International Journal of Information Technology & Decision Making
© World Scientific Publishing Company

A METHODOLOGY FOR ECONOMIC EVALUATION OF CLOUD-BASED WEB APPLICATIONS

JOSEP DOMENECH, RAUL PEÑA-ORTIZ, JOSE A. GIL, ANA PONT
Universitat Politècnica de València, Camí de Vera s/n
46022 Valencia, Spain

Received 20 March 2015

Cloud technology is an attractive infrastructure solution to optimize the scalability and performance of web applications. The workload of these applications typically fluctuates between peak and valley loads and sometimes in an unpredictable way. Cloud systems can easily deal with this fluctuation because they provide customers with an almost unlimited on-demand infrastructure capacity using a pay-per-use model, which enables Internet-based companies to pay for the actual consumption instead of peak capacity. In this paradigm, this paper links the business model of an Internet-based company to the performance evaluation of the infrastructure. To this end, the paper develops a new methodology for assessing the costs and benefits of implementing web-based applications in the cloud. Traditional performance models and indexes related to usage of the main system resources (such as processor, memory, storage, and bandwidth) have been reformulated to include new metrics (such as customer losses and service costs) that are useful for business managers.

Additionally, the proposed methodology has been illustrated with a case study of a typical e-commerce scenario. Experimental results show that the proposed metrics enable Internet-based companies to estimate the cost of adopting a particular cloud configuration more accurately in terms of the infrastructure cost and the cost of losing customers due to performance degradation. Consequently, the methodology can be a useful tool to assess the feasibility of business plans.

Keywords: Cloud computing; performance analysis; economic evaluation.

1. Introduction

Global Internet usage has grown by 741% since 2001, and currently reaches three billion users, which is exactly ten times more than 14 years ago according to Internet World Stats[1]. New online applications and services are partially responsible for this huge growth. As an illustrative example, Facebook had more users in 2012 than the entire Internet in 2004, the year when the social network was founded[2]. The related business activity has also grown accordingly. For instance, e-commerce sales reached $1.298 trillion worldwide in 2012, while eMarketer[3] estimates a compound annual growth rate in Internet sales of 13.8%. At the same time, The New York Times declared 2012 as the dawn of the “Age of Big Data” because of the massive amounts of information generated every day[4].

In such a scenario, the infrastructure provided by the cloud seems to be the best solution for storing big data and hosting e-services and applications with a certain level of reliability and capacity[5]. The National Institute of Standards and Technology defines cloud computing (CC) as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction[6]. Moreover, the term CC has evolved to refer to a wide offer, and a new terminology is being used by vendors and service providers, such as: infrastructure as a service (IaaS); platform as a service (PaaS); or software as a service (SaaS)[7]. Companies are now able to adopt this service on demand. By renting as many cloud services as they need at any given moment, companies convert fixed capital costs to variable operational costs. This avoids under or over-resource provisioning, and enables minute-by-minute flexibility[8]. The cloud computing model has a major impact on firm cost structures, and consequently, on production possibilities. This is especially relevant for small and medium sized-enterprises, since they can access large infrastructures that would otherwise be unattainable[9].

The paradigm of cloud computing makes performance evaluation studies especially complex: yet such studies are becoming increasingly important for planning, administering, and tuning infrastructure and services. Performance evaluation studies are traditionally focused on computation, storage, and communication utilization. In our opinion, this offers a partial view of the system and only enables managers to make decisions on system infrastructure and its organization without capturing the intrinsic economic implications. Given the close relationship of new business models to the cloud infrastructure, the traditional approach is inadequate for decision making in a cloud environment.

Performance evaluation studies need to include customer satisfaction as a main business metric[10] because dissatisfied customers represent lost sales and generate poor reputation, which finally result in economic losses. New indexes related to economic variables are required. Such metrics should include those indicators related to quality of service (QoS), sales, or cost of losing customers[11]. Taking this new perspective into consideration, our work focuses on extending the traditional performance evaluation methodologies by including business related aspects. That is, this paper proposes a new methodology for assessing the economic costs and benefits of implementing and maintaining an e-service based on cloud technology. To this end, traditional performance models relying on indexes related to the usage of the main system resources (i.e. processor, memory, storage and bandwidth) are reformulated to include new metrics that are useful for business managers: such as customers abandonments and service costs.

This reformulation enables cloud consumers to accurately estimate the cost of adopting a particular cloud configuration, both in terms of infrastructure cost and the cost of losing customers due to any performance degradation. This view enables cloud consumers to easily consider issues related to QoS, capacity planning, and system flexibility in order to dynamically provide server resources in response to flash events.

The main contributions of this paper are: i) a new methodology for the economic evaluation of cloud-based applications from the point of view of cloud consumers as an instrument for the business model; ii) an example of applying this methodology to help a cloud e-commerce site reach its business goals; and iii) an accurate web user characterization to represent the system workload.

The remainder of this paper is organized as follows. Section 2 explains our motivation and reviews the state of the art in this field. Section 3 proposes a methodology for evaluating cloud-based applications from an economic point of view. Section 4 includes an illustrative example of applying such methodology to a cloud-based e-commerce site. Finally, Section 5 includes some concluding remarks and points to future work directions.

2. Background and motivation

It is well-known that Internet traffic fluctuates over time. Floyd et al.[12] observed that the diurnal pattern of network activity is one of the few invariant properties of the Internet. A similar observation can be made from the point of view of web-based application workloads. Indeed, most web applications fluctuate unpredictably between peak and valley loads (for example, during flash crowds such as the Slashdot effect[13]). Vallamsetty et al.[14] also found this fluctuation in e-commerce traffic, with peaks related to seasonal effects or the availability of different services.

The cloud technology adopted by well-known websites is an attractive infrastructure solution to tackle this fluctuation in loads because: i) cloud technology provides an almost unlimited on-demand infrastructure capacity; ii) a cloud infrastructure is much larger than most enterprise data centers; and iii) the pay-per-use model allows cloud consumers to pay for the actual consumption instead of paying for the peak capacity. Considering the model complexity of these systems, it is particularly interesting to include the infrastructure costs in the evaluation of system performance. Some efforts to test the performance of applications running in real cloud platforms can be found in the literature. Unfortunately, they do not consider the economic aspects related to each implementation. For instance, Tudoran et al.[15] present a traditional performance evaluation study of two cloud architectures under a workload representative of scientific applications. More recently, Mukherjee et al.[16] introduce an approach to evaluate the Amazon Cloud Platform for web applications. Wee et al.[17] present a performance study comparing different cloud providers as platforms for web applications, and propose a load balancing architecture using a scalable cloud storage service.

Only a few papers go a step beyond and consider the economic implications of the cloud. Kiruthika et al.[18] define a quality measurement model for cloud-based e-commerce applications. Their proposal is aimed at enabling companies to measure quality and align their business goals by using key performance indicators for measuring and monitoring revenues, gross margins, return on investment, productivity and customer satisfaction. Chihoub et al.[19] evaluate the monetary cost of consistency in the cloud, and reveal a noticeable cost variation when different consistency levels are considered. Moreover, some attempts have been made to model the pricing strategies that cloud providers should apply. Wang et al.[20] distinguish three types of pricing schemes that providers can follow: *pay as you go*, in which consumers are charged a fixed price per unit of resource consumption; *spot market*, in which server resources are auctioned periodically; and *subscriptions*, in which consumers pay for reserving a resource for a certain period of time. A pricing scheme similar to the spot market is proposed by Kantere et al.[21] to adaptively adjust the price that cloud providers should charge. In contrast, research by Hong et al.[22] takes the perspective of the cloud consumer to predict and minimize the cost of using cloud infrastructure.

