Sustainable Computing: Informatics and Systems 4 (2014) 225–240

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

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

Quality of service modeling for green scheduling in Clouds

Tom Guérout[a,b,c,*], Samir Medjiah[a,d], Georges Da Costa[b], Thierry Monteil[a,c]

[a] CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France
[b] IRIT/Toulouse University, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
[c] Univ de Toulouse, INSA, LAAS, F-31400 Toulouse, France
[d] Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France

ARTICLE INFO

Article history:
Received 1 March 2014
Received in revised form 7 July 2014
Accepted 11 August 2014

Keywords:
Cloud computing
Quality of service
Energy-efficiency
Dynamic Voltage and Frequency Scaling
Multi-objective optimization
Simulation

ABSTRACT

Most Cloud providers support services under constraints of Service Level Agreement (SLA) definitions. The SLAs are composed of different quality of service (QoS) rules promised by the provider. Thus, the QoS in Clouds becomes more and more important. Precise definitions and metrics have to be explained. This article proposes an overview of Cloud QoS parameters as well as their classification, but also it defines usable metrics to evaluate QoS parameters. Moreover, the defined QoS metrics are measurable and reusable in any scheduling approach for Clouds. The use of these QoS models is done through the performance analysis of three scheduling approaches considering four QoS parameters. In addition to the energy consumption and the Response Time, two other QoS parameters are taken into account in different virtual machines scheduling approaches. These parameters are dynamism and robustness, which are usually not easily measurable. The evaluation is done through simulations, using two common scheduling algorithms and a Genetic Algorithm (GA) for virtual machines (VMs) reallocation, allowing us to analyze the QoS parameters evolution in time. Simulation results have shown that including various and antagonist QoS parameters allows a deeper analysis of the intrinsic behavior and insight of these three algorithms. Also, it is shown that the multi-objective optimization allows the service provider to seek the best trade-off between service performances and end user's experience.

1. Introduction

Cloud **Q**uality **o**f **S**ervice (QoS) [1] parameters are used by Cloud providers to establish their **S**ervice **L**evel **A**greement (SLA) contract. From the user's point of view, SLA parameters correspond to the facilities they can have using a service, mainly in terms of performance, virtual machines renting cost or availability. From the service provider point of view, the QoS analysis can be very different. Two main goals can be distinguished: Cloud providers define their SLA with the aim to satisfy their users by achieving high performance and ensuring a safe functioning. This first aim is related to the user point of view, but the second one is more dedicated to providers interests as the energy consumption can help to secure certain amount of benefits, while for example, trying to minimize the amount of virtual machine provisioning to respect SLA guarantees. Many studies have been carried out in these domains using benchmarks, metrics and technical solutions [2–5] and they are also widely proposed in Cloud scheduling studies.

In current Cloud quality of service research, the main objective metrics are the energy consumption, the virtual machine provisioning, services response-time, and, from an economical view, the cost can be taken into account. Indeed, these objectives are relevant. The energy consumption is a key issue for the environment as data-centers are energy consuming, and the cost and performance of Cloud services are the most important parameters the users are looking for.

Nowadays, the analysis of Clouds quality of service parameters can be performed in depth. The main contribution of this article is the definition of Clouds QoS parameters, which are organized into four categories. Each QoS parameter is described, ranging from the standard network performance analysis to more non-common Clouds parameters. The contribution proposed in this article is threefold. First, this article proposes a large list, but also defines as much as possible metrics that allow to evaluate them. These metrics are intended to be measurable and reusable by anyone who wants to take care about Cloud quality of service in research studies. Second, this article highlights how this quality of service study can be used in any scheduling approach, and how it could help

[*] Corresponding author at: LAAS-CNRS, 7 avenue du Colonel Roche, 31400 Toulouse, France. Tel.: +33 561336924.
E-mail addresses: tguerout@laas.fr (T. Guérout), smedjiah@laas.fr (S. Medjiah), dacosta@irit.fr (G. Da Costa), monteil@laas.fr (T. Monteil).

http://dx.doi.org/10.1016/j.suscom.2014.08.006
2210-5379/© 2014 Elsevier Inc. All rights reserved.

to enrich multi-objectives algorithms evaluation. Finally, the article proposes a Cloud architecture modeling, including the use of the DVFS [6] (Dynamic Voltage and Frequency Scaling) tools which allow to dynamically change the CPU frequency and gives an interesting trade-off between performance and energy-efficiency.

This article is organized as follows: the related work section surveys some references of different Cloud modeling, different approaches of Clouds scheduling and SLA definitions of the current best known Cloud providers. Section 3 describes the contributions of this article. It details the Cloud architecture model defined for this study. It also contains the enumeration of quality of service parameters, and explains how the proposed QoS metrics can improve scheduling approaches for Clouds. The validation methodology, in Section 4, presents in detail the Genetic Algorithm (GA) implemented for the evaluation phase, and two other basic algorithms. In Section 5, the simulations results of these three scheduling approaches are compared and analyzed. Finally, Section 6 concludes this article and discusses future works.

