Future Generation Computer Systems 74 (2017) 1–11

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Efficient task scheduling for budget constrained parallel applications on heterogeneous cloud computing systems

Weihong Chen [a,b], Guoqi Xie [a,b,*], Renfa Li [a,b], Yang Bai [a,b], Chunnian Fan [a,c], Keqin Li [a,d]

[a] College of Information Science and Engineering, Hunan University, Changsha, Hunan 410008, China
[b] The National Supercomputing Center in Changsha, Changsha, Hunan 410008, China
[c] Nanjing University of Information Science and Technology, Nanjing, Jiangsu 410008, China
[d] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

HIGHLIGHTS

- We convert the budget constraint of an application into tasks using the budget level.
- We propose the MSLBL algorithm with low-time complexity.
- We validate that MSLBL performs better than existing algorithms under different conditions.
- We propose the algorithm called minimizing the schedule length using the budget level (MSLBL).
- MSLBL can generate less schedule lengths than existing algorithm under different conditions.

ARTICLE INFO

Article history:
Received 22 October 2016
Accepted 28 February 2017
Available online 5 April 2017

Keywords:
Budget constraint
Heterogeneous clouds
Parallel application
Schedule length

ABSTRACT

As the cost-driven public cloud services emerge, budget constraint is one of the primary design issues in large-scale scientific applications executed on heterogeneous cloud computing systems. Minimizing the schedule length while satisfying the budget constraint of an application is one of the most important quality of service requirements for cloud providers. A directed acyclic graph (DAG) can be used to describe an application consisted of multiple tasks with precedence constrains. Previous DAG scheduling methods tried to presuppose the minimum cost assignment for each task to minimize the schedule length of budget constrained applications on heterogeneous cloud computing systems. However, our analysis revealed that the preassignment of tasks with the minimum cost does not necessarily lead to the minimization of the schedule length. In this study, we propose an efficient algorithm of minimizing the schedule length using the budget level (MSLBL) to select processors for satisfying the budget constraint and minimizing the schedule length of an application. Such problem is decomposed into two sub-problems, namely, satisfying the budget constraint and minimizing the schedule length. The first sub-problem is solved by transferring the budget constraint of the application to that of each task, and the second sub-problem is solved by heuristically scheduling each task with low-time complexity. Experimental results on several real parallel applications validate that the proposed MSLBL algorithm can obtain shorter schedule lengths while satisfying the budget constraint of an application than existing methods in various situations.

1. Introduction

1.1. Background

Today's large-scale scientific applications, such as e-commerce, automotive control, and traffic state predication, have drawn a great amount of demand for the design of high performance computing systems [1]. Such applications comprised of many interdependent modules are usually executed in heterogeneous parallel and distributed environments. Computing grids have been used by researchers from various areas of science to execute complex scientific applications [2]. With the emergence of cloud computing and rapid development of cloud infrastructures, more and more scientific computing applications have been migrated to the cloud, on which a pay-as-you-go paradigm is established and on-demand computational services with difference performance and quality of service (QoS) levels can be offered [3]. In this computing model, users pay only for what they use. Accordingly,

* Correspondence to: College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China.
E-mail addresses: paney@126.com (W. Chen), xgqman@hnu.edu.cn (G. Xie), lirenfa@hnu.edu.cn (R. Li), by1990310@qq.com (Y. Bai), fcn@nuist.edu.cn (C. Fan), lik@newpaltz.edu (K. Li).

http://dx.doi.org/10.1016/j.future.2017.03.008
0167-739X/

time and cost become two of the most important factors cared by users. Typically, the execution speed of the powerful resource is directly proportional to the unit price [4]. Thus, the trade-off between time and cost is the key to DAG scheduling. In this paper, we aim at researching the scheduling of the budget constrained application such that its schedule length is minimized.

1.2. Motivation

The scheduling problem of minimizing the schedule length of budget constrained parallel applications has become challenging. In cloud computing systems, service providers and customers are the two types of roles with conflicting requirements by the server-level agreement (SLA) [5]. For service providers, minimizing the schedule length of an application is one of the most important concerns. For customers, the budget constraints of an application is one of the most important QoS requirements. Many studies have been conducted recently to minimize the schedule length while satisfying budget constraints [6–13]. As heterogeneous systems scale up, distributed and parallel applications with precedence-constrained tasks, i.e., large-scale scientific workflows, are represented by directed acyclic graphs (DAGs), in which the nodes represent the tasks and the edges represent the communication messages between tasks [14,15].

