Future Generation Computer Systems 55 (2016) 473–479

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Optimising resource costs of cloud computing for education

Fernando Koch [a], Marcos D. Assunção [b], Carlos Cardonha [c], Marco A.S. Netto [c,*]

[a] SAMSUNG Research Institute, Brazil
[b] INRIA, LIP, ENS de Lyon, France
[c] IBM Research, Brazil

HIGHLIGHTS

- Context-aware algorithm for allocating computing resources for class- rooms.
- Experiment setup based on real-world school data.
- Evaluation analysis considering security margin, costs, and QoS.

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form
11 March 2015
Accepted 12 March 2015
Available online 23 March 2015

Keywords:
Cloud computing
Education
Resource allocation
QoS

ABSTRACT

There is a growing interest around the utilisation of cloud computing in education. As organisations involved in the area typically face severe budget restrictions, there is a need for cost optimisation mechanisms that explore unique features of digital learning environments. In this work, we introduce a method based on Maximum Likelihood Estimation that considers heterogeneity of IT infrastructure in order to devise resource allocation plans that maximise platform utilisation for educational environments. We performed experiments using modelled datasets from real digital teaching solutions and obtained cost reductions of up to 30%, compared with conservative resource allocation strategies.

1. Introduction

Digital teaching requires new methods to continuously evaluate student performance [1]. These methods revolve around collecting, classifying, and understanding events that happen during in-classroom activities [2]. They require the instrumentation of learning environments to generate multi-dimensional signals capable to define key contextual elements. This approach generates large amounts of data that need intense computing power and storage. As a solution, Sclater envisions that "the majority of educational services will be hosted in the cloud and institutions no longer host their own data centres with expensive hardware, power bills, staff salaries and computing resources which are rarely fully utilised" [3]. A challenge in this context is to balance resource demand, expected quality of services, and operational costs thus making the use of technology viable for the education environment.

In terms of cloud computing, this means to minimise the number of allocated resources subject to keeping quality of service at an acceptable level [4]. We claim that, given the unique features of digital education, one can devise mechanisms of resource allocation tailored for this domain. For instance, it is possible to estimate the number of resources required by a classroom during a specific class based on information such as features of the learning objects in digital education material, number of students, and historical resource demand. Traditional methods, however, estimate the *peak usage* and allocate resources considering a safety margin over the worst-case scenario; this over-allocation approach results in large and undesired resource waste. The work presented by Koch et al. compared different allocation strategies and evaluated their expenditures and impact upon Quality of Service (QoS) [5].

In order to optimally exploit the cost-effectiveness of cloud computing in education, we propose a probabilistic method that allows fine-grained adjustments of load forecast models and hence enables significant cost reductions in a pay-as-you-go business model. We consider the number of resources w_c for delivering a class c, a prime α_r of demand fluctuation based on limitations of the cloud infrastructure supporting activities in classroom r, and a

[*] Corresponding author.
E-mail address: mstelmar@br.ibm.com (M.A.S. Netto).

http://dx.doi.org/10.1016/j.future.2015.03.013
0167-739X/© 2015 Elsevier B.V. All rights reserved.

prime β of safety margin which is adjusted according to the confidence level of the estimations. Special care must be taken with such methods, though, as they bring larger risks to QoS. The research questions addressed in this paper are:

- How to adjust prime α_r in order to optimise resource allocation?
- How to adjust prime β to achieve acceptable QoS levels?

To address these research questions, we constructed scenarios based on real-world digital teaching initiatives. In particular, we evaluated these scenarios considering fluctuations of infrastructure availability, which is typical in developing countries. The proposed method is analysed via discrete-event simulations considering various numbers of classrooms and using resource savings and QoS violation as metrics.

In summary, the contributions of the paper are the following:

- A probabilistic resource allocation method for educational institutions to optimise their use of cloud resources;
- An evaluation of the method under several scenarios using data based on existing digital teaching initiatives.

The article is structured as follows. Section 2 introduces the motivation of this work by describing how resource demand behaves in a real world implementation of a digital teaching platform and presents the formal description of the problem. Section 3 describes the probabilistic algorithm used for resource allocation. Results of a computational evaluation involving the proposed method are presented in Section 4. Section 5 contains the description of related work in the literature, and Section 6 presents our conclusion.

