A Cost Analysis of
Cloud Computing for Education

Fernando Koch, Marcos D. Assunção, Marco A. S. Netto
IBM Research
Sao Paulo, Brazil

Abstract. Educational institutions have become highly dependent on information technology to support the delivery of personalised material, digital content, interactive classes, and others. These institutions are progressively transitioning into Cloud Computing technology to shift costs from locally-hosted services to a “renting model” often with higher availability, elasticity, and resilience. However, in order to properly explore the cost benefits of the pay-as-you-go business model, there is a need for processes for resource allocation, monitoring, and self-adjustment that take advantage of characteristics of the application domain. In this paper we perform a numerical analysis of three resource allocation methods that work by (i) pre-allocating resource capacity to handle peak demands; (ii) reactively allocating resource capacity based on current demand; and (iii) proactively allocating and releasing resources prior to load increases or decreases by exploring characteristics of the educational domain and more precise information about expected demand. The results show that there is an opportunity for both educational institutions and Cloud providers to collaborate in order to enhance the quality of services and reduce costs.

1 Introduction

Education has evolved in a way that institutions have become highly dependent on information technology to improve the delivery of personalised material, and to offer digital content and interactive classes. As mentioned in the 2010 UNESCO Report [23], “the economies of scale and other features of cloud computing are likely to mean an increasing shift way from institutionally-hosted services”. We are heading towards a future where “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”.

In this context, the role of Cloud Computing is to assist educational organisations in reducing costs and focusing on their core business [2, 6, 14, 24]. For instance, by following the pay-as-you-go model, educational institutions are charged only for the services and resources they use (e.g. computing and storage resources, educational and specialised scientific software systems, and lecture material), whereas providers bear the costs of hardware and software provision.
In this model, pricing may vary depending on factors such as the time of day when resources are used, peaks in demand, and electricity costs. Cloud computing users may therefore optimise the resource utilisation by executing tasks when the costs are lower.

To optimally exploit the cost-effectiveness of this business model, educational institutions require methods to estimate demand and promptly adjust the resource allocation in response to both pre-determined and fluctuating loads. This includes being able to allocate more resources just prior to the delivery of resource-demanding classes and releasing resources right after the classes end. Although educational institutions, specially small and mid-size schools, have limited understanding on their actual IT resource demand, they often have deep understanding of domain characteristics, such as class and teacher timetables and teacher and student profiles. Cloud computing resources can be provisioned and automatically managed by leveraging the understanding of the application domain for different schools’ configurations. In this way, resource allocation is adjusted on-demand, depending on load fluctuations, on how students utilise the applications, timetables, and features of the educational material.

This paper evaluates the impact of more refined demand predictions when provisioning Cloud resources to educational institutions. We claim that if the proper method for dynamically allocating resources based on knowledge of the application is implemented, it can yield significant optimisation in capital and operational costs without compromising quality of service. Dynamic resource allocation has been highly investigated [4, 5, 10, 11, 18]. Therefore, we leverage existing work to study the effects of educational information published to Cloud providers in order to optimise resource allocation.

Our evaluation is based on a numerical analysis of three resource allocation methods that work by (i) pre-allocating resource capacity to handle peak demands; (ii) reactively allocating resource capacity based on current demand; and (iii) proactively allocating and releasing resources prior to load increases or decreases by exploring characteristics of the educational domain and more precise information about expected demand. The work assesses the impact of using specific domain information to assist resource allocation considering both IT costs and quality of service. The main contributions of this paper are the following:

– Impact evaluation of three resource allocation methods on monetary costs and quality of service. One of the methods explores domain understanding to provide better demand predictions on resource provisioning. This method considers specific features of this domain of application, such as when and how students utilise the applications, timetable, features of the educational material, and others.

– A system architecture containing the specialised elements and interactions to instrument the process of self-regulating resource allocation in Cloud computing for education.
– Analysis of the allocation methods based on the reservation of “safety margin” considering the cost/benefits of having more margin and how that can contribute to the overall QoS.

This project builds upon the business demands of the IBM Smarter Education project, which is part of the IBM Smarter Planet program. We envisage a collection of Cloud Computing based services ranging from operations, education tracking, delivery, and classroom instrumentation. In this environment, educational institutions will be able to contract services on-demand, reducing the time for and improving the cost-effectiveness of digitising the education system.