The problem of minimizing the schedule length of a budget constrained application with precedence-constrained tasks has been solved recently in a number of studies [4,10]. However, these studies focus on homogeneous cloud system only. The same problem has been studied for heterogeneous cloud computing systems based on the earliest finish time (EFT). Arabnejad et al. [11] presented the heterogeneous budget constrained scheduling (HBCS) algorithm by presupposing tasks with the minimum execution cost to satisfy the budget constraint of the application, then the spare budget of the application was used by prioritized tasks in turn. Although a quantized measurement is used in HBCS, its limitations still exist as follows.

(1) Prioritized tasks select processors with the highest worthiness value by a series of calculations, which results in tasks with higher priority having more chance to use the spare budget of the application. It is unfairness for tasks with lower priority.

(2) Preassigning the minimum execution cost to each task is not always effective with respect to schedule length minimization if there is not enough budget for executing an application. The tasks that have low priority and not enough budget available tend to select the processor with the minimum execution cost, which will lead to longer schedule length.

1.3. Our contributions

In this paper, we focus on the scheduling problem of budget constrained applications on heterogeneous cloud computing systems, and a fair scheduling algorithm with low-time complexity using the budget level is proposed. The objective is to minimize the schedule length of applications under the budget constraint. The contributions of this study are as follows:

(1) The problem of minimizing the schedule length of a budget constrained parallel application in heterogeneous cloud environments is decomposed into two sub-problems, namely, satisfying the budget constraint and minimizing the schedule length. We illustrate the proof of converting the budget constraint of an application into budget constraints of tasks using the budget level.

(2) The algorithm of minimizing the schedule length using the budget level (MSLBL) is proposed, which minimizes the schedule length of budget constrained parallel applications by preassigning tasks with the budget level cost while not violating the precedence constraints between tasks and budget constraint of the application. It has high-performance and low-time complexity.

(3) We do extensively experiments with both Fast Fourier transform and Gaussian elimination parallel applications. Experimental results validate that the proposed MSLBL algorithm can generated less schedule lengths than the state-of-the-art algorithm under different budget constrained and scale conditions.

The rest of this paper is organized as follows. Section 2 reviews related studies. Section 3 presents related models and preliminaries. Section 4 analyzes the existing algorithm and presents the MSLBL algorithm. Section 5 presents the verification of the MSLBL algorithm. Section 6 concludes this study.

2. Related works

Many studies have investigated the DAG scheduling problem in various computing environments. The QoS-aware scheduling considers the optimization of parameters, such as time and cost, using cloud computing to execute workflows. Hu et al. [14] formulated the task scheduling on parallel processors as a DAG scheduling problem, where a heuristic algorithm was proposed to minimize the schedule length (makespan) of a DAG for a bounded number of processors. The author also proved that the optimal makespan can be obtained and a lower bound on the minimum makespan can be determined for a DAG with arbitrary dependency constraints. Heuristic algorithms are widely accepted because DAG scheduling is an NP-complete problem [15]. Scheduling problems are extended to many forms of constraints and environment settings [16–19]. For time-critical parallel applications, the IaaS Cloud Partial Critical Paths (IC-PCP) method for deadline-constrained applications has been proposed on heterogeneous cloud environments [16]. In [20–22], an extensive study on the scheduling algorithms of grid computing was presented toward performance and makespan minimization. However, financial cost is another important parameter in the cloud. For cost-critical parallel applications, cost-aware scheduling algorithms have been proposed for minimizing execution cost or satisfying the budget constraint on heterogeneous systems [1,23]. However, very few papers have targeted heterogeneous cloud computing environment and designs for minimizing the schedule length of budget constrained applications. In [24], a heuristic algorithm of deadline early tree was presented. It minimized the cost of deadline constrained applications without considering the communication time between tasks. In [4], Wu et al. proposed a critical-greedy (CG) algorithm to minimize the end-to-end delay of budget constrained parallel applications. In this work, the CG algorithm defines a global budget level (GBL) parameter and preassigns tasks with the budget-level execution cost. However, this algorithm is for homogeneous cloud environments, where the communication time between tasks is assumed zero, which is not the practice on heterogeneous cloud computing systems.

In addition to single QoS parameter optimization scheduling, the scheduling problem becomes more challenging when two QoS parameters (i.e., time and cost) are considered simultaneously [12,25–27]. Workflow scheduling to satisfy multiple QoS parameters is an attractive area in cloud computing. Malawski et al. [13] proposed DPDS, WA-DPDS, and SPSS algorithms for workflow ensembles on clouds to satisfy budget and deadline constraints, but their aim was to maximize the number of user-prioritized workflows. Arabnejad et al. [12] proposed a deadline-budget constrained scheduling algorithm (DBCS), which aimed to find a feasible schedule within the budget and deadline constraints. In this work, DBCS transfers the deadline and budget constraints of the application to that of each task by defining the CL and DL of each task. The obtained scheduling may or may not satisfy the deadline constraint while satisfying the budget constraint of the application.