The Journal of Systems and Software 99 (2015) 20–35

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment

Huangke Chen [a], Xiaomin Zhu [a,*], Hui Guo [b], Jianghan Zhu [a], Xiao Qin [c], Jianhong Wu [d]

[a] Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, PR China
[b] School of Computer Science and Engineering, University of New South Wales, NSW 2052, Australia
[c] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849-5347, USA
[d] Department of Mathematics and Statistics, York University, Toronto M3J1P3, Canada

ARTICLE INFO

Article history:
Received 13 January 2014
Received in revised form 4 July 2014
Accepted 28 August 2014
Available online 6 September 2014

Keywords:
Green cloud computing
Uncertain scheduling
Proactive and reactive

ABSTRACT

Green cloud computing has become a major concern in both industry and academia, and efficient scheduling approaches show promising ways to reduce the energy consumption of cloud computing platforms while guaranteeing QoS requirements of tasks. Existing scheduling approaches are inadequate for real-time tasks running in uncertain cloud environments, because those approaches assume that cloud computing environments are deterministic and pre-computed schedule decisions will be statically followed during schedule execution. In this paper, we address this issue. We introduce an interval number theory to describe the uncertainty of the computing environment and a scheduling architecture to mitigate the impact of uncertainty on the task scheduling quality for a cloud data center. Based on this architecture, we present a novel scheduling algorithm (PRS[1]) that dynamically exploits proactive and reactive scheduling methods, for scheduling real-time, aperiodic, independent tasks. To improve energy efficiency, we propose three strategies to scale up and down the system's computing resources according to workload to improve resource utilization and to reduce energy consumption for the cloud data center. We conduct extensive experiments to compare PRS with four typical baseline scheduling algorithms. The experimental results show that PRS performs better than those algorithms, and can effectively improve the performance of a cloud data center.

1. Introduction

Cloud computing has become a paradigm for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell and Grance, 2009). To satisfy such a soaring demand of computing services, IT companies (e.g., Google and Facebook) are rapidly deploying distributed data centers in different administrative domains around the world.

Consequently, tens of thousands of hosts in these data centers consume enormous energy for computing and equipment cooling operations (Luo et al., 2012). It is reported that the energy consumed in data centers is about 1.5% of the global electricity in 2010, and the percentage will be doubled by 2020 if the current trends continue (Koomey, 2011). Apart from the operating cost, high energy consumption will result in low reliability of the system since the failure rate of hosts doubles for every 10-degree increase in temperature (Cameron et al., 2005). In addition, high energy consumption has a negative impact on environment because generating electrical energy from fossil fuels produces a large amount of CO_2 emissions, which are estimated to be 2% of the global emissions (Pettey, 2007). Therefore, reducing energy consumption or conducting green computing has become a grand challenge when deploying and operating cloud data centers.

With the development of virtualization technology (Barham et al., 2003), a single physical host can run multiple virtual machines (VMs) simultaneously. In addition, the VMs can be relocated by live operations, such as VM creation, VM live migration and VM deletion, to achieve fine-grained optimization of computing resources for cloud data centers. This technology offers significant opportunities for green computing (Beloglazov et al., 2012). Leveraging

* Corresponding author.
E-mail addresses: hkchen@nudt.edu.cn (H. Chen), xmzhu@nudt.edu.cn (X. Zhu), huig@cse.unsw.edu.au (H. Guo), jhzhu72@gmail.com (J. Zhu), xqin@auburn.edu (X. Qin), wujh@mathstat.yorku.ca (J. Wu).

[1] Proactive and Reactive Scheduling.

http://dx.doi.org/10.1016/j.jss.2014.08.065
0164-1212/

the capabilities of virtualization technology, one can scale up or down VMs rapidly according to the current workloads in the system. When the system is overloaded, more VMs are added; when the system is underloaded, the VMs can be consolidated to a minimal number of physical hosts and the idle hosts can be turned off. Hosts in a completely idle state can dissipate over 70% as much power as when they fully utilized (Ma et al., 2012). Turning-off idle hosts, therefore, means significant power savings.

Nevertheless, the virtualization also brings about new challenges to the resource management in clouds due to the fact that multiple VMs can share the hardware resources (e.g., CPU, memory, I/O, network, etc.) of a physical host (Kong et al., 2011). The resource sharing may cause the performance of VMs subjecting to considerable uncertainties in cloud computing environments mainly due to I/O interference between VMs (Bruneo, 2014; Armbrust et al., 2010) and hosts are overloaded (Beloglazov and Buyya, 2013). For example, the ready time and the computing capacity of a VM arbitrarily varies over time, which makes it difficult to accurately measure the execution timing parameters and resource usage of VMs. Such dynamic and non-deterministic characteristics of the VM computing cause great difficulties for efficient resource management in clouds.

