average and standard deviation per country](image)
Numergy. We used the small instances with no special demand on bandwidth. The table II explains the cost of each virtual machine per month for each cloud provider (in euros).

C. Evaluation of CP selection mechanism

This section explains and demonstrates the benefits of the use of an adaptive algorithm in cloud selection. It compares the adaptive selection approach and the simple selection approach. To illustrate the difference, we take an example of a client with some specifications: 50% of his business is done in USA, 30% in France and 20% in others countries. According to these specifications, our approaches give weight to the measured data as follows: 0.5 for measured data from the United States, 0.2 for measured data from France and 0.3 for other data. The experience started at midnight. First the platform starts by saving the data in order to have some experience from which the algorithm can make a decision. Then, the selection algorithm picks up a provider, the simple approach goes on with the same provider. The dynamic approach executes the selection algorithm each time a break point is detected.

![Fig. 4. Adaptive Provider Selection](image)

The figure 4 represents this variation through time and compares the score calculated by the ACPS and the SCPS. Indeed, each time a break point is detected, the selection algorithm calculates the new score from the current benchmark results and if there is a benefit in changing to another CCP, the algorithm switches the client to the new cloud provider. The figure shows that the results start at 4 o’clock. In fact, the measures start at midnight, then after 4 hours of measuring, we estimate that the algorithm has enough experience and knowledge to make decision depending on past experiences. The figure shows that the dynamic approach gives the best choice and the best score through time and in worse case it gives the same results as the simple selection algorithm. In fact, the beginning of the experience shows that the two methods had the same results. This was explained by the fact that both of them chose the best provider. The SCPS continues with that provider, but the ACPS executes the selection algorithm a second time if a decrease in quality is detected.

These results are made in 24h. However, we believe that the difference will be much bigger as the time increases. As a matter of fact, the ACPS has a great impact on the enterprise choice of CCP as it provides them with the best performance and the cheapest services. Given the difference in scores between the two approaches, we can insist on the fact that changing the cloud provider may save money to customers.

V. Discussion

This proposal focuses on the definition of cloud services and measures the service performance from an end point perspective. These measures are taken from different points around the globe. The second step consisted in the evaluation of these measures. We used a continuous method for benchmark score retrieving and the Page Hinkley algorithm that ensures the watchdog job for the variation of quality of services depending on the client exigency. The real time cloud provider selection is interesting in several cases especially when there is fluctuating demand together with a small change in the data (which greatly reduces the cost of migration). For some use cases, it is obvious that we must take into account the cost of migration between cloud providers. In our case, we did not consider this factor for two reasons:

- The cost of migration depends on the nature and the quantity of the hosted data
- This cost can be greatly reduced by maintaining multiple instances hosted on different providers and synchronized with one another. The extra cost is not necessarily important because unused instances require minimal resources.

VI. Conclusion

The selection of cloud provider is classified as a problem of multi choice selection because it deals with complex structure and correlated indicators. This paper proposes a new adaptive approach to solve the problem of cloud provider selection. Initially, we fix and measure the KPI from an end-user perspective. Our main focus when setting these parameters is the quality of experience. Once the KPI is defined, the focus is placed towards the definition of the knowledge plane that has the ability and knowledge to make decisions. However, applying this approach solely in the case of Internet network supposes that the network is stationary. Finally, more attention is paid to the dynamic status of the network as well as the network switching. In this context, the Page Hinkley algorithm is implemented to track the network changes and detect the break point of nonconformity between the client needs and the specifications. At this point, we only deal with a dynamic factor without trying to modulate it in a formal way. In future work, the focus will be on this part of network nonconformity and non-linearity. We will try to implement an algorithm that predicts the network evolution as a function of time from previous knowledge. Under such conditions, the algorithm is expected to choose the cloud provider that best responds to

<table>
<thead>
<tr>
<th>Provider</th>
<th>Amazon</th>
<th>HP</th>
<th>Microsoft</th>
<th>Numergy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly cost in euros</td>
<td>43.20</td>
<td>49.90</td>
<td>43.80</td>
<td>35.77</td>
</tr>
</tbody>
</table>

**TABLE II. INSTANCES COSTS**
the subscriber needs. The aim is to improve performance of the offered service and reduce its cost.

REFERENCES