Linking web performance metrics with business performance is not a new topic. Galleta et al.[23] find in an experimental setting that user intentions are affected by server response time when it exceeds four seconds. From the business point of view, users who are dissatisfied due to high latencies abandon their purchases, and do not return[5], thus damaging reputation and lowering future sales. Kou et al. [24] include financial criteria to evaluate clustering algorithms using MCMD methods. With this contribution, authors stand out the need of considering financial risk analysis in the software selection thus extending their previous proposal for classification algorithm selection [25].

Poggi et al.[26] analyze the relationship between long response times and sales. They describe two thresholds for user behavior: the *toleration* threshold, when the response time is between 7 and 10 seconds, which results in a 5% decrease in sales; and the *frustration* threshold, in which long response times (between 10 and 20 seconds) make users leave the page and so reduce sales by 53%. Following this work, the same authors introduce a methodology to determine how sales are affected when response times increase. This methodology can then be used in capacity planning studies[10].

For take capacity planning it is necessary to represent how users interact with web applications. Indeed, it is well-known that one of the main challenges when evaluating web applications is to devise a representative workload for the study. While real traces faithfully reproduce the load for certain client and server conditions, they cannot be easily scaled to larger systems, or different settings, when behavior largely depends on the responsiveness of the website [27]. For this reason, an accurate workload model is required so that these interactions are explicitly taken into account. There are several efforts in the literature to generate representative workloads for specific web applications by accurately modelling user behavior. Menascé and Almeida[28] introduced the customer behavior model graph (CBMG), which describes patterns of user behavior in e-commerce sites. The model was extended by Shams et al.[29] to capture application inter-request and data dependencies. Duarte et al.[30] applied this model for defining the workload of a blogspace. Similarly, Benevenuto et al.[31] introduced the clickstream model to characterize user behavior in online social networks.

In this context, in a previous work[32] we proposed the dynamic web workload model (DWEB model) to generalize workload modeling for multiple applications. This model makes it possible to characterize and reproduce the behavior of real web users in performance evaluation studies. DWEB model introduces two concepts to consider different levels of user dynamism. Firstly, the *user navigation concept* enables us to represent the dynamic reactions of users when they interact with web content and services. Secondly, the *user role concept* defines the behavior of users and their continuous changes. A workload generator was developed to mimic the behavior of the real web user community by implementing these two concepts. It also represents the physical distribution of users on the web, and improves workload accuracy by providing a distributed architecture.

This paper provides a new methodology for assessing the costs and benefits of implementing cloud-based web applications. This methodology integrates both infrastructure and business performance by taking three perspectives into account: end-users (customers); cloud consumers (enterprises); and cloud providers (suppliers of technology). To this end, the DWEB model and the workload generator[33] have been extended to include new metrics that are useful for business managers such as customer losses or sales volumes (enabling us to focus on the economic value of user abandonments). Furthermore, traditional performance indexes have been reformulated to define new business metrics related to service costs.

3. Methodology for Economic Evaluation of Cloud-based Applications

This section presents a new methodology to economically evaluate any cloud-based application. Our methodology can be used as an instrument to establish a close relationship between the underlying business model and the monetary cost of cloud infrastructure. A business model defines how an enterprise creates and delivers value to customers, and then converts payments received into profits[34]. It also outlines an architecture of revenues, costs, and profits associated with the business enterprise delivering that value.

3.1. *Participants description*

The performance evaluation of a cloud-based application from an economic point of view requires understanding the characteristics and objectives of the participants. To do so, we firstly define the main subjects or participants involved in a cloud-based architecture:

Customers: These are the web application users, and can also be considered as the final users. When using any e-service, their objective is to obtain as much value as possible. The perceived value mainly depends on the services, contents, and web design offered, and the value can be lessened by a poor web server performance that

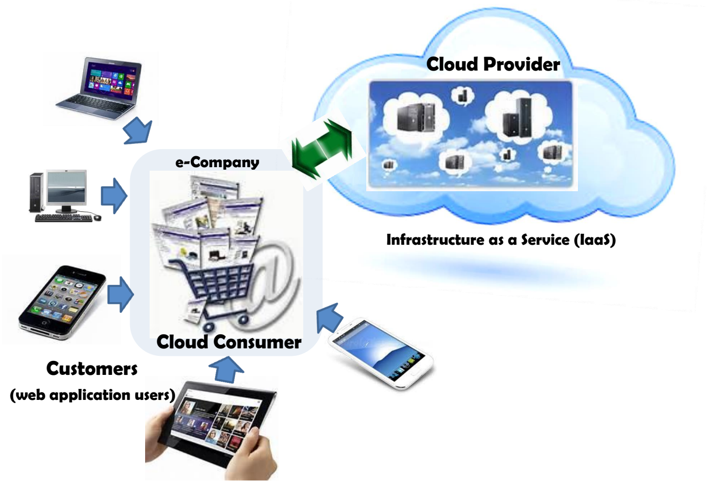


Fig. 1. Relationship among cloud participants

causes long latencies or reduced availability. As a consequence, low QoS may cause customer abandonment, which results in a loss of revenue for the company offering the e-service (the cloud consumer).

Cloud consumers: These are companies that rent cloud infrastructure to provide an e-service. Their objective is to provide consumers with as much value as possible at a minimum cost. The most relevant cost for the objective of this paper is the cost of the cloud infrastructure. Therefore, cloud consumers must trade off infrastructure costs for user abandonment due to poor service. That is, the cost of the cloud infrastructure versus the cost of losing customers. Cloud consumers play an intermediate role between customers and cloud providers.

Cloud providers: These are the owners and managers of the cloud infrastructure. From the point of view of this study they offer IaaS and make it available to the cloud consumers using a set of price plans for various infrastructure configurations and virtualization techniques. Their objective is to maximize the difference between the revenues obtained from cloud consumers and the cost of maintaining infrastructure (including hardware, energy, etc).

Although cloud providers are not directly bound to customers (final users), they are partially responsible for their satisfaction. Figure 1 illustrates the relationship between cloud business participants.

3.2. *Methodology basis*

The proposed methodology for performance evaluation combines infrastructure and business performance metrics to estimate the costs and benefits of providing a web application or service in the cloud from a wider perspective. This method is focused on the point of view of cloud consumers, as they are a central element of the cloud architecture. The methodology provides cloud consumers with a guide for deciding which cloud configuration best suits their business needs. However, our method could be extended to the perspective of cloud providers.

3.2.1. *Participants characterization*

The methodology is developed around the three participants described above and starts by defining the main objectives and requirements of each participant.

Customers. They are characterized by the *customers requirements and behaviors (CRB)* that represents their needs or expectations regarding a specific web application and their behavior depending on the service offered (e.g. loyalty, shopping actions, or abandonment). Particularly, *CRB* define the users navigational patterns and QoS conditions (mainly in terms of latency perceived) before giving up the session.

Cloud consumers. They are characterized by their *Business Goals (BG)* which are the observable and measurable end results that must be achieved by the company according to its business plan *BG* specify the target number of users, the benefits derived from serving an average customer, as well as the costs of losing him/her because of poor QoS. The latter cost is usually greater than sales lost in a session, because dissatisfied customers may not return and may spread their bad experience, so creating a poor reputation for the site and dissuading other potential customers.

