Future Generation Computer Systems 28 (2012) 155–162

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

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

Empirical prediction models for adaptive resource provisioning in the cloud

Sadeka Islam [a,b,*], Jacky Keung [a,b,c], Kevin Lee [a,b], Anna Liu [a,b]

[a] National ICT Australia, Sydney, Australia

[b] School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

[c] Department of Computing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 9 March 2011
Received in revised form
11 May 2011
Accepted 28 May 2011
Available online 25 June 2011

Keywords:
Cloud computing
Resource provisioning
Resource prediction
Machine learning

ABSTRACT

Cloud computing allows dynamic resource scaling for enterprise online transaction systems, one of the key characteristics that differentiates the cloud from the traditional computing paradigm. However, initializing a new virtual instance in a cloud is not instantaneous; cloud hosting platforms introduce several minutes delay in the hardware resource allocation. In this paper, we develop prediction-based resource measurement and provisioning strategies using Neural Network and Linear Regression to satisfy upcoming resource demands.

Experimental results demonstrate that the proposed technique offers more adaptive resource management for applications hosted in the cloud environment, an important mechanism to achieve on-demand resource allocation in the cloud.

1. Introduction

Cloud computing is one of the most popular buzzwords in today's enterprise. An intrinsic feature of the cloud that differentiates it from traditional hosting services is its seemingly infinite amount of resource capacity (e.g. CPU, memory, Network I/O, disk etc.) offered at a competitive rate. It provides opportunities for start-up companies to host their applications in the cloud; thus, eliminating the overhead of procuring traditional infrastructure resources which typically takes several months.

However, a near-infinite resource pool for scale and flexibility is not the only potential of the cloud. Cloud hosting providers offer this near-infinite resource on demand using different pricing models, e.g. pay-per-use model for workloads with unforeseeable characteristics, reserved instance pricing model with long-term commitment of availability and spot instance model for workloads with flexible completion time. Therefore, application providers are able to choose an appropriate pricing model based on the anticipated workload characteristics and provision the resources accordingly in the cloud. The pay-as-you-go model and dynamic resource provisioning features of the cloud reduce the overhead associated with static provisioning, which is not considered as a cost effective option because of over-provisioning or under-provisioning of infrastructural resources at any particular moment. The time it takes to instantiate a new virtual machine (VM) instance on demand is relatively small, for instance, 5–15 minutes [1] as compared to a traditional month-long procurement process. A contemporary research challenge is to devise an intelligent way towards dynamic provision of resources in the cloud which is effective in terms of both cost and performance.

It is intuitive that if the dynamic resource scaling system is a reactive one, it might not be able to scale proportionally with the *Slashdot effect* [2] or sudden *Traffic surge* resulting from special offers or market campaigns; thus turning out to be catastrophic for application performance, leading to an unacceptable delay in response time and in the worst case, application unavailability. Therefore, proactive prediction-based resource scaling is required in order to cope up with the ever fluctuating resource usage pattern of e-commerce applications.

Predictive analysis of resource usage is the key to several crucial system design and deployment decisions such as, workload management, system sizing, capacity planning and dynamic rule generation in the cloud. Hence, our proposed prediction framework uses statistical models which are able to speculate the future surge in resource requirement; thus enabling proactive scaling to handle temporal bursty workload in a controllable way. For the prediction approach, we resort to machine learning algorithms (e.g. Neural Network and Linear Regression) and the sliding window technique, which have proved successful in the financial and health informatics area [3]. For training and testing the prediction model, we use data generated from running

* Corresponding author at: National ICT Australia, Sydney, Australia. Tel.: +61 293762194.

E-mail addresses: Sadeka.Islam@nicta.com.au (S. Islam), Jacky.Keung@comp.polyu.edu.hk (J. Keung), Kevin.Lee@nicta.com.au (K. Lee), Anna.Liu@nicta.com.au (A. Liu).

0167-739X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2011.05.027

the TPC-W benchmark [4] in the Amazon EC2 cloud. Finally, the effectiveness of the prediction framework is validated by introducing evaluation metrics e.g. Mean Absolute Percentage Error (*MAPE*), *PRED*(25) etc. We show that our framework is not only able to make accurate projections but also skilled enough to forecast resource demand ahead of the VM instance setup time.

2. Related work

The unique business model of cloud computing provides subscription-based pay-per-use virtual services that allows IT to increase capacity or add capability in real time over the Internet, a way to extend IT's existing resources on demand without investing in new infrastructure and personnel associated costs.

The challenge here is that applications hosted in the cloud need to be elastic in order to achieve economy of scale while preserving the application-specific Service Level Agreements (SLAs) such as, response time, throughput etc. The usage prediction and dynamic provisioning of resources is one of the fundamental research challenges, because a balanced trade-off between the business-level SLAs and other constraints (e.g. VM setup overhead, cost effectiveness etc.) needs to be achieved. Recent research on dynamic provisioning in the cloud explores some interesting prediction techniques and heuristics as well as some measurement metrics to evaluate the approaches from different perspective; we strive to focus on the research problem of developing resource prediction models for facilitating proactive scaling in the cloud so that hosted applications are able to withstand the variation in workload with least drop in performance and availability. This section provides an overview of some of the related techniques.