2 Background and Related Work

To meet business demands, educational institutions have become highly dependent on IT, thus constantly requiring substantial investments and skills to maintain and operate their IT systems. Keeping IT infrastructure up-to-date is important in delivering today’s educational material, providing quality of service, and maintaining students’ satisfaction [25].

Cloud Computing is being used by educational institutions as a platform for affordably offering modern and up-to-date IT resources to students [15,16]. This is particularly important in developing countries [12, 17] and for meeting the limited budgets that institutions often have as a result of the current economic turmoil [20, 25]. Cloud computing offers opportunities for cost reduction due to the economies of scale, thus resulting in a shift away from locally-hosted services [23]. The 2010 UNESCO Report [23] on Cloud Computing in Education highlights the benefits of cloud computing for institutions and students. Apart from the claimed benefits of cost reduction, elasticity, and concentration on core business, the report mentions enhanced resource availability, better end-user satisfaction, and augmented learning process and collaboration.

Another study [26] has focused on the opportunities of cloud computing to increase collaboration among multiple institutions. In addition, as discussed by Sultan [25], there are several examples of educational institutions that have adopted cloud computing to not only rationalise the management of IT resources, but also to make the education process more efficient.

Cost reductions and quality of service are key factors for educational institutions. Such factors are impacted by how Cloud providers manage their resources, and having appropriate tools for doing so is an important differentiator. The following projects have investigated aspects related to Service Level Agreements (SLAs) and load prediction methods for optimising resource management. For instance, Emeakaroha et al. [8] investigated monitoring time intervals for detecting SLA violations wherein their proposed architecture can be used to determine whether an SLA is violated and then inform the resource allocation system. The solution is reactive and does not use service workload for proactively predicting resource consumption. Li et al. [18] introduced an approach to optimal virtual-machine placement for predictable and time-constrained load peaks. The
solution, although focuses on a proactive resource allocation using prediction techniques, does not leverage specific information about the workload domain. Similar approaches were investigated by Ali-Eldin et al. [1].

Bodenstein et al. [5] have focused on resource allocation decisions, ignoring application information to predict when resource allocation should be adapted. Gong et al. [11] introduced a system called PRESS (PRedictive Elastic ReSource Scaling), which aims at avoiding resource waste and service level objective violations in the context of Cloud computing. Their goal is to avoid the use of application profiling, model calibration, and understanding of user applications. Our work takes another direction where Cloud customers provide information about their workloads in order to avoid SLA violations and reduce resource waste. Gmach et al. [10] also investigated capacity planning using historical data, but without considering the nature of the workload. Other projects [7, 9] have also explored the use of resource consumption prediction to better allocate resources. However they have not considered IT cost reductions and QoS in their studies. Adaptive resource allocation and demand prediction have also been explored in Grid and Cluster computing environments in the past [3, 4, 21, 27].

Furthermore, there are several projects on Cloud Computing and resource allocation. A key difference of our work is that it assesses the impact of using specific domain information of a workload to assist resource allocation considering both IT costs and Quality-of-Service for educational institutions.

3 Resource Allocation for Educational Institutions

This section presents three methods for resource allocation that educational institutions can use. The method described in Section 3.3 aims at enhancing allocation performance, thus reducing IT costs and increasing QoS, by exploring the understanding that education intuitions might have about their application domains.

3.1 Method to Pre-allocate Resources

The method is based on predetermined or off-line review of education environment requirements. For instance, estimate the amount of resources—e.g. cores, memory, disk space—required to execute a set of applications \(\{a, b, \ldots, z\}\) for a school with \(X\) number of students, \(Y\) classrooms, among other requirements. This can be done by either simple calculation or based on historical resource demand.

This is the solution of choice by small- and mid-size schools being the simplest to implement. The drawback is that in order to guarantee quality of service the pre-allocation is done based on the demands during the peak hour plus a given margin. That is, outside the peak period, which depends on the school’s timetable, the resources are largely idle, although the Cloud provider will still charge for their use. In this scenario, idle resources represent a waste of money for schools.
3.2 Method for Traditional Dynamic Resource Allocation

This method adjusts the resource allocation in reaction to fluctuations in demand but does not consider domain specific parameters, such as timetable, resource requirements for specific classes, student feedback, among others. For instance, the regulating method allocates more resources in response to a raise in demand. This method has clear advantages compared to pre-allocation, being more flexible and adaptable to load fluctuations. The drawbacks are: (i) resource allocation usually lags behind increases in demand, where resources are allocated a $\Delta_{\text{start}}$ amount of time after the raise in demand and remain allocated a $\Delta_{\text{stop}}$ amount of time after the demand decreases, and; (ii) this process does not perform very well under highly dynamic workloads with constant peaks and falls in demand. By using this method, a customer accepts the risks of not having the resources specified in the SLA [22].