Cloud Providers. For the purpose of the performance evaluation from the cloud consumers point of view, cloud providers are characterized as a set of possible *infrastructure specifications (ISs)*. The prices and characteristics of the virtual or physical machines that are offered define each IS. This includes the main hardware and software resources for each machine, such as processor, memory, storage, network, or operating system.

3.2.2. *Economic evaluation procedure*

Once the different participants are described, a performance evaluation study can be made to find the optimal infrastructure decision for the cloud consumer. This can be done by taking the following steps:

STEP 1. The first step is to model the behavior of final users according to their CRB in the particular web application under study. Describing customer requirements and behavior implies modeling traffic patterns and mimicking how end-users interact with the web application when trying to fulfill their objectives. It can also represent user abandonments due to high response times or due to other navigational difficulties such as HTTP errors or service unavailability.

STEP 2. Once user behavior is modeled, the second step is to conduct a classical system performance tuning to find the number of users that can be served with a given cloud infrastructure. This can be done by iteratively increasing the number of simulated users and keeping track of the number of users that were served and how many of them abandoned. We define *services* as the number of users that surfed the e-service website and reached a given target page (e.g., finish the order). Similarly, *abandonments* are defined as the number of users that stopped navigating the site because of long latencies and slow server performance. When the number of abandonments increases, system bottlenecks must be identified to better select the rental of a new cloud *infrastructure specification (IS).*

STEP 3. The third step is to estimate the opportunity cost of lost sales because of user abandonments. This cost has direct and indirect components. On the one hand, the direct costs are those related to the missed revenues from the users who abandoned. On the other hand, the indirect costs are related to the missed revenues due to the dissuasion effect generated by unsatisfied users. It requires an economic analysis of the business and depends on the particularities of each e-service.

STEP 4. The fourth step, which can run parallel to the third step, is an estimation of the increase in the cloud cost due to the upgrade of the new cloud infrastructure proposed in the second step. This cost should include the service cost of running the new cloud configuration, as well as the cost of changing the configuration.

After both economic costs are estimated, the decision on whether or not to test a better infrastructure should be made by comparing the opportunity cost of lost customers with the cost of upgrading cloud infrastructure. If the former is greater than the latter, the cloud configuration will be upgraded and then evaluated again as the second step suggests. Otherwise, the same cloud infrastructure will be evaluated with more incoming users. The opportunity cost of lost customers will eventually rise to the point where an upgrade of the infrastructure is worthwhile.

The performance evaluation process finishes when the number of incoming users reaches a predefined threshold.

Algorithm 3.1 provides a schematic view of the proposed methodology.

4. Case study

This section introduces a case study to illustrate how the described economic evaluation methodology can be applied to select the best choice of infrastructure specification for a cloud-based business. The presented case does not come from a real production environment but illustratively presents how our approach can be used in a real business application. The evaluated application is an online bookstore which is a well-known example of a transactional web e-commerce system. Although this type of site in its more traditional dimension does not usually require powerful hardware resources, it has been chosen because this testbed scenario is based on the TPC Benchmark$^{\text{TM}}$ W (TPC-W)[35], which is a popular benchmark and widely used by the scientific community.

There are many open platforms that enable cloud providers to offer IaaS so-

Algorithm 3.1 Economic evaluation of a cloud system

```
{STEP 1}
Model site users (CRB)
**while** $incoming\_users < max\_users$ **do**
  {STEP 2}
  Increase $incoming\_users$ until $abandonments$ increase
  Detect performance bottleneck
  Propose new IS
  {STEP 3}
  $Lost\_Customer\_Value \Leftarrow$ Economic value of the abandonments
  {STEP 4}
  $Cloud\_Upgrade\_Cost \Leftarrow$ Cost of upgrading to new cloud setting
  **if** $Lost\_Customer\_Value > Cloud\_Upgrade\_Cost$ **then**
    Upgrade cloud to new IS
  **end if**
**end while**
```

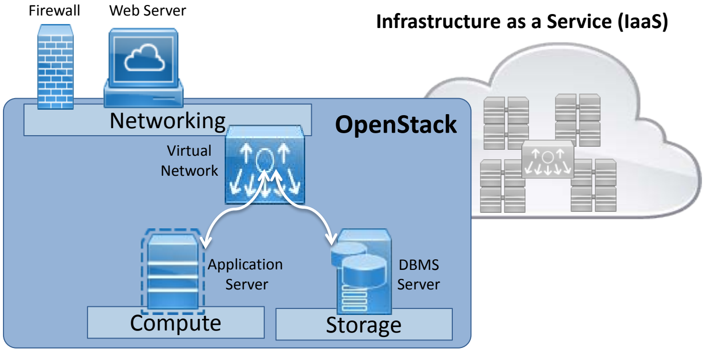


Fig. 2. OpenStack general view

lutions. In this case study, we assume that the provider offers IaaS on top of an OpenStack[36] public cloud (e.g. Rackspace). This implementation allows cloud consumers to upload their virtual machine images and instantiate them in one of the virtual hardware templates (or flavors) previously defined by the provider. These flavors correspond to the *infrastructure specifications (ISs)* introduced in Section 3.2.1. Figure 2 represents the OpenStack services that a bookstore such as the example analyzed in this case study could use.

We consider a baseline scenario where the web application resides in a simple *IS* and supports peak traffic of 50 simultaneous users. Departing from this situation, the bookstore management decides to conduct a promotional campaign to attract new visitors and increase sales. Before doing so, the company needs to know which

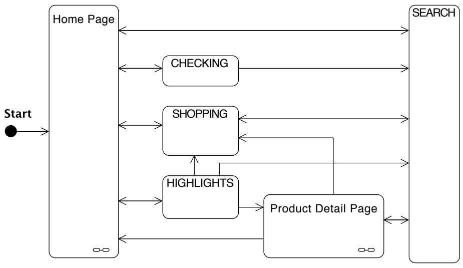


Fig. 3. Online bookstore simplified website map

IS is the best choice to serve the extra traffic at a reasonable cost in order to maximize profits. In this section we provide a detailed description of the testbed used to recreate such a scenario and then apply our methodology as a tool for business decision-making

4.1. *Testbed overview*

The testbed used in this work is the result of integrating our workload generator and TPC-W, which models an online bookstore. It was developed and validated with the aim of supporting system performance evaluation when considering dynamic workloads generated with the DWEB model[37].

Figure 3 depicts a simplified website map for the online bookstore, where pages with related functionality are included in the same group: *checking, shopping, highlights* and *search.* Navigation hyperlinks among them are indicated by arrows.

The *search* group provides the site with a book search engine. It is composed of a search page to enter the query and a result page that shows the list of returned results. The *highlights* group embraces the best-sellers and the new product pages, which arrange the bookstore catalogue according to the sales and publication date, respectively. The *shopping* group is the largest set of pages and provides: i) shopping cart with sale functionality; ii) purchase request and confirmation; and iii) payment process through secured navigation. The *checking* group includes a set of pages that enable users to check the order status. Finally, the most referred pages (home page and product detail page) are included.

This experimental setup is a typical two-tier configuration that consists of a back-end tier implementing the application functionality over a Linux Ubuntu Server hosted by Open Stack, and a front-end tier implementing the workload generation over a Linux Ubuntu Desktop. Figure 4 illustrates the hardware/software platform of the experimental setup used in this work.

