Information Sciences 360 (2016) 57–72

Contents lists available at ScienceDirect

Information Sciences

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

TRS: Temporal Request Scheduling with bounded delay assurance in a green cloud data center

Jing Bi[a,b], Haitao Yuan[c,*], Wei Tan[d], Bo Hu Li[e]

[a] School of Software Engineering, Beijing University of Technology, Beijing 100124, China
[b] Beijing Engineering Research Center for IoT Software and Systems, Beijing 100124, China
[c] School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China
[d] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
[e] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 17 February 2016
Revised 6 April 2016
Accepted 19 April 2016
Available online 26 April 2016

Keywords:
Scheduling
Cloud data center
Particle swarm optimization
Simulated annealing
Energy management

ABSTRACT

The growing deployment of Internet services in cloud data centers significantly increases the grid energy cost of cloud providers. Considering the environmental effect, many of current cloud providers migrate to green cloud data centers (GCDCs), and seek to reduce the usage of brown energy by partially (or entirely) adopting renewable energy sources. However, the temporal diversity in the grid price, wind speed and solar irradiance makes it a big challenge to minimize the grid energy cost of a GCDC while meeting the performance of each delay bounded request. This work proposes a Temporal Request Scheduling algorithm (TRS) that jointly considers the temporal diversity. TRS considers the long tail in real-life requests' delay, and can provide strict delay assurance to all arriving requests by scheduling them to execute within their delay bound. Besides, this work explicitly provides mathematical modeling of the relation between the service rate in a GCDC and the refusal of delay bounded requests. Specifically, TRS solves a constrained nonlinear optimization problem by a hybrid meta-heuristic in each of its iterations. Compared with some existing scheduling methods, TRS can achieve higher throughput and lower grid energy cost for a GCDC while meeting each request's delay requirement.

1. Introduction

Recently, a growing number of Internet services, e.g., social networking and e-commerce, are running in large-scale cloud infrastructure in a cost effective way [3]. However, the usage of electric power in cloud data centers, e.g., Amazon and Google, becomes an increasingly important concern to cloud providers in recent years. In 2011, the electric power consumed by these centers comprises roughly 3% of the total power consumption in U.S., and the percentage may reach 15% in the future [13]. Therefore, such power consumption significantly increases the cost of cloud data centers. Besides, the usage of brown energy leads to severe damages to the environment. Over 57% of the electricity power in the U.S. are produced with coal in 2009 [11]. Therefore, an increasing number of large green cloud data centers (GCDCs) (e.g., Google, Facebook, Microsoft) intend to reduce the usage of brown energy by adopting renewable energy sources, e.g., wind and solar energy.

* Corresponding author. Tel.: +8618813119708.
E-mail address: cityu.yuan@gmail.com (H. Yuan).

http://dx.doi.org/10.1016/j.ins.2016.04.024
0020-0255/

In this way, their carbon footprint can be reduced. For example, to produce solar energy, Apple recently has built a solar farm close to the existing iCloud data center.

Typically, a GCDC aims to serve all users' requests in a cost-effective way while guaranteeing a user-specified delay bound. Similar to [22,44], this work focuses on delay bounded requests that have a relaxed and relatively long delay bound. Typical requests in a cloud include massive-scale data analysis [7], scientific computing [17], etc. During the delay bound of users' requests, multiple factors show the temporal diversity. In the real-life market, the price of grid energy often varies during the delay bound of requests. In addition, wind speed and solar irradiance change with time [8]. Therefore, temporal diversity in these factors brings a big challenge of how to minimize the cost of grid energy consumed by a GCDC while guaranteeing a given delay bound for users' requests.

Growing efforts are made to reduce the grid energy cost of a GCDC [6,18,22,41,42]. Some papers explore the temporal diversity in electricity price to reduce the energy cost of data centers [22,41]. Other papers adopt the geographical diversity in electricity price of different regions to decrease energy cost of distributed data centers [18,42]. Different from such early studies, this work achieves the minimization of the grid energy cost by jointly considering the temporal variation of grid price, wind speed, and solar irradiance during the delay bound of requests. Besides, most of existing scheduling methods can only meet the average delay bound of all arriving requests [6]. However, the long tail in the delay of real-life requests makes it possible that the delay performance of some requests is not guaranteed [43]. Therefore, different from prior methods, this work seeks to provide delay assurance for all arriving requests by proposing a Temporal Request Scheduling (TRS) algorithm.

To avoid overload of data centers, many existing papers selectively admit arriving requests, and therefore choose to refuse excessive requests, e.g., in [22,39]. However, these papers do not provide an explicit analysis of the relation between the service rate in a data center and the refusal of delay bounded requests. Different from these papers, this work focuses on delay bounded requests, and explicitly investigates the mathematical modeling of this relation. Specifically, this work formulates the grid energy cost minimization as a constrained nonlinear optimization problem, and solves it with the proposed TRS based on the combination of typical meta-heuristics. It can provide strict delay assurance for each arriving request by smartly scheduling all requests to execute within their corresponding delay bound. The proposed TRS is evaluated by trace-driven simulation based on experimental data including the realistic trace from 1998 World Cup website [2], the grid price and renewable energy resources. Extensive simulation results are presented to assess the effectiveness of TRS, and show that it outperforms some existing request scheduling methods in terms of grid energy cost and throughput.

The remainder of the work is organized as follows. Section 2 gives a brief discussion of related work. Section 3 presents the TRS framework in a GCDC. Based on it, Section 4 gives the formulation of a problem in a GCDC. Then, Section 5 presents the design of the TRS. Section 6 provides the performance evaluation of the proposed TRS based on trace-driven simulation. Section 7 concludes this work and discusses future directions.

2. Related work

This section gives a summary of some previous studies related to the research issue in this work, and reveals the differences between this work and them.

2.1. Request scheduling

Request scheduling in data centers is a challenging topic that was studied in the past [9,21,37,47,48]. In [9], to mitigate interference effects for concurrent data-intensive applications, a framework that adopts control and modeling methods based on statistical machine learning is presented to significantly enhance the system performance. In [21], the problem of task scheduling in the mobile cloud computing is studied. To reduce energy consumption, its proposed algorithm performs task migration and dynamic voltage and frequency scaling after minimal-delay scheduling. In [37], a service-oriented workflow scheduling algorithm that applies a hybrid metric based on recommendation trust and direct trust is proposed to tackle the challenge brought by the unreliability and uncertainty of workflow scheduling in a cloud. In [47], to realize the suboptimal scheduling of multitask jobs over a long period, a method of iterative ordinal optimization is presented and adopted in every iteration. In [48], to lower the execution time of tasks, a fine-grained MapReduce scheduler that divides each task into several phases is introduced. This scheduler considers the high variation in resource requirements of tasks, and schedules tasks in each phase that has a specific profile of resource usage. However, request scheduling methods in these studies do not consider the cost minimization problem of data centers.

2.2. Green cloud

Recently, several studies to focus on the application of widely available renewable energy in large-scale cloud [1,11,13,14,20]. The work [1] investigates the adoption of green energy in a cloud, and guarantees that carbon emissions of cloud providers cannot exceed a predefined bound. The proposed framework of resource management aims to reduce operational cost by allocating resources across geo-distributed data centers. The work [11] designs a novel online algorithm to jointly perform load scheduling and power management for distributed clouds under highly dynamic user demand. Based on the theory of Lyapunov optimization, the algorithm aims to minimize the eco-aware power cost of cloud providers provided that requests' quality of service (QoS) is ensured. The work [13] proposes a convex optimization-based strategy to maximize