3.3 Workload-Aware Dynamic Resource Allocation

This *workload-aware* method considers application domain specific parameters such as: (i) when and how students utilise their applications, (ii) class timetables, (iii) features of the education material, among others. This allows for assuming more “confidence” in the resource allocation adjusting process, leading to finer granularity. For example, by knowing that there is a resource demanding class about to start, the system can pre-allocate the estimated resources prior to the event. Reversely, resources can be released during blank periods in the timetable. Note that this method fits well into both physical and virtual lecture periods. The method would require adaptation when considering students accessing resources and services out of such periods.

Figure 1 depicts the proposed architecture. The core component of the architecture is the resource allocation assistant for education. This component is

![Fig. 1. Architecture for dynamic resource allocation using education information.](image-url)
responsible for passing information to the resource allocation system which allocates and releases resources according to the school's demand. To do so, the allocation assistant relies on an analytics module that leverages the following information:

- **Class Schedule**: The times when classes start and finish;
- **Class Profile**: The expected set of applications and workloads to be used in the class, how many users are expected, and the profile of the users (e.g. how interactive they are in respect to the digital devices);
- **Interaction Patterns**: How students are interacting with their devices [13, 19].

Additional information could be used to enhance the resource demand prediction. For instance, student interaction patterns collected from application interface can refine the class profile. Such information can determine how students are interacting with the application, and how much more content they are willing to consume from the Cloud. The pseudocode of the dynamic resource allocation method, which runs inside the analytics module, is described in Algorithm 1.

The algorithm starts by selecting a class according to the Class Schedule (Line 1). The second step collects the expected load (Line 2) and provision delay (Line 3) for the class. These two values are obtained from the class profile. The initial setup of the class profile can be done manually, and updated and refined during the class. Once the initial provisioning is performed (Line 4), the algorithm keeps monitoring the resources and user interaction (Lines 5-9) to determine whether resource allocation needs to be adjusted (Line 8). During this process, the class profile can be updated (Line 9). Once the class finishes, the resources are released (Line 10) and the updated class profile is stored (Line 11).

---

**Algorithm 1**: Pseudo-code for dynamic resource allocation method using educational information.

```
Input: Class Schedule, Class Profile, Device Interaction
Output: Updated Class Profile
1 class ← selectClass(classSchedule)
2 load ← getExpectedLoad(classProfile)
3 provisionDelay ← getExpectedProvisionDelay(classProfile)
4 provisionResources(class.startTime, provisionDelay, classProfile)
5 while demand for resources do
6   monitor user interaction and resource consumption
7   if demand changed then
8     adjust resources
9     update(classProfile)
10 return classProfile
11 release resources
```

---

- **Class Schedule**: The times when classes start and finish;
- **Class Profile**: The expected set of applications and workloads to be used in the class, how many users are expected, and the profile of the users (e.g. how interactive they are in respect to the digital devices);
- **Interaction Patterns**: How students are interacting with their devices [13, 19].
### 4 Case Study

In order to estimate the add-value of applying the three allocation methods described in the previous section, we elaborated a case study by comparing the opportunity cost in terms of resource utilisation between the dynamic allocation method that considers application domain information and the other two more common practices presented in Section 3 viz (1) Pre-allocate Cloud Computing resources and (2) Traditional Dynamic Resource Allocation Method.

#### 4.1 Illustrative Scenario

Figure 2 depicts the case study scenario, which consists of a school contacting a Cloud provider that offers Educational Applications and IT Resources. For instance: Collaboration Applications, e.g. instant messenger, social network, email; Digital Education Material, e.g. interactive content material to be executed on tablet devices, and; Remote execution for resource intensive applications, such as CAD and planning systems.