The back-end provides a cloud infrastructure that runs the online bookstore, and whose core is a Java web application deployed on the Tomcat web application server. Static content requests by the web application, such as images, are served by the Apache web server, which redirects requests for dynamic content to the Tomcat

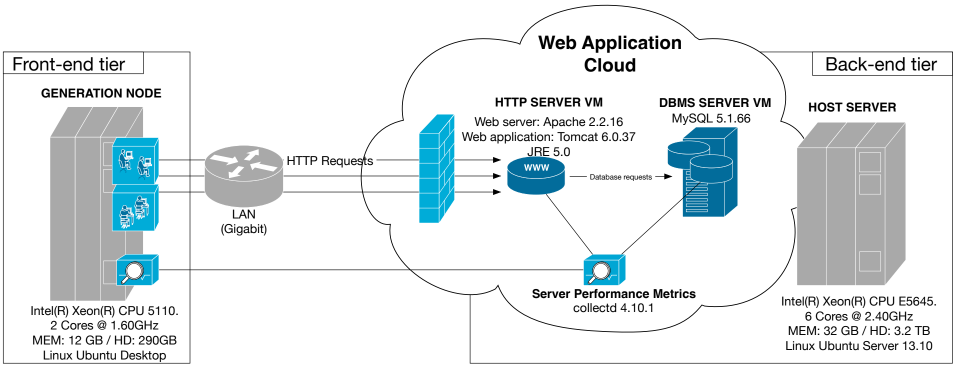


Fig. 4. Experimental setup

server. The web application generates the dynamic content by fetching data from the MySQL database.

Following the usual separation in current web systems, our back-end distributes the presentation control and data layers in two virtual machines (VM). The HTTP server VM hosts the Apache web server and the Tomcat application server, while the DBMS server VM hosts the MySQL database. In an OpenStack public cloud, each VM image would be instantiated in one of the flavors. By choosing the flavor in which they are instantiated, the number of processors and memory size used by each VM are specified. To conduct the experiments described below, the back-end tier was emulated by hosting both VMs in the same 6-CPU node server (which limits the total number of processors that can be virtualized between both VMs). The front-end tier, which generates the workload according to the DWEB model, is hosted in a different node.

Given the multi-tier configuration of this environment, system parameters (both in the server and workload generators) have been properly tuned to prevent middleware and infrastructure bottlenecks interfering with the results. The online bookstore has been configured with 500 users and a large number of items (100,000 books) that forces us to balance database accesses (e.g. the pool connection size), static content service by Apache (number of processes to attend HTTP requests), and dynamic content service by Tomcat (number of threads providing dynamic contents). Each experiment was 35 minutes long and consisted of a 15-minute warm-up period followed by a 20-minute data-collecting phase in which measurements were performed. The simulation experiments consisted of a group of simultaneous users surfing the bookstore according to the behavior modeled with DWEB model. An example of user behavior is described in the next section. All experiments were run 30 times to obtain results with a 95% confidence interval.

The main performance metrics collected at the front-end are the *number of abandonments* and *sales per session*, both related to the business performance and provided by our workload generator. At the back-end, performance metrics for VMs were gathered by `collectd`[38]. The main metrics collected for both VMs are: i) CPU utilization, which is related to the number of processors in the VM; and ii) virtual memory page faults as a metric of memory utilization.

4.2. *STEP 1. User Modeling*

This section presents how user behavior was modeled to generate the workload used in the experimental study. Although a detailed study of user navigation in an e-commerce site is not the main aim of this paper, a realistic workload is crucial to evaluate cloud-based applications from an economic point of view. Therefore, the final user model must gather the different ways consumers behave when interacting with a similar application.

According to Menascé et al.[28] there are three scenarios for web e-commerce workloads: shopping, browsing, and ordering. The shopping-intensive scenarios include user browsing and ordering activities. A browsing-intensive scenario consists of significant browsing activity but relatively little ordering activity. The ordering-intensive scenario, however, mixes significant ordering activity and relatively little browsing activity.

Based on the shopping scenario, DWEB model was used to accurately model the customers (i.e., the web-application users) according to their requirements and behavior (CRB). We define a realistic workload where customers make decisions according to their own navigation objectives and the results of previously visited pages.

The workload has been designed to include: i) customer behavior in which requests depend on the content retrieved before; ii) the use of back button and parallel browsing tabs; and iii) customer abandonment when long response times are perceived. The final workload has been composed by combining several customer navigations according to the different roles that a user can adopt when surfing this type of site. One of these navigations through the TPC-W benchmark website is illustrated in Figure 5. Nevertheless, since the aim of this paper is not user modeling, we did not include all the navigation diagrams. Readers interested in this topic are referred to Ref.[39].

Table 1 summarizes CRB features included in the workload characterization. The first feature defines a dynamic thinking time closer to real navigation[40]. This thinking time depends on the number of items in the search results.

The second and third features consider the use of the back button and the facility of opening parallel tabs or windows, respectively[41]. Finally, the fourth feature introduces customer abandonments, according to the thresholds of tolerance to the response time as defined by Poggi et al.[26]. Within the *toleration* threshold (response time of between 7 and 10 seconds), the customer abandons the website in 5% of cases. The *frustration* threshold (long response time between 10 and 20 seconds) users leave the site in 53% of cases. If response time is more than 20 seconds, the customer leaves the website.

