EAI Endorsed Transactions
on Cloud Systems
Review Article EAI.EU

Task scheduling in cloud computing based on meta-heuristic techniques: A review paper

Rasha A. Al-Arasi[1,*], Anwar Saif[2]

[1]Sana'a University, Department of Computer Science, Sana'a, Yemen
[2]Sana'a University, Department of Information Systems, Sana'a, Yemen

Abstract

Cloud computing delivers computing resources like software and hardware as a service to the users through a network. Due to the scale of the modern datacentres and their dynamic resources provisioning nature, we need efficient scheduling techniques to manage these resources. The main objective of scheduling is to assign tasks to adequate resources in order to achieve one or more optimization criteria. Scheduling is a challenging issue in the cloud environment, therefore many researchers have attempted to explore an optimal solution for task scheduling in the cloud environment. They have shown that traditional scheduling is not efficient in solving this problem and produce an optimal solution with polynomial time in the cloud environment. However, they introduced sub-optimal solutions within a short period of time. Meta-heuristic techniques have provided near-optimal or optimal solutions within an acceptable time for such problems. In this work, we have introduced the major concepts of resource scheduling and provided a comparative analysis of many task scheduling techniques based on different optimization criteria.

Keywords: Cloud computing, Resource scheduling, Optimization criteria, Scheduling, Meta-heuristic techniques, Task scheduling.

Received on 10 September 2019, accepted on 16 January 2020, published on 30 January 2020

doi: 10.4108/_____

*Corresponding author. rasha.ali66@gmail.com

1. Introduction

Cloud Computing (CC) is the latest technology with a fast outgrowth in the field of distributed computing. It confers the users with high reliability, security, scalability, cost-effective mechanism, group collaboration and ease of access to various applications and resources [1]. It is a model for enabling appropriate, on-demand provisioning of computing resources such as software, hardware, applications, and services that can be fast provisioned and freed with least management overhead or interaction from service providers [2]. Cloud computing offers three primary types of service models namely Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) [3]. Cloud computing can be implemented as a layered architecture and comes in four main development models namely public, private, community, and hybrid clouds [4].

The major concept used in cloud computing is virtualization. Virtualization is a technique by which the user can easily access the computing resources without considering the complexity and internal details of the system [5]. It enables the user to create Virtual Machines (VMs) on physical servers [6], which leads to reducing the required hardware equipment and improving physical resources utilization in cloud computing. There are several advantages provided by clouds to cloud users and the service providers, the major advantages of cloud computing are described in [5, 7-10] and listed below:

- Reducing the cost by providing computing resources on-demand based on a pay-as-you-go system.

utilization of the available resources and minimize the total execution time. Assume that the number of tasks is more than the number of available resources ($n > m$), and tasks are not allowed to migrate between resources [44]. To formulate the problem, consider the set of tasks defined as T_i={1,2,…n where n is the number of independent tasks and R_j={1,2,…m} where m is the number of computational resources. Therefore, cloud resource scheduling problem is to get an optimal mapping (OM) of tasks (T_i) to resources (R_j) OM: $T_i \rightarrow R_j$. The definition of this problem is depicted in Figure 2, where, two or more tasks may share one resource [45].

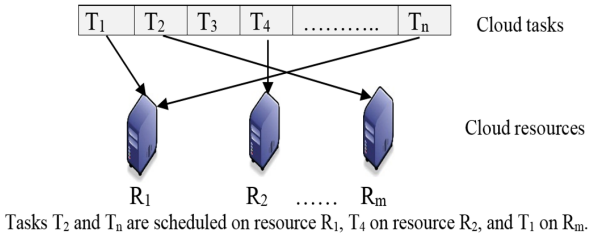

Figure 2. Cloud Resource Scheduling Problem

3.4. Optimization Criteria

This section explains the parameters used to measure the effectiveness of scheduling. The existing works have addressed different kinds of optimization criteria such as makespan, cost, budget, deadline, resource utilization, throughput, load balancing, and energy efficiency.
Generally, these optimization criteria are categorized into two desires based on cloud service: cloud users desire and cloud service providers desire, figure 3 [47]. These optimization criteria are addressed from most of the reviewed works, thus this work tries to demonstrate the way these criteria are studied in a comparative method.

3.4.1. User Desire Criteria

- **Makespan (completion time):** Makespan is defined as the completion time of the last task that is required to complete and leave the cloud system[48].
- **Cost:** cost is the total amount the user pay to a service provider on the basis of their resource usage[49].
- **Budget:** it indicates the constraints on completing the tasks within the budget[50].
- **Deadline:** it represents the termination of running tasks at a certain time[51].

3.4.2. Provider Desire Criteria

- **Resource utilization:** making the most of the available resources and keep resources as busy as possible. It is useful for service providers to get gain by leasing the finite resources to the cloud user on-demand [52].
- **Throughput:** it measures the number of completed tasks per unit time[53].
- **Load balancing:** load balancing in cloud computing is the distributions of loads evenly between the VMs over physical resources. Many techniques have been introduced by the authors in [54-56].
- **Energy efficiency:** energy efficiency can be defined as a reduction of energy consumed by a task [57].

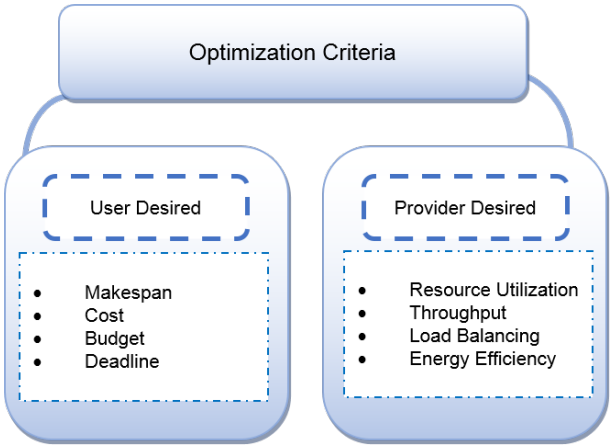

Figure 3. Optimization Criteria

Optimization problems can be divided into discrete and continuous problems. The decision variables for a combinatorial problem have discrete values; while the decision variables for a continuous optimization problem can take up values within the domain of real values (R_i) [58, 59]. According to the number of criteria involved in the optimization problem, this can be divided into single-criterion and multicriteria. The task of single- criterion optimization is to find the optimal solution according to only one criterion function. When the optimization problem