In this environment, (1) students are equipped with tablet devices through which they access these resources. The Cloud provider has a (2) Cloud Monitor that tracks the resource utilisation per time. Here we consider that the school has a fixed timetable for classes and an estimate of required resources per class, based on historical information and/or any other evaluation. Thus, it is possible to (3) derive the estimate resource demand per time by cross-relating this information.
Let us also consider that the school contracts its Cloud Computing service based on allocated resources. That is, it pays for the allocated resources, such as CPU cores, network, and storage, regardless of their utilisation. In order to maximise the cost effectiveness of the Cloud Computing environment, the school must implement a solution to allocate resources just prior to foreseeable demand. Reversely, the solution must release resources when the system is expected to be idle.

The experiments presented in this section consider the three allocation methods described earlier and measure the following metrics:

- **Cost**: Money spent by the school to allocate a given number of resources;
- **QoS Violations**: Number of times the resource demand is higher than the allocated resources;
- **Allocated Resource**: Number of allocated resources.

**CPU Cost**: Regarding CPU allocation, utilisation, and price, we considered process utilisation in terms of “SPEC in use per time”. One can calculate this metric based on the average CPU utilisation versus the provided SPEC performance for the CPU in use*. For the sake of calculation, we consider 40 SPEC per processor core at a cost of $0.05/core/hour. The absolute numbers are not relevant in this study, as they are directly related to the domain parameters. We provide absolute numbers for the sake of illustration. The argumentation builds upon the comparison between different scenarios.

**Margin**: Another variable considered in the evaluation is the cost margin, that is how much more from the estimated resources the school is willing to buy. The standard value for cost margin in the market is 30%, we therefore evaluated the margin of 10%, 30%, and 50%.

For the traditional dynamic allocation method, we configured the average utilisation of resources for the past 60 minutes, especially because some resource providers allocate virtual machines per hour basis.

### 4.2 Result Analysis

We first analyse the cost and quality of service considering the three methods. The results are summarised in Figure 3. We added a “safety margin” of 30% in this experiment.

The *Method of Pre-Allocating Cloud Computing Resources* (see item 4 in Figure 2) is based on pre-determined or off-line review of education environment demand (Section 3). This method is completely proactive as it calculates the resource requirements to support the estimated demands of the peak hour (plus

---

* SPEC performance: this metric is provided by the Standard Performance Evaluation Corporation (SPEC); the benchmarks are available at [http://www.spec.org/](http://www.spec.org/).
margin). In Figure 2 item (7), it is clear that there is an excess in allocation outside the peak hour period. As Cloud providers charge for allocated resources per time (i.e. regardless of utilisation), this excess translates into additional costs.

This is the method of choice by small- and mid-size schools and other corporations due to its simplicity to setup and maintain. As one can conclude from the summary in Figure 3, this approach:

- Implies higher costs due to the proactive allocation feature since resources are allocated considering the demands of the peak-hour plus margin for the whole day.
- Delivers the best quality of service, virtually zeroing the possibility of QoS violations. It happens as the pre-allocation based on the peak-hour demands results in plenty of resources available, even in the event of temporary utilisation peaks.

The Method of Traditional Dynamic Resource Allocation (see Figure 2 item (5)) works by adjusting the allocated resources in reaction to fluctuation in demand. It implements a fine-granular model that adjusts the allocation by calculating the average demand in the past minutes. That is, this method is
reactive to fluctuations of demand, adjusting the resource allocation. Figure 2 item (8) depicts the balance between past demand versus adjusted allocation. The major highlights of Figure 3 for this method are that it:

- Delivers significantly better cost-effectiveness when compared to the Pre-Allocation method. This result is directly related to (i) the precision of the estimated allocation, (ii) the fine-granular reactivity to fluctuations of demand, and (iii) the safety margin.
- Provides the worst quality of service due to the large number of QoS Violation. These events are clear in Figure 2 item (8), where the “demand line” surpasses the “allocated line”. Again, this situation is directly related to the attributed safety margin.
- One can conclude that it is possible to mitigate the poor QoS issue by allocating more safety margin. This is definitively the case. Nonetheless allocating more margin implies increasing costs to a threshold where this method is no longer cost-effective.