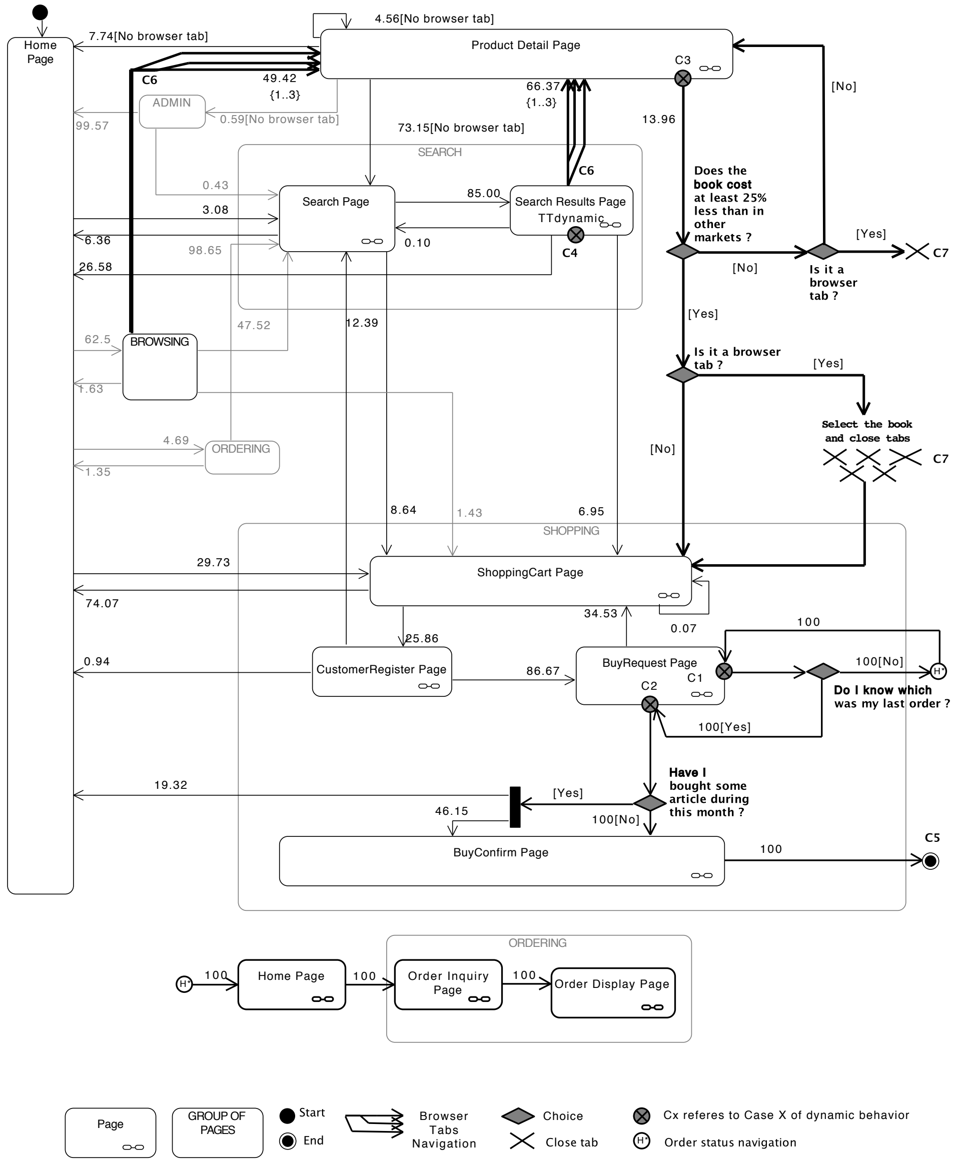


Fig. 5. Example of a loyal customer navigation

Figure 6 summarizes the session length distribution (measured as the number of visited pages) for the workload generated according to the previously described CRB when the system is not overloaded.

Table 1. Customer requirements and behavior (CRB) included in the workload characterization

Feature	Description
1	The greater the number of provided search results, the longer the time that a user takes to read and think about them.
2	A customer can return to recently visited listings of books (browsing listings) without repeating a request to the web server by using the back button.
3	When a user reviews the result of a search, he/she might begin a parallel tab browsing session.
4	A customer leaves the website when the response time is too long. • *Toleration threshold:* $7 <$ *response time* ≤ 10. • *Frustration threshold:* $10 <$ *response time* ≤ 20.

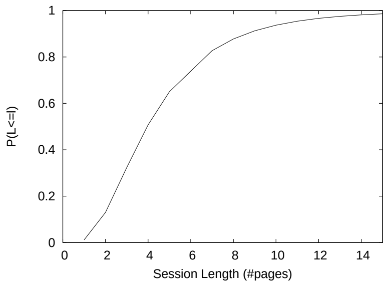


Fig. 6. Session length distribution

4.3. *STEP 2. System Performance Tuning*

The bookstore in this case study plans to initially serve up to 50 simultaneous customers. This situation can only occur during peak hours, and the number of simultaneous customers is much lower the rest of the time. Nevertheless, the system (software and hardware) must be tuned for these peak hours. Accordingly, a basic virtual infrastructure based in the cloud is rented.

A new business strategy aims to improve profits by increasing the number of customers and raising sales. However, this strategy has to take into account that more customers could overload the infrastructure and so make some of them abandon the site. Thus, it requires investing in better infrastructure to avoid such abandonments, and this action may raise costs.

To discover how many customers can be served by the current infrastructure without degrading the quality of service, a process of traditional system performance evaluation and tuning is carried out. The current basic virtual infrastructure, named I1 in this study, corresponds to the tiny flavor in OpenStack, which is also the minimum configuration that several cloud providers currently offer: one CPU and 512 MB of memory both for the HTTP server VM and DBMS server VM. According to the CRB modeled, we found that virtual infrastructure I1 can serve up to 150 simultaneous customers without any noticeable percentage of abandoned sessions. Figure 7(a) shows when the number of simultaneous customers increases from 50 to 250. From an economic point of view, this can also be observed in the evolution of the percentage of customer sessions that finish with a buying order. Figure 7(b) depicts the percentage of sessions with one or more sales for an increasing number of simultaneous users in the system. Notice that this percentage noticeably decreases when more than 150 customers surf the site at the same time.

To understand which element in the cloud infrastructure acts as a bottleneck and prevents the system from serving more customers within the considered quality parameters, we evaluated the performance of the main components, i.e. the HTTP server VM and the DBMS server VM. Figures 7(c) and 7(d) show the CPU and memory utilization for the Apache server. The use of memory is indirectly represented by the number of virtual memory page faults per minute. These graphs show a bottleneck in this server when the number of simultaneous users increases. The small memory size generates a huge number of page faults, resulting in a loss of performance. When the Apache server acts as a bottleneck, the activity in the DBMS server VM decreases, as seen in Figure 7(e), which shows its CPU time. The configuration of the DBMS server is not a problem in this case because the number of virtual page faults is negligible, as Figure 7(f) illustrates.

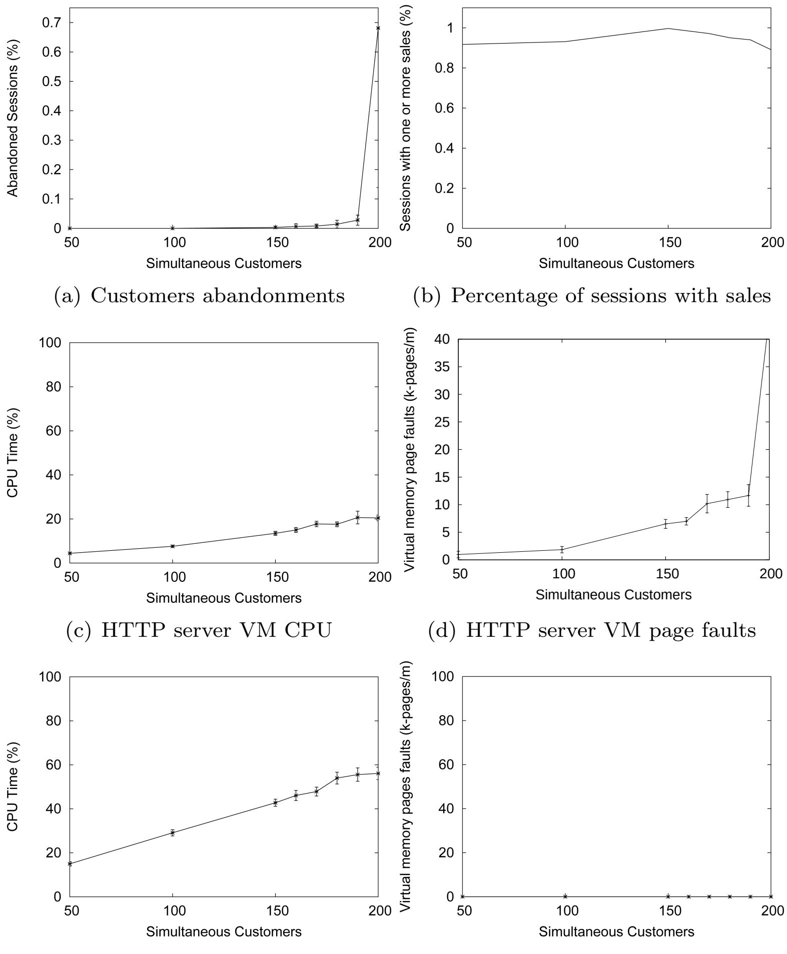
According to these results, it seems that the current basic virtual infrastructure (I1) is not suitable for the business expansion plans of the company. Serving more simultaneous customers would involve a new cloud infrastructure (I2) with more memory in the HTTP server VM.