2.1. Resource provisioning

The problem of resource provisioning in the cloud has been investigated from the platform hosting provider's perspective. Van et al. [5] presented solutions to automate the management of virtual machines for service hosting platforms while optimizing a global utility function that integrates the application-level SLAs and the operating cost of the platform provider. They resorted to the Constraint Programming approach to solve the optimization problem by defining business level SLAs of the application and the resource exploitation cost of the hosting provider as constraints. The proposed management system attempts to maximize the performance of the hosted applications with an optimal operating cost for the hosting provider. Throughout the study in [5], the important assumption on the performance model of the hosted application is omitted, whether it is a web application with stringent QoS requirements or a pure analytical one.

Quiroz et al. [6] introduced a Decentralized Online Clustering (DOC) mechanism for autonomic VM provisioning that introduces the unique challenges of enterprise grids and clouds. They also recognized the problem of inaccurate cloud client resource requests that lead to over-provisioning and therefore integrated a workload modeling technique, called Quadrature Response Surface Model (QRSM) with VM provisioning. This technique is used to model the application dynamics in the cloud environment so that a feedback on the appropriateness and the number of the requested resources with respect to the application-specific SLA and QoS requirements can be provided.

On the other hand, Silva et al. [7] proposed a heuristic for optimizing the number of machines for processing an analytical job with predefined number of independent tasks so that maximum speedup can be achieved within a limited budget. However, the traffic of a web application is dynamic and random in nature; hence predicting the optimal number of machines (instances) for the fulfillment of the application level SLAs in real time and within budget is not a trivial task.

Alternatively, Lim et al. [8] addressed the problem of building an effective external controller for automated adaptive scaling of applications deployed in the cloud. They recommended the Proportional Thresholding approach which dynamically adjusts the target range (i.e. high and low thresholds) based on the number of accumulated virtual machine instances. Thus the relative effect of allocating resources becomes finer as the number of accrued resources increases; eventually resulting in being adaptive and more resource efficient.

2.2. Resource prediction techniques

Although the works of Silva et al. [7] and Lim et al. [8] presented different new algorithms for adaptive scaling, their approach is not proactive and hence more susceptible to degradation of performance due to VM instance provisioning and booting delay in the cloud. In this regard, Caron et al. [9] initiated the groundwork for a new approach to workload prediction algorithms based on past usage patterns. Since dynamic allocation and de-allocation of virtual machine instances include some overheads such as, instance setup time, performance improvement and responsiveness could be achieved if the system can predict and scale in advance to adapt with the changing workload. Based on similar characteristics of web-traffic, they proposed a pattern matching algorithm that is used to identify closest resembling patterns similar to the present resource utilization pattern in a set of past usage traces of the cloud client. The resulting closest patterns are then interpolated by using a weighted interpolation to forecast approximate future values that are going to follow the present pattern. This prediction finally aids in making dynamic scaling decisions in real-time. However, their approach has several problems; firstly, searching for similar patterns each time over the entire set of historical data is inefficient. And secondly, it may lead to over-specialization, thus turning out to be ineffective.

Finally, Kupferman et al. [10] recommended a set of scoring metrics to measure the effectiveness and efficiency of dynamic scaling algorithms in terms of availability and cost. Their analysis revealed the fact that dynamic provisioning provides significant improvement over static allocation algorithms in terms of cost with least drop in availability. Again, prediction of resource provisioning based on past usage and intelligent de-allocation using "smart kills" further enhance the efficiency of the algorithm in the face of sharp and random spikes in traffic.

2.3. Resource prediction in the cloud

Our work aims at analyzing the problem of resource provisioning from the application provider's viewpoint so that the hosted application is capable of making autonomic scaling decisions by evaluating the future resource utilization (such as, CPU, memory, Network I/O etc.) in real time and thus request for additional virtual instances beforehand through intelligent prediction to maximize performance and availability. A majority of the existing works have been in the investigation of resource provisioning from the cloud service provider perspective.

The current scope of our work is bound to the development of a performance prediction model and its statistical validation only; in future work, we plan to implement and evaluate the model on top of the public cloud. Our proposed prediction technique is more prospective and robust in terms of resource prediction in the cloud for several reasons. First, we generated the historical data by running a standard client–server benchmark, such as TPC-W [4] on Amazon EC2 and used that data for training and testing of the forecasting models. Second, our prediction framework is developed using statistical learning algorithms and the sliding window mechanism which are simple yet effective. And finally, the accuracy of the prediction framework is assessed using our proposed evaluation metrics and cross-validation [11,12].