The Workload-Aware Dynamic Method (see Figure 2 item (6)) considers domain specific parameters such as: (i) when and how students utilise the applications, (ii) timetable, (iii) features of the education material, and others. Hence, this method combines both reactive adjustment to fluctuations of demand and proactive techniques by calculating the resource requirement to support the estimated demand based on domain parameters. Moreover, it adds a safety margin. Figure 2 item (9) depicts the balance between fluctuating demand and self-adjusted allocation. Compared to the Traditional Method, it is clear that the allocation “follows closer” the fluctuations of demand. The results can be deducted from the summary in Figure 3, as this method:

- Delivers cost-effectiveness similar to reactive methods (i.e. the Traditional Dynamic Method, thus significantly better than the Pre-Allocation method. The reasons are the same: (i) the precision of the estimated allocation, (ii) the fine-granular reactiveness to fluctuations of demand, and (iii) the safety margin.
- Provides significantly better quality of service when compared to the Traditional Dynamic Method, but not as good as the Pre-allocation Method.

Thus, the Workload-Aware Dynamic Allocation Method provides a clear add-value solution leading to a better balance between costs and quality. As we discuss below, the quality of service is directly influenced by the allocated safety margin. The higher the margin, the better the quality of service. However, the operational costs are higher as well. It is natural to ask: “how to balance costs and quality in the Workload-Aware Method?”

The Influence of Safety Margin. As mentioned, it is natural to think that allocating more safety margin to the calculations can mitigate the quality-of-service problem. Figure 4 depicts the allocated resources by applying three distinct safety margins in the Workload-Aware Method: (1) Margin 10%, (2) Margin
Fig. 4. Example of number of allocated and required resources for different safety margins.

30%, and (3) Margin 50%. The visible difference is the “buffer” between the “allocated line” and the “demand line”: the larger the margin, the larger the safety net, but with cost that needs to be considered.

The practical results can be deducted from the summary in Figure 5:

- Configuring more safety margin significantly improves the quality of service. As one can infer from Figure 5(b), increasing the safety margin reduces the
QoS violations exponentially. This results in having more “manoeuvre room” in case of temporary peak loads, as one can see in the examples in Figure 4.

– Reversely, adding more safety margin increases costs. This is intuitive, as more resources are allocated. However, as one can infer from Figure 5(a), the cost increase is linear, whereas the QoS improvement is exponential. Again, the absolute values are domain dependent and the numbers being provided work for the sake of comparison. But with the parameters and workload information being provided, it shows that with an increase of around 23% in costs the system delivers an improvement of over 10 times fewer QoS violations. Multiple tests with different cost parameters yield to similar results, leading to conclude the “incremental more costs versus exponential better quality” trend.

Therefore, we conclude that the Workload-Aware Method provides an economically viable solution for delivering quality services to any-size schools. This is achieved by balancing proactive and reactive behaviour in estimating the demands for dynamic resource allocation. The side effect of poorer Quality-of-Service can be mitigated by adding more safety margin to the allocation. We demonstrated that this approach yields significantly better quality without compromise the cost-effectiveness of the solution.
5 Final Remarks

This paper evaluated three Cloud resource allocation strategies that educational institutions can use to meet their IT demand. The strategies are: (i) resource pre-allocation based on peak demands; (ii) reactive resource allocation based on current demand; and (iii) proactive resource allocation that considers workload characteristics and parameters of the domain, in our case, education.

For the evaluation we considered two metrics: quality-of-service and resource costs. Our main finding is that the workload-aware proactive allocation method provides an economically viable solution for delivering quality services to schools. The quality-of-service provided by this method can be highly increased with a minor addition in the safety margin to the allocation. Our results show that increasing the safety margin reduces the QoS violations exponentially. This results in having more “manoeuvre room” in case of temporary peak loads. However, adding more safety margin increases costs. Although this is intuitive, the cost increase is linear, whereas the QoS improvement is exponential. In our experiments, we showed that increasing costs around 23% makes the system reduce QoS violations by over 10 times. Multiple tests with different cost parameters yield to similar results, leading to conclude the “incremental more costs versus exponential better quality” trend.

Therefore, the allocation method that explores domain specific information for better resource consumption predictions yields significantly better quality without compromising the cost-effectiveness of the solution. We find this result relevant and it serves as an incentive for educational institutions and Cloud providers to collaborate and understand better the schools demand in order to optimise the resource allocation strategies.

Acknowledgment

This material is based upon work supported by the FINEP under Contract 03.11.0371.00, MCT/FINEP/FNDCT 2010, related to the project “Platform for the Development of Accessible Vocational Training”. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of FINEP or any other related institution.

References