4.4. *STEP 3. Opportunity cost estimation*

Once a significant number of abandonments occurs, a more powerful cloud infrastructure could be considered to serve more simultaneous customers. The opportunity cost of the customers lost must be estimated to assess the economic impact on the overall business strategy.

As the working case study is fictitious, we need to make some assumptions to estimate the economic value of the customers who abandon the site due to poor server performance. These assumptions are shown in Table 2.

According to the performance evaluation made in STEP 2, the virtual infrastructure I1 begins to produce abandonments when 150 customers are simulated. At this load level, 0.032% of the sessions ended with an abandonment (see Figure 8(a)), which roughly means that one session is being prematurely aborted every 3.4 minutes. If the daily average peak load lasts 30 minutes, then an average of 8.8



(a) Customers abandonments

(b) Percentage of sessions with sales

(c) HTTP server VM CPU

(d) HTTP server VM page faults

(e) DBMS server VM CPU

(f) DBMS server VM page faults

Fig. 7. Performance metrics for basic virtual infrastructure (I1)

Table 2. Assumptions for the economic evaluation of the customers lost because of the abandonments

Average peak load time (daily)	30 min
Days reaching the peak load (monthly)	20 days
Average profit for each sale	$3

sessions are abandoned due to server capacity constraints. In monthly terms, and considering that each month has 20 working days in which the peak performance is reached, this means that about 176 sessions are abandoned every month. According to Figure 7(b), the usual percentage of sessions with sales is about 0.95%, meaning that 1.68 sales (with a value of $5.03) are lost every month, as shown in Figure 8(b). This lost value represents the opportunity cost of the lost customers. Although this is a simplification, a real business manager might also want to estimate how poor performance damages business reputation.

4.5. *STEP 4. Estimating the cost of upgrading infrastructure*

To decide whether or not to upgrade the cloud setting, the opportunity cost of the lost customers is compared with the cost of upgrading the virtual infrastructure. This cost consists of two main components: a fixed cost for changing the infrastructure specification, that is, the cost of migrating from one cloud setting to another; and a service cost that is periodically paid to the cloud provider. Regarding the former, one of the advantages of a cloud system is that such fixed costs are usually low (or almost zero) as an upgrade simply requires allocating more physical resources to a virtual machine (assuming that the cloud provider and the architecture of the system are not changed). For this reason, we consider that this cost is negligible in this case.

The service cost depends on the price plans offered by the cloud providers. That is, the cost for the cloud consumer of every possible infrastructure specification that can be contracted. To simplify the example, we consider that each upgrade of the infrastructure has the same cost of $400. As this cost is lower than the opportunity cost of lost customers (as calculated in STEP 3), the company should not upgrade its cloud configuration when it expects 150 simultaneous customers during peak times.

4.6. *Iterative evaluation*

After performing the four steps detailed above, the evaluation finishes if the number of users that the cloud infrastructure can serve with the required quality of service is in line with the expansion plans of the company. Otherwise, this process should be repeated until finding the most convenient virtual infrastructure to appropriately serve all the customers.

In this case, as the cost of upgrading the infrastructure exceeds the opportunity cost of lost customers when 150 simultaneous users are considered, the virtual infrastructure I1 should be tested with more simultaneous users. Figure 8 summarizes the performance evaluation study from an economic point of view. When the bookstore receives 170 customers, the abandonment ratio is 0.08%, which means that the opportunity cost is less than $14, which is still below the cost of upgrading the infrastructure. Similarly, the opportunity cost associated with serving 190 customers with I1 is $54.42 (0.028% of abandonments). However, with 200 simultaneous users, the opportunity cost suddenly raises to $1397 (0.68% of abandonments), making the infrastructure upgrade profitable. Thus, the virtual I1 infrastructure is suited for workloads with up to 190 simultaneous customers.

The next step is to repeat the performance evaluation process for a new virtual infrastructure named I2, which is identical to I1 with the exception that the memory in the HTTP VM is 2GB. With the virtual infrastructure I2, the performance degradation starts at 200 simultaneous users (0.023% of abandonments), with an opportunity cost of $46, as Figure 8 shows. Here, the bottleneck is found to be caused by the CPU of the DBMS server, which becomes unable to serve the number of queries generated by all these customers. Although the processor utilization in HTTP VM increases with the number of customers, it maintains quite low values, as shown in Figure 9(c).

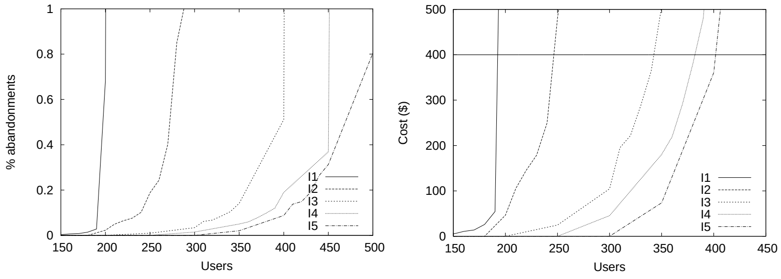
When increasing the number of customers, the opportunity cost remains below $400, while the number of simultaneous users is held below 240 (0.102% of abandonments). With 250 customers, however, the opportunity cost increases to $481. Thus, upgrading the infrastructure is recommended. A similar procedure can be followed to determine that the virtual infrastructure I3 can serve up to 340 simultaneous customers (see Figure 10); while I4 can serve up to 380 customers (see Figure 11), and I5 can serve up to 400 customers (see Figure 12).

Table 3 shows the settings that were successively chosen for the HTTP server and the DBMS server virtual machines. With these results the iterative process finishes and the company can select the best choice for its expansion plans.

5. Conclusions

This paper addresses the challenge of including economic indexes in the performance evaluation of cloud-based applications as a tool for business decision making. Cloud platforms are of special interest for companies providing e-services because virtual infrastructures can be easily and flexibly adapted to the demands of their users, peaks of traffic, QoS requirements, etc. Furthermore, IaaS solutions are now a common practice, especially among small and medium-sized enterprises, because of the impact of pay-per-use pricing on firms economic models. Converting fixed capital costs into variable operating costs dramatically reduces the cost of accessing technology for SMEs, which in turn offers great opportunities in times of global crisis.

There is a need for assessing the use of this type of cloud service from an economic and business opportunity point of view. This is of special interest in a context where the rapid development and penetration of cloud infrastructures has led to a massive but relatively disordered adoption, without a clear methodology that provides companies with the accurate parameters for their efficient use. In this vein, this work has proposed a methodological approach to include economic aspects in performance evaluations of cloud-based applications (from the point of view of cloud consumers), thus supporting business decision-making. This methodology has been applied to an online bookstore to illustrate how to select the best infrastructure specification to serve a cloud-based business. To do so, the DWEB model and our workload generator were extended to accurately represent customer behavior when browsing a site. This made it possible to include economic metrics that are



(a) Customers abandonments

(b) Opportunity cost

Fig. 8. Summary of the performance evaluation from an economic point of view

Table 3. Cloud infrastructures

	HTTP Server		DBMS Server	
Id	CPUs	Memory (MB)	CPUs	Memory (MB)
I1 (Basic)	1	512	1	512
I2	1	1024	1	512
I3	1	2048	2	512
I4	1	2048	3	512
I5	1	2048	5	512

related to customer satisfaction in the performance evaluation study, which is the key aspect of our methodology.

