ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

System resource utilization analysis and prediction for cloud based applications under bursty workloads

Jianwei Yin [a], Xingjian Lu [a,*], Hanwei Chen [a], Xinkui Zhao [a], Neal N. Xiong [b,c]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China
[b] School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China
[c] School of Computer Science, Colorado Technical University, Colorado Springs, Colorado 80907, USA

ARTICLE INFO

Article history:
Received 17 November 2013
Received in revised form 10 March 2014
Accepted 29 March 2014
Available online xxxx

Keywords:
Burstiness
Workload generation
Fined-grained prediction
Index of dispersion for count

ABSTRACT

Performance analysis and prediction need a solid understanding of the system workload. As a salient workload characteristic, burstiness has critical impact on resource provisioning and performance of cloud based applications. Thus performance analysis and prediction under bursty workloads are of crucial importance to cloud based applications. However, it is yet challenging for such analysis and prediction, since no accurate and effective bursty workload generator exists, as well as the fine-grained bursty workload analysis and prediction method. In this article, to deal with these challenges, a bursty workload generator has been proposed for Cloudstone (a cloud benchmark) based on 2-state Markovian Arrival Process (MAP2). Then based on this generator, a fine-grained performance analysis method, which can be used to predict the probability density function of CPU utilization, has been suggested for cloud based applications, to support better resource provisioning decision making and system performance optimization. Finally, extensive experiments are conducted in a Xen-based virtualized environment to evaluate the accuracy and effectiveness of the two methods. By comparing the actual value of Indices of Dispersion for Count with the target value deduced from MAP2 model, the experiments show the precision of our method is superior to existing works. By comparing the real and predicted system resource utilization under a variety of bursty workloads generated by the proposed generator, the experiments also demonstrate the effectiveness and accuracy of the proposed fine-grained system resource utilization prediction method.

1. Introduction

Workload has become critical in the cloud journey. On one hand, workload analysis is an important step in determining what can run effectively in a cloud environment. It is important to understand the characteristics of the workloads to determine whether it's suitable to be delivered as a cloud service. On the other hand, for applications that have been moved to clouds, workload analysis is the premise of many major operations such as software optimization, profiling and performance evaluation. To analyze the performance of a cloud system, one needs a solid understanding of its workload.

* Corresponding author. Tel.: +86 15168470230.
E-mail addresses: zjuyjw@zju.edu.cn (J. Yin), zjulxj@zju.edu.cn (X. Lu), chw@zju.edu.cn (H. Chen), zhaoxinkui@zju.edu.cn (X. Zhao), xiongnaixue@gmail.com (N.N. Xiong).

http://dx.doi.org/10.1016/j.ins.2014.03.123
0020-0255/

Some characteristics such as burstiness (which means highly variable request arrival rate or service time) of workloads have critical impact on resource provisioning strategies and performance of cloud based applications. For example, flash-crowd service requests can cause resource allocation problems and seriously degrade cloud system performance; High variance in incoming traffic and service time distributions can collapse the system in few seconds [13]; Simultaneously launching jobs for different cloud based applications, which are no longer single-program-single-execution applications, during a short time period can immediately aggravate resource competitions and load unbalancing among computing sites [33]. So performance analysis and prediction under bursty workloads is significant to cloud based application performance optimization.

However, performance analysis and prediction for cloud based applications under bursty workloads are still challenging. Firstly, there is no accurate and effective generator for bursty workloads. Though burstiness, which has been observed in Ethernet LAN [22], Web applications [6], storage systems [29], grid systems [23] and cloud systems [33], is characterized by many mathematical methods, including Self-similarity [22,6], Peakedness [9], Peak-to-mean Ratio, Coefficient of Variation, and Indices of Dispersion for Count (IDC) [14,4], a few existing works can support its generation and later analysis. Geist [17], SWAT [20] and the method proposed in [24] were developed to provide mechanisms for burstiness injection, but the accuracy, controllability and effectiveness are still far away from being satisfactory. Geist, which focuses on the characteristics of bursty workload can only control the distribution and correlation of inter-arrival times for open systems. SWAT that was built on top of Httperf, can only introduce burstiness by using a highly variable session length and think time distribution in the session mode. By using IDC, average think time and population of clients as inputs of 2-state Markovian Arrival Process (MAP2), the workload generation method proposed in [24] can inject burstiness into arrival stream in a controllable manner, however, it is difficult for users to provide the 'magic' value of IDC in practice. And the request arrival rate in bursty state was approximated inappropriately so that it often resulted in low accuracy.