In addition, a primary fraction of computing applications in cloud data centers are real-time tasks. The arrival times of these tasks are dynamic and the predictions of their execution duration can also be difficult and sometimes impossible (Van den Bossche et al., 2010), since most real-time tasks are fresh and no much information is available to help the accuracy of the predictions. The imprecise execution prediction and dynamic task arrival time leave the associated scheduling timing constraints (i.e., start time, execution time and finish time) under considerable uncertainty. Furthermore, real-time tasks often need deadlines to guarantee their timing requirements, which further exacerbates the problem of efficient task scheduling and resource management.

Motivation: Due to the dynamic and uncertain nature of cloud computing environments, numerous schedule disruptions (e.g., shorter or longer than expected task execution time, variation of VM performance, arrival of urgent tasks, etc.) may occur and the pre-computed baseline schedule may not be executed and may not be effective in real execution. Unfortunately, the vast majority of researches did not consider the uncertainties of clouds, which may leave a large gap between the real execution behavior and the behavior initially expected. To address this issue, we study how to describe these uncertain parameters, how to control uncertainties' impact on scheduling results, and how to reduce the energy consumption in cloud data centers, while guaranteeing the timing requirements of real-time tasks.

Contributions: The major contributions of this work are:

- An uncertainty-aware architecture for scheduling real-time tasks in the cloud computing environment.
- A novel algorithm named PRS that combines proactive with reactive scheduling methods for scheduling real-time tasks and computing resources when considering the uncertainties of the system.
- Three system scaling strategies according to dynamic workloads to reduce energy consumption.
- The experimental verification of the proposed PRS algorithm based on randomly generated test instances and real world traces from Google.

The remainder of this paper is organized as follows. The related work in the literature is summarized in Section 2. Section 3 presents the scheduling model and the problem formulation. The energy-aware scheduling algorithm for real-time tasks considering the uncertainties of the system is introduced in Section 4. Section 5 conducts extensive experiments to evaluate the performance of our algorithm by comparing it with four baseline scheduling algorithms. Section 6 concludes the paper with a summary and future directions.

2. Related work

In recent years, the issue of high energy consumption in cloud data centers has attracted a great deal of attention. In response to that, a large number of energy-aware scheduling algorithms have been developed. Among them, there are two typical approaches. One is DVFS (dynamic voltage and frequency scaling) based and another is machine virtualization based.

The DVFS technique makes trade-offs between processor power and performance and has been commonly used to reduce the power consumption of data centers. For example, Garg et al. (2011) proposed near-optimal energy-efficient scheduling policies that leverages DVFS to scale down the CPU frequency as far as possible to minimize the carbon emission and maximize the profit of the cloud providers. Li and Wu (2012) proposed a Relaxation-based Iterative Rounding Algorithm (RIRA) for DVFS-enabled heterogeneous multiprocessor platforms to minimize overall energy consumption while meeting tasks deadlines. Rizvandi et al. (2011) focused on the issue of high energy consumption in cluster, and presented the MVFS-DVFS algorithm to fully utilize slack times and reduce energy consumption on processors. Zhu et al. (2013) proposed an energy-efficient elastic (3E) scheduling strategy to make trade-offs between users' expected finish time and energy consumption by adaptively adjusting CPU's supply voltages according to the system workload. However, the DVFS is mainly implemented on host processor machines and their energy consumption contributes about one-third of the total system power (Ahmad and Vijaykumar, 2010). In addition, only the dynamic power (about 30% of the processor power Beloglazov et al., 2012) can be moderated by DFVS. Due to those limitations of DVFS, the virtualization technique, used to consolidate VMs for low energy consumption of data centers, is becoming popular and is the focus in this paper.

The vast majority of the energy-aware scheduling research efforts over the past several years have concentrated on dynamical consolidation of VMs according to the system workload, to reduce the number of physical hosts so that the idle hosts can be switched off for low energy consumption. Hermenier et al. examined the overhead of migrating a VM to its chosen destination, and proposed a resource manager named Entropy for homogeneous clusters. The Entropy can dynamically consolidate VMs based on constraint programming and explicitly takes into account the cost of the migration plan (Hermenier et al., 2009). Younge et al. (2010) developed a novel green framework where green computing was realized by energy efficient scheduling and VM management components. Srikantaiah et al. (2008) explored the inter-relationships among energy consumption, resource utilization, and performance of consolidated workloads, and designed a heuristic scheme for minimizing system energy consumption while meeting the performance constraint. Hsu et al. (2014) studied how to dynamically consolidate tasks to increase resource utilization and reduce energy consumption, and presented an energy-aware task consolidation (ETC) method to optimize energy usage in cloud systems. Our work also leverages the VM consolidation technique to reduce the systems' energy consumption. However, unlike the above existing approaches, where uncertainties of tasks' execution times and VMs' performance were not considered, our design takes the uncertainties into account, and we employ proactive and reactive methods to mitigate the impact of uncertainties on the scheduling quality for cloud data centers.