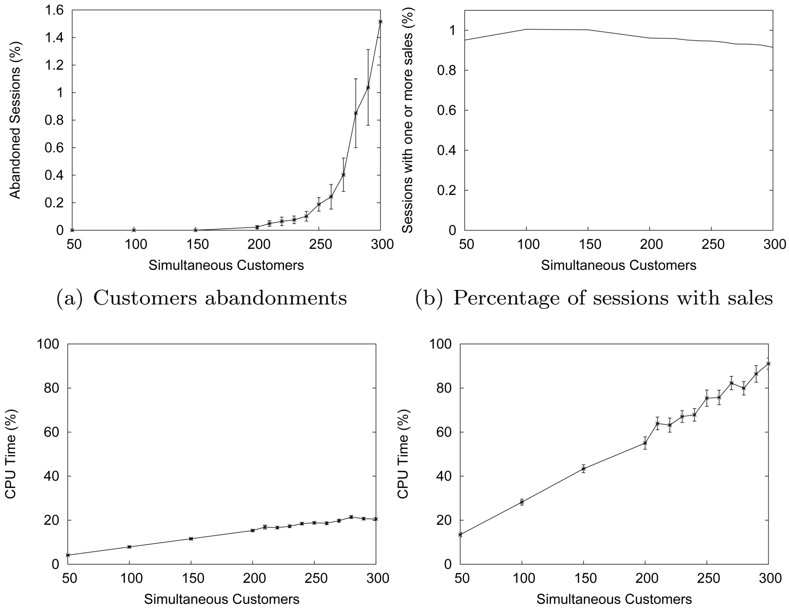
As for future work, we plan to extend our research in different ways. We intend to broaden the methodology so as to also focus on the cloud provider point of view. With this aim, it will be necessary to identify the performance metrics that best represent the cloud provider business goals and establish its relationship with the rest of the participants. Furthermore, we plan to develop new testbeds that represent other models of e-services that usually have greater infrastructure needs, such as social and collaborative networks. A detailed mathematical modelling could give a complementary perspective of the proposed methodology and is also scope for future work.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under Grant TIN2013-43913-R.

References

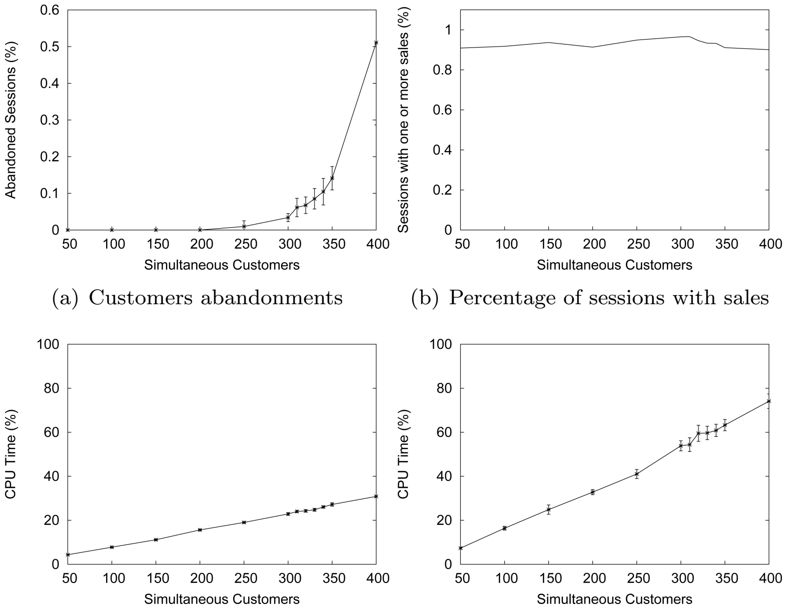
1. Miniwatts Marketing Group, Internet World Stats: Usage and Population Statistics (2014).



(a) Customers abandonments (b) Percentage of sessions with sales

(c) HTTP server VM CPU (d) DBMS server VM CPU

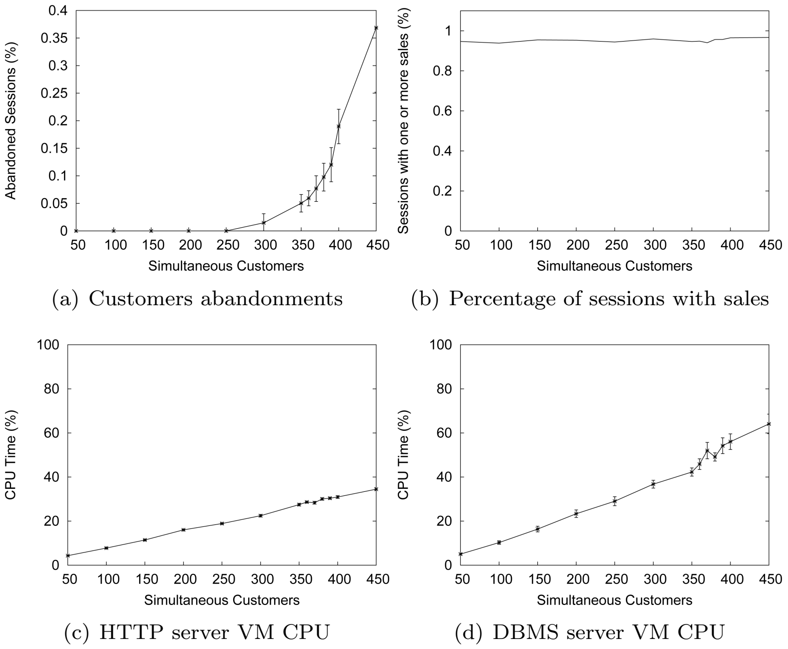
Fig. 9. Performance metrics for virtual infrastructure I2



