Simulation Modelling Practice and Theory 57 (2015) 142–160

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

PIASA: A power and interference aware resource management strategy for heterogeneous workloads in cloud data centers

Altino M. Sampaio [a], Jorge G. Barbosa [b,*], Radu Prodan [c]

[a] Instituto Politécnico do Porto, Escola Superior de Tecnologia e Gestão de Felgueiras, CIICESI, Felgueiras, Portugal
[b] LIACC, Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
[c] University of Innsbruck, Institute of Computer Science, Innsbruck, Austria

ARTICLE INFO

Article history:
Received 16 February 2015
Received in revised form 30 June 2015
Accepted 1 July 2015
Available online 17 July 2015

Keywords:
Performance interference
Energy efficiency
CPU-intensive load
I/O intensive load
SLA
QoS

ABSTRACT

Cloud data centers have been progressively adopted in different scenarios, as reflected in the execution of heterogeneous applications with diverse workloads and diverse quality of service (QoS) requirements. Virtual machine (VM) technology eases resource management in physical servers and helps cloud providers achieve goals such as optimization of energy consumption. However, the performance of an application running inside a VM is not guaranteed due to the interference among co-hosted workloads sharing the same physical resources. Moreover, the different types of co-hosted applications with diverse QoS requirements as well as the dynamic behavior of the cloud makes efficient provisioning of resources even more difficult and a challenging problem in cloud data centers. In this paper, we address the problem of resource allocation within a data center that runs different types of application workloads, particularly CPU- and network-intensive applications. To address these challenges, we propose an interference- and power-aware management mechanism that combines a performance deviation estimator and a scheduling algorithm to guide the resource allocation in virtualized environments. We conduct simulations by injecting synthetic workloads whose characteristics follow the last version of the Google Cloud tracelogs. The results indicate that our performance-enforcing strategy is able to fulfill contracted SLAs of real-world environments while reducing energy costs by as much as 21%.

1. Introduction

There is growing interest in the use of cloud computing, which has been progressively adopted in different scenarios, such as business applications, social networking, scientific computation and data analysis experiments [18,38]. However, clouds now host a wider range of applications with diverse resources and QoS requirements. From the consumers viewpoint, it is essential that the cloud provider offer guarantees about service delivery. Typically, consumers detail the required service level through QoS parameters, which are described in service level agreements (SLAs) established with providers [25]. More concretely, SLAs specify all the expectations and obligations of the service to be provided in terms of metrics as well the penalties for violating those expectations agreed upon by all parties. Thus, SLAs are a key element for supporting and empowering QoS in cloud environments.

* Corresponding author.
E-mail addresses: ams@estgf.ipp.pt (A.M. Sampaio), jbarbosa@fe.up.pt (J.G. Barbosa), radu@dps.uibk.ac.at (R. Prodan).

http://dx.doi.org/10.1016/j.simpat.2015.07.002
1569-190X/© 2015 Elsevier B.V. All rights reserved.

Providing QoS guarantees in current cloud data centers is a very difficult and complex task due to the dynamic nature of the environment and the applications workload characteristics. The problem becomes even more complicated when considering efficient resource usage and technological limitations. Cloud computing environments are very dynamic by nature in that end customers share a large, centrally managed pool of storage and computing resources. At any one time, a substantial number of end users can be inactive (e.g., submitted jobs just finished, low utilization due to specific day of the week), which allows a service provider to opportunistically consolidate, multiplex and even transfer resources among virtual machines rented to different users. Moreover, as operational costs become more relevant, the implementation of policies for maximizing the efficiency, cost-effectiveness, and utilization of resources becomes paramount. However, balancing QoS guarantees with efficiency and utilization becomes extremely challenging because virtualization does not guarantee performance isolation between VMs. For example, an applications performance can change due to the existence of other co-resident VMs that share the last-level cache (LLC). This phenomenon is known as performance interference [13,15,19,24]. Furthermore, different applications demand different QoS requirements. For example, non-interactive batches require completion time, while transactional web applications are concerned with throughput guarantees. Different application workloads demand a diverse type and amount of resources. In particular, batch jobs tend to be relatively stable, while web applications tend to be highly unpredictable and bursty [10].

In this paper, we present a dynamic resource management strategy that optimizes power efficiency and considers the SLAs of two different types of workloads: CPU-bound (i.e., batch jobs) and network I/O-bound (i.e., transactional web) applications. We propose a mechanism that estimates the slowdown in co-hosted deadline-driven CPU-bound applications due to contention in on-chip resources and a second mechanism that responds to changes in demand for network I/O-bound applications. A scheduler algorithm is also proposed to compensate for deviations from the required performance in both types of applications. The algorithm applies readjustments in the VM to PM mapping to correct such performance deviations. The assignment of resources to VMs is also refined to optimize the energy efficiency of the underlying infrastructure.

The rest of the paper is organized as follows. Section 2 discusses related work in the area of resource provisioning in cloud data centers. Section 3 introduces the architecture of the power- and interference-aware mechanism to address the QoS of heterogeneous workloads. Section 4 considers the metrics used to assess the performance of the proposed mechanism and describes the workloads and performance deviation characteristics. Section 5 presents and discusses the results. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related work

Energy efficiency optimization and assurance of application performance have two opposite objectives. While energy wastages can be mitigated through consolidation, performance deviations are caused by interference among co-hosted applications due to technological limitations (i.e., inefficient virtualization isolation) or by abrupt demand variation in the workloads. The occurrence of either makes the previous provisioning of resources inappropriate, and corrective actions must be performed to fulfill the application QoS requirements. This section presents notable efforts to understand the causes of performance deviation in co-hosted applications and to define strategies to guarantee QoS requirements for diverse applications while maximizing energy efficiency during runtime.

2.1. Performance interference estimation in virtualized environments

Determining the relationship between allocated resources and high-level metrics in a dynamic cloud environment is not trivial and has led to intensive research in related topics. For example, Koh et al. [19] have studied this phenomenon by measuring the effects of the consolidation of two VMs running diverse resource-bound applications (i.e., CPU, memory, and I/O). Key findings include the following: (i) the performance of I/O-bound applications degrades much less when co-located with CPU- or memory-bound applications; (ii) CPU- and memory-bound applications consume a large amount of CPU resources; and (iii) the correlation between workload characteristics and performance is not linear. Huang and Lee [15] analyzed the adverse impact of performance interference from a security perspective. For this purpose, the authors exhaust one type of hardware resource by co-locating a misbehaved VM with the victim VM in different configurations. The results show the following: (i) the TCP throughput can fall below 70% when the malicious VM uses memory intensively; (ii) the CPU execution time increases by 60% when a malicious VM uses the disk I/O intensively; and (iii) the memory bandwidth for the victim VM decreases from 20% to 80% when the malicious VM heavily uses disk I/O (the decrease is closer to 80% when the malicious VM uses network I/O). Mars et al. [24] explored the impact of contention in cache, memory bandwidth, and prefetcher resources in the performance of different types of applications. The authors showed that when the pressure in the cache exceeds saturation, the impact on an application QoS no longer increases. The experiments used real Google workloads and took place in a Google cluster. In addition, Hashimoto et al. [13] measured the performance degradation due to contention in network I/O, disk I/O, and on-chip resources. The results indicated that an application that uses more resources degrades more, and the impact on performance can change between 1% and nearly 190%. Additionally, the results indicate that running two programs with high memory usage or network I/O causes high overhead in the hypervisor. Kousiouris et al. [20] studied the effect of critical parameters on the performance of VMs. Their study concluded that there is a nearly linear relationship between CPU share (i.e., quantum) and performance improvement. Additionally, application performance can