J. Parallel Distrib. Comput. 74 (2014) 2757–2769

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

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

Detecting similarities in virtual machine behavior for cloud monitoring using smoothed histograms

Claudia Canali, Riccardo Lancellotti *

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy

HIGHLIGHTS

- We propose a technique to automatically cluster VMs with similar behavior in a cloud system.
- We exploit smoothed histograms to represent VM behavior.
- We propose a technique to automatically select useful information for VM behavior representation.
- We quantify the potential reduction in data collected for monitoring.
- We demonstrate that the proposed technique outperforms the state of the art.

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form
12 February 2014
Accepted 23 February 2014
Available online 28 February 2014

Keywords:
Cloud computing
Virtual machine clustering
Bhattacharyya distance
Histogram smoothing
Spectral clustering

ABSTRACT

The growing size and complexity of cloud systems determine scalability issues for resource monitoring and management. While most existing solutions consider each Virtual Machine (VM) as a black box with independent characteristics, we embrace a new perspective where VMs with similar behaviors in terms of resource usage are clustered together. We argue that this new approach has the potential to address scalability issues in cloud monitoring and management. In this paper, we propose a technique to cluster VMs starting from the usage of multiple resources, assuming no knowledge of the services executed on them. This innovative technique models VMs behavior exploiting the probability histogram of their resources usage, and performs smoothing-based noise reduction and selection of the most relevant information to consider for the clustering process. Through extensive evaluation, we show that our proposal achieves high and stable performance in terms of automatic VM clustering, and can reduce the monitoring requirements of cloud systems.

1. Introduction

The cloud computing paradigm has emerged in the last few years as a way to cope with the demands of modern application exploiting virtualization techniques in large data centers. Cloud data centers based on Infrastructure as a Service (IaaS) paradigm typically host several customer applications, where each application consists of different software components (e.g., the tiers of a multi-tier Web application). Each physical server in a cloud data center hosts multiple virtual machines (VMs) running different software components with complex and heterogeneous resource demand behavior. Many customers are outsourcing services and moving their applications from internal data centers to cloud platforms exploiting *long-term commitments*, purchasing several VMs for extended periods of time (for example, integrating a data center with the Amazon so-called *reserved instances*). As this scenario is, and is expected to be in the next future, a significant part of the cloud ecosystem [16], we assume in the present study that customer VMs do not change frequently the software component they are running and that a single software component is typically deployed on several different VMs for reliability and scalability purposes.

As cloud data centers grow in size and complexity to accommodate an increasing number of customers, the process of monitoring VMs resource usage to support management strategies in cloud systems becomes a major challenge due to scalability issues. As VMs are traditionally considered as independent black boxes, management strategies require to collect information about each single VM of the data center. This means that gathering data about VMs exhibiting similar behaviors results in the collection of redundant information, thus hindering the scalability of monitoring tasks for the cloud system.

* Correspondence to: Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Via Vignolese 905/b, Modena, 41125, Italy.
E-mail addresses: claudia.canali@unimore.it (C. Canali), riccardo.lancellotti@unimore.it, lancellotti.riccardo@unimore.it (R. Lancellotti).

http://dx.doi.org/10.1016/j.jpdc.2014.02.006
0743-7315/

We claim that automatically clustering together VMs with similar behaviors may improve the scalability of the monitoring process. However, this approach opens novel issues about how to represent VMs behavior and to measure their similarity. The main contribution of this paper is the proposal of a technique, namely *Smoothing Histogram-based clustering* (or *SH-based clustering* for short), to group VMs showing similar behavior in a cloud data center. The proposed technique exploits histograms of resource usage to model VM behaviors, and applies a smoothing algorithm to cope with the quantization error introduced by the histogram-based representation. The VMs similarity is determined through the Bhattacharyya distance [7], that is a statistical technique measuring the similarity of discrete probability distributions. A further qualifying contribution of the *SH-based* clustering technique is the automatic selection of the specific information that is useful for the clustering process, to avoid considering data that do not carry any meaningful information and may degrade the clustering performance due to the presence of spiky or noisy behaviors.

To the best of our knowledge, the automatic clustering of VMs with similar behavior is a problem only recently analyzed in [8,9]. In [8] clustering is based on the correlation coefficients among resource usage, which leads to highly sensitive performance with respect to the length of resource usage time series. In [9] the authors exploit an approach based on the Bhattacharyya distance requiring a separate clustering step for every VM resource, thus resulting in a non-negligible computational cost of the clustering process. On the other hand, the technique proposed in this paper outperforms the previous attempts in terms of both quality and computational cost of the clustering solution. A preliminary version of the present study was published by the authors in [10]; however, the *SH-based* proposal is a clear step ahead with respect to the original work in terms of methodological improvements (that is, use of histogram smoothing and analysis of a wider set of information to describe VM behavior) and novel experimental testbed.

We apply the proposed technique to two case studies: a first dataset coming from a cloud provider hosting VMs running Web servers and DBMS, and a second dataset obtained from a synthetic benchmark deployed on a cloud infrastructure. We show that our technique achieves high and stable performance in clustering VMs on the basis of their resource usage monitored over different time periods; in particular, the proposed clustering is effective even when the VM resource monitoring covers short periods of time (e.g., one day). Furthermore, our results demonstrate that blindly feeding every available information into the clustering process does not necessarily improve the clustering performance, demonstrating the advantage of automatically selecting relevant information.

The remainder of this paper is organized as follows. Section 2 describes the proposed technique for VM clustering. Section 3 discusses the application of the *SH-based* technique to a cloud data center. Section 4 describes the experimental testbeds used to evaluate our technique, while Section 5 presents the results of the experimental evaluation. Finally, Section 6 discusses the related work and Section 7 concludes the paper with some final remarks.

2. SH-based clustering technique

Management strategies in cloud data centers typically try to predict VM workload over a planning period of time (e.g., hours or days) based on resource usage patterns observed on past measurements, that are usually carried out with a fine granularity (e.g., 5 min intervals) [1,29]. Since management strategies consider each VM as a stand-alone object with independent resource usage patterns, the amount of information that needs to be collected represents a challenge for the scalability of the monitoring system.

The SH-based clustering technique aims to address this scalability issue by automatically grouping similar VMs based on resource behavior. The main goal is to cluster VMs of the same customer application which are running the same software component (e.g., VMs belonging to the same tier of a Web application), and therefore show similar behaviors in terms of resource usage. Then, the monitoring system can exploit a fine-grained data collection about few *representative* VMs as a representation of the behavior of a larger VM cluster [8].

In the rest of this section we describe and formalize the *SH-based* technique to automatically cluster similar VMs in a IaaS cloud system. The proposed technique is based on the following steps:

- Extraction of a *quantitative model* to describe the VM behavior through selected useful information
- *Smoothing* step to remove noisy contributions from the VM behavior description
- Definition of a *distance matrix* representing VMs similarities
- *Clustering* based on the distance matrix to identify classes of similar VMs

Each step is now described in detail, providing insight on the main design choices and their motivation.

2.1. VM behavior quantitative model

We now formally define the quantitative model chosen to represent the behavior of VMs and discuss some critical design choices involved in this step. We call the usage of a resource on a VM a *metric* and we use the probability distributions of the metrics to describe the VM behavior. Specifically, we represent such probability distributions using *histograms*. For formalization purposes, we consider N VMs and M metrics, so that $n \in [1, N]$ is a generic VM and $m \in [1, M]$ represents a generic metric.

We now explain how the histogram is built. Let $(\mathbf{X}_1^n, \mathbf{X}_2^n, \ldots, \mathbf{X}_M^n)$ be a set of time series, where $\mathbf{X}_m^n$ is the time series consisting of the samples for metric m on VM n. The corresponding probability density function $p(\mathbf{X}_m^n)$ is represented through normalized histograms. Each histogram consists of a specified number of *bins*, where each bin is associated to an interval of values the samples can take and represents the sample density for the interval, that is the fraction of samples in the time series falling within the interval.

If B_m is the number of bins considered for metric m, the histogram for metric m on VM n is the set $\mathbf{H}_m^n = \{h_{b,m}^n \forall b \in [1, B_m]\}$, where $h_{b,m}^n$ is the density associated to the b-th histogram bin and defined as:

$$h_{b,m}^n = \frac{|\{x \in \mathbf{X}_m^n : x > X_m^l(b), x \leq X_m^U(b)\}|}{|\mathbf{X}_m^n|}$$

where $|\{x \in \mathbf{X}_m^n : x > X_m^l(b), x \leq X_m^U(b)\}|$ is the number of samples in the range $(X_m^l(b), X_m^U(b)]$ and $|\mathbf{X}_m^n|$ is the number of samples in the time series. The bin upper and lower bounds are defined as: $X_m^l(b) = Xmin_m + (b-1)\Delta x_m$ and $X_m^U(b) = Xmin_m + b\Delta x_m$, where $Xmin_m$ is the minimum value of metric m for every VM, $Xmax_m$ is the maximum value of metric m for every VM, and Δx_m is the width of a bin for metric m, that is $\Delta x_m = \frac{Xmin_m - Xmax_m}{B_m}$. Fig. 1 provides a graphical example of the above defined histogram.

This definition ensures that for each metric m the number of bins is the same for every VM, which is required to compare histograms of different VMs. It is worth to note that different statistical techniques are popularly used to automatically estimate the number of bins of an histogram, such as Scott, Sturges and Freedman–Diaconis rules [28,17]. In this paper, we consider the Freedman–Diaconis [17] rule, which was proven to be the best choice to generate resource usage histograms capturing VMs behavior [10]. This rule is particularly suitable to cope with