(a) Customers abandonments (b) Percentage of sessions with sales

(c) HTTP server VM CPU (d) DBMS server VM CPU

Fig. 10. Performance metrics for virtual infrastructure I3



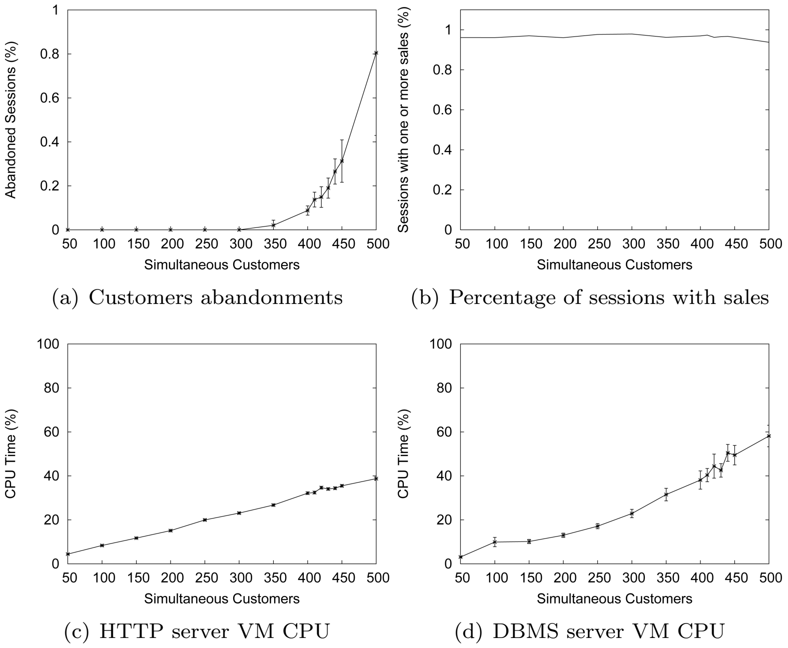
(a) Customers abandonments

(b) Percentage of sessions with sales

(c) HTTP server VM CPU

(d) DBMS server VM CPU

Fig. 11. Performance metrics for virtual infrastructure I4



(a) Customers abandonments

(b) Percentage of sessions with sales

(c) HTTP server VM CPU

(d) DBMS server VM CPU

Fig. 12. Performance metrics for virtual infrastructure I5

2. Royal Pingdom, World Internet population has doubled in the last 5 years (2012).
3. eMarketer, Ecommerce Sales Topped $1 Trillion for First Time in 2012 (2013).
4. Steve Lohr, The Age of Big Data (2012).
5. M. Armbrust, I. Stoica, M. Zaharia, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, and A. Rabkin, *A view of cloud computing*, *Communications of the ACM* **53** (2010) p. 50.
6. P. Mell and T. Grance, The NIST definition of Cloud Computing NIST Special Publication 800-145, National Institute of Standards and Technology (NIST) (2011).
7. B. P. Rimal, E. Choi, and I. Lumb, A Taxonomy and Survey of Cloud Computing Systems, in *Fifth International Joint Conference on NC, IMS and IDC* (2009) 44–51.
8. E. Bayrak, J. P. Conley, and S. Wilkie, *The economics of cloud computing*, *The Korean Economic Review* **27** (2011) 203–230.
9. F. Etro, *The economic impact of cloud computing on business creation, employment and output in Europe*, *Review of Business and Economics* **54** (2009) 179–208.
10. N. Poggi, D. Carrera, R. Gavaldà, E. Ayguadé, and J. Torres, *A methodology for the evaluation of high response time on E-commerce users and sales*, *Information Systems Frontiers* (2012)
11. Forrester Research, eCommerce Web site performance today: an updated look at consumer reaction to a poor online shopping experience Akamai White Paper (2009).
12. S. Floyd and V. Paxson, *Difficulties in simulating the Internet*, *IEEE/ACM Transactions on Networking* **9** (2001) 392–403.
13. Wikipedia, the free encyclopedia, Slashdot effect (2013).
14. U. Vallamsetty, K. Kant, and P. Mohapatra, Characterization of E-commerce traffic, in *IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems* (2002) 137–144.
15. R. Tudoran, A. Costan, G. Antoniu, and L. Bougé, A performance evaluation of Azure and Nimbus clouds for scientific applications, in *Proceedings of the 2nd International Workshop on Cloud Computing Platforms* (2012)
16. J. Mukherjee, M. Wang, and D. Krishnamurthy, Performance testing web applications on the cloud, in *Software Testing, Verification and Validation Workshops (ICSTW), 2014 IEEE Seventh International Conference on* (2014) 363–369.
17. S. Wee and H. Liu, Client-side load balancer using cloud, in *ACM Symposium on Applied Computing* (2010)
18. J. Kiruthika, G. Horgan, and S. Khaddaj, Quality measurement for cloud based e-commerce applications, in *11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science* (2012) 209–213.
19. H.-E. Chihoub, S. Ibrahim, G. Antoniu, and M. S. Pérez, Consistency in the Cloud: When Money Does Matter!, in *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (2013) 352–359.
20. W. Wang, B. Li, and B. Liang, Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing, in *IEEE 32nd International Conference on Distributed Computing Systems* (2012) 425–434.
21. V. Kantere, D. Dash, G. Francois, S. Kyriakopoulou, and A. Ailamaki, *Optimal Service Pricing for a Cloud Cache*, *IEEE Transactions on Knowledge and Data Engineering* **23** (2011) 1345–1358.
22. Y.-J. Hong, J. Xue, and M. Thottethodi, Dynamic server provisioning to minimize cost in an IaaS cloud, in *ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems* (2011) 147–148.
23. D. F. Galletta, R. Henry, S. McCoy, and P. Polak, *Web site delays: How tolerant are users?*, *Journal of the Association for Information Systems* **5** (2004) 1–28.

24. G. Kou, Y. Lu, Y. Peng, and Y. Shi, *Evaluation of classification algorithms using MCDM and rank correlation*, *International Journal of Information Technology & Decision Making* **11** (2012) 197–225.
25. G. Kou, Y. Peng, and G. Wang, *Evaluation of clustering algorithms for financial risk analysis using MCDM methods*, *Information Sciences* **275** (2014) 1–12.
26. N. Poggi, D. Carrera, R. Gavaldà, and E. Ayguadé, Non-intrusive Estimation of QoS Degradation Impact on E-Commerce User Satisfaction, in *IEEE International Symposium on Network Computing and Applications* (2011) 179–186.
27. M. Al-Ghamdi, A. Chester, and S. Jarvis, Predictive and Dynamic Resource Allocation for Enterprise Applications, in *IEEE 10th International Conference on Computer and Information Technology* (2010) 2776–2783.
28. D. A. Menascé and V. A. F. Almeida, *Scaling for E-Business: Technologies, Models, Performance, and Capacity Planning* (Prentice Hall, 2000).
29. M. Shams, D. Krishnamurthy, and B. Far, A model-based approach for testing the performance of web applications, in *International Workshop on Software Quality Assurance* (2006) 54–61.
30. F. Duarte, B. Mattos, J. Almeida, V. A. Almeida, M. Curiel, and A. Bestavros, Hierarchical characterization and generation of blogosphere workloads tech. rep. (2008).
31. F. Benevenuto, T. Rodrigues de Magalhães, M. Cha, and V. A. Almeida, Characterizing user behavior in online social networks, in *Internet Measurement Conference* (2009) 49–62.
32. R. Peña-Ortiz, J. Sahuquillo-Borrás, A. Pont-Sanjuán, and J. A. Gil-Salinas, *DWEB model: representing Web 2.0 dynamism*, *Computer Communications* **32** (2009) 1118–1128.
33. R. Pena-Ortiz, J. Domenech, J. A. Gil, and A. Pont, An approach for economic evaluation of cloud-based applications, in *Cloud Networking (CloudNet), 2014 IEEE 3rd International Conference on* (2014) 281–287.
34. D. J. Teece, *Business models, business strategy and innovation*, *Long Range Planning* **43** (2010) 172–194.
35. Transaction Processing Performance Council, TPC Benchmark$^{\text{TM}}$ W Specification. Version 1.8 tech. rep. (2002).
36. OpenStack, OpenStack open source cloud computing software (2015).
37. R. Peña-Ortiz, J. A. Gil, J. Sahuquillo, and A. Pont, *Providing TCP-W with web user dynamic behavior*, *CLEI electronic journal* **15** (2012)
38. Collectd The system statistics collection daemon (2010).
39. R. Peña-Ortiz, J. A. Gil, J. Sahuquillo, A. Pont, and J. Domenech, Generating realistic workload for web performance studies in *Modeling and Simulation of Computer Networks and Systems* (M. S. Obaidat, P. Nicopolitidis, and F. Zarai, eds.), 157–186, Waltham, MA, USA: Morgan Kaufmann (2015).
40. R. Peña-Ortiz, J. A. Gil-Salinas, J. Sahuquillo-Borrás, and A. Pont-Sanjuán, *Analyzing web server performance under dynamic user workloads*, *Computer Communications* **36** (2013) 386–395.
41. R. Peña-Ortiz, J. A. Gil, J. Sahuquillo, and A. Pont, *Surfing the web using browser facilities: a performance evaluation approach*, *Journal of Web Engineering* **14** (2015) 3–21.