Secondly, traditional performance analysis and prediction methods often focus on resource utilization calculated by mean value, which is not accurate or beneficial enough to support optimal resource provisioning and scheduling decision making. This inaccurate analysis and prediction may cause more critical impacts on cloud system resource provisioning and scheduling particularly under bursty workloads, since burstiness often means high variability on request arrival rate or service time even if the workload seems to be stable from the view of mean value. Without detailed information, such as the distribution of system resource utilization, we cannot succeed to learn the real workload demands. Furthermore, due to huge number of users, various kinds of applications, and the constantly expanding scale of clouds, the probability and intension of burstiness will be definitely enhanced in clouds. Thus the fine-grained performance analysis and prediction under bursty workloads makes much sense to cloud based applications, for supporting performance optimization, better provisioning and scheduling decision making. For instance, the authors of [37] suggested that server consolidation should considerate the fine-grained probability density function (pdf), for reducing the risk of performance violation. Though a large number of approaches for performance analysis and prediction have been proposed, few of them considers the fine-grained performance analysis and prediction methods for cloud based applications under bursty workloads.

In order to deal with these challenges, we develop a synthetic bursty workload generator (MAP2_Generator) for Cloudstone [32] (a cloud benchmark) based on 2-state Markovian Arrival Process (MAP2) in this paper. The model of our bursty workload generator is simple and it is easy to use. With some plain parameters that can be straightforwardly derived from real system logs or provided by performance analysts, the proposed generator can produce workloads with different intension of burstiness as you specified. Then, based on this generator, a fine-grained performance analysis and prediction method under bursty workloads is suggested for cloud based applications. This method can be used to predict the pdf of CPU utilization, to support cloud performance optimization, better provisioning and scheduling decision making. Extensive experiments are conducted in a Xen-based virtualized environment, to evaluate the accuracy and effectiveness of the two methods. By comparing the actual value of IDC estimated from system logs with the target value deduced from MAP2 model, the experiments show our generation method is more accurate than existing works. By comparing the real and predicted system resource utilizations under a variety of bursty workloads generated by the proposed generator, the experiments also demonstrate the effectiveness and accuracy of the proposed fine-grained system resource utilization prediction method.

The remainder of this paper is organized as follows. Section 2 introduces the bursty workload generation method for cloud benchmark Cloudstone. Section 3 describes more details the fine-grained resource utilization prediction method for both of bursty and non-bursty workloads. Section 4 uses two groups of experiments to evaluate the accuracy and effectiveness of the bursty workload generation method and the fine-grained utilization prediction method. The extensive experiment results demonstrate the effectiveness and accuracy of the two methods. Section 5 describes the related work. And Section 6 gives the conclusions of this paper.

2. Bursty workload generation

Cloudstone consists of two main parts: the client, which is a workload driver developed based on Faban,[1] and the server, which is a typical Web 2.0 application Olio.[2] Since the generation tool for bursty workloads is not provided by Faban, we design and implement a bursty workload generator (MAP2_Generator) for Cloudstone based on MAP2. The main notations used in this paper are summarized in Table 1.

[1] http://java.net/projects/faban/.
[2] http://incubator.apache.org/olio/.

Please cite this article in press as: J. Yin et al., System resource utilization analysis and prediction for cloud based applications under bursty workloads, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.03.123