2. Motivation and problem description

Digital teaching provides means for instrumenting learning environments and novel methods to collect, classify, and understand in-classroom learning activities. The workload of such systems varies over time depending on the context and elements composing the delivered education material. For example, there are demand peaks throughout the delivery of a class when the material comprises videos, pictures, tests, screen sharing, and so forth. Moreover, fluctuations of network availability highly influence the flow of incoming requests, which leads to an undesired decrease in resource demand. When using resources from a cloud, these nuances must be considered when optimising allocation of resources in order to minimise cost and avoid waste.

The scenario considered here is that of a service provider – or educational organisation – that needs to automatically allocate resources from a cloud to deliver education services required by a school or university. We considered Samsung School solutions in this article, a real world implementation of a digital teaching platform. The addressed problem can formally be defined as follows. Let $\mathcal{R}$ denote a set of classrooms and $\mathcal{C}$ denote a set of classes. We assume that all classes are presented in each classroom exactly once over T time-slots, so that we denote by $S_{r,t}$ the class c being taught in classroom r at the t-th time-slot, $1 \leq t \leq T$. Let us consider that there is a set of learning objects $L(c)$ associated to class c. Each object l is a media element of type $m(l)$, where type in this context may refer to text, image, and video. Let us denote the amount of resources consumed by learning objects of type m by $w(m)$. The sequence of events are as follows:

1. Educator signs into a classroom r at time-slot t and confirms that class $c = S_{r,t}$ will be delivered.
2. Students located in r sign in and the applications running on their devices load the links to content in $L(c)$.
3. Educator starts the class.
4. Educator requests students to go to specific objects or pages, act upon objects, respond to tests, watch videos, *etc.*
5. Students react to educator's command in heterogeneous ways, depending on the behaviour of the cloud infrastructure supporting material delivery in r and their level of engagement.
6. The cycle loops to Step 4 until the class ends (*i.e.*, until the end of time-slot t).
7. Applications upload log files reporting all activities during the class.
8. Students and educators are prepared to start activities scheduled for time-slot $t + 1$.
9. The cycle loops to Step 1, if more classes exist.

It is clear that peak load can happen at distinct points, such as when the applications load the material (Step 2), when students act upon the content (Step 5), and when the application uploads the log files for processing (Step 7). Thus, the maximum resource demand per student throughout class c is roughly $\max_{l \in L(c)} w(m(l))$, and if the number of students located in classroom r is $n \in \mathbb{N}$, then the maximum resource demand of c is given by

$$w_c = n \left(\max_{l \in L(c)} w(m(l)) \right).$$

One may infer c from r and t given S. We will use $w_{r,t}$ and w_c interchangeably whenever $c = S_{r,t}$. We remark that maximum resource demand w_c is achieved if all students access the most resource-demanding learning content simultaneously.

Ideally, each class c is *expected* to have maximum demand w_c, independently from the classroom where it is being presented. However, as classrooms may be located in different regions and, consequently, subject to different IT infrastructure, deviations on w_c may occur. For instance, in places where data transmission is inefficient, students may not be able to access the content smoothly. In these situations, *allocated resources may be underused*.

The goal of an optimum resource allocation method is to maximise system utilisation while delivering good QoS, where QoS is harmed whenever the number of allocated resources is insufficient for the load requirements.

We propose a method that considers usage variations caused by cloud infrastructure issues to reduce over-allocation and set adequate safety margins that reduce risk of QoS degradation. This technique is specially useful in emerging countries such as Brazil, where fluctuations of data communication availability is a common reality; moreover, a successful implementation of this technique will rationalise the cost factor around cloud computing for education.

3. Probabilistic workload-aware dynamic resource allocation

Let $w'_{r,t} = w'_c$ denote the actual resource consumption demand of class $c = S_{r,t}$ for classroom r, as explained in Section 2. This value can be smaller than w_c in cases where the underlying IT infrastructure r does not deliver optimal service. That is, quality of infrastructure influences directly upon resource utilisation. This scenario emerges in schools with poor Internet connection, as execution of specific content may be impacted by the slow communication—thus students give up from playing the content, consequently curbing the data load. Moreover, for each time-slot t, let

$$w_t = \sum_{r \in \mathcal{R}} w_{r,t}$$

denote the total expected number of resources and

$$w'_t = \sum_{r \in \mathcal{R}} w'_{r,t}$$

denote the actual resource demand at t.

We assume that deviations on w_c for each class c presented in classroom r are given by a multiplicative factor α_r drawn from a Gaussian distribution $\mathcal{N}(\mu_r, \sigma_r^2)$ whose values are truncated on