2. Related work

This section presents previous work in three different Clouds fields: modeling, scheduling studies and an analysis of real Cloud providers SLA proposition. In the domain of Clouds modeling, the main issue is to identify a model which points out clearly the essential characteristics. In the survey [7], some models of Clouds have been presented and widely analyzed. The following Clouds modeling related works section summarizes the more relevant articles which propose both energy and quality of service models or approaches that make use of CPU frequency scaling. In Cloud scheduling studies, a detailed analysis of previous works has been carried out in [8] with a focus on works that combine complex energy-efficient solutions and QoS parameters evaluation. Since the modeling of Quality of service in Clouds is one of the main focus of this article, an analysis of Software-as-a-Service (SaaS) providers' SLAs is also proposed.

2.1. Cloud modeling

Abdelsalam et al. [9] have analyzed the mathematical relationship between SLAs and the number of servers used. The energy consumption is also taken into account and their Cloud model uses homogeneous hosts with DVFS enabled, allowing each physical machine to use different frequencies. Their study includes the number of users and their needs, and the average Response Time of users requests.

Islam et al. [10] have developed an elasticity model for Cloud instances. They have assumed that each resource type (CPU, memory, network bandwidth, etc.) can be allocated in units and the users are aware of the allocated resources and the relevant QoS metrics for their requests, as in the case of Amazon CloudWatch.[1] Their proposed model combines the cost of provisioned but underutilized resources and the performance degradation cost due to under-provisioning. The consumer's detriment in over-provisioning is the difference between chargeable supply and demand, while the cost of under-provisioning is quantified through the percentage of rejected requests. Authors have also assumed that the customers are able to convert the latter into the estimated their financial impact.

Gelenbe et al. [11] have formulated an optimization problem for load sharing between a local and a remote Cloud service. Their study defines a multi-objective function formulated to optimize Response Time and energy consumption per job. Their approach also includes a Poisson job arrivals rate.

Baliga et al. [12] proposes a study of the energy consumption for processing large amounts of data, management, and switching of communications. Their article associated three Cloud Computing services, namely storage as a service, software as a service and processing as a service. The energy consumption was considered as an integrated supply chain logistics problem involving processing, storage, and transport and has been analyzed in both public and private Clouds.

Garg et al. [13] have modeled various energy characteristics, such as Energy Cost, carbon emission rate, workload and CPU power efficiency. Their model considers homogeneous CPUs, a cooling system which depends on a coefficient of performance (COP), and the use of the DVFS. Their performance evaluation includes many metrics and many parameters: average energy consumption, average carbon emission, profit gained, urgency class and arrival rate of applications and data transfer cost.

Beloglazov et al. [14] have studied the performance degradation of virtual machines taking into account CPU, memory and bandwidth utilization. The virtual machine provisioning is defined as a QoS parameter, and an SLA violation occurs when a virtual machine does not have the required amount of CPU.

Mi et al. [15] have formulated the multi-constraint optimization problem of finding the optimal virtual machine consolidation on hosts while minimizing the power consumption. An application load prediction is computed and they proposed a heuristic based on Genetic Algorithms in order to find a near optimal reconfiguration policy. The objective function (i.e. fitness) is composed of the power consumption function and a penalty function to keep the CPU utilization between two threshold values.

2.2. Cloud green scheduling

Beloglazov et al. [16] propose a resource management system for efficient power consumption that reduces operating costs and provides quality of service. Energy saving is achieved through the continued consolidation of virtual machines according to resource utilization. The QoS is modeled by the amount of resource needed in Millions Instructions Per Second (MIPS) for the CPU, the amount of Memory (RAM), and by the network bandwidth rate. An SLA violation occurs when a VM cannot have the required three amounts. In [17] multiple energy-aware resource allocation heuristics are presented. However, only simulation-based results based on simple migration and energy-cost models are presented, only considering the CPU load.

Duy et al. [18] design, implement, and evaluate a scheduling algorithm integrating a predictor of neural networks to optimize the power consumption of servers in a Cloud. The prediction of future workload is based on demand history. According to the prediction, the algorithm turns off unused servers and restarts them to minimize the number of servers running, thus also minimizing the energy consumption.

Binder and Suri [19] propose an algorithm of allocation and dispatching of tasks that minimizes the number of required active servers, managing to reduce energy consumption which is inversely proportional to the number of concurrent threads running in workloads.

Srikantaiah et al. [20] study how to obtain consolidation of energy efficiency based on the interrelationship among energy consumption, resource utilization, and performance of consolidated workloads. It is shown that there is an optimal point of operation among these parameters based on the Bin Packing problem applied to the problem of consolidation.

In [21], Dasgupta et al. pose workload normalization across heterogeneous systems, such as Clouds, as a challenge to be addressed.

[1] http://aws.amazon.com/fr/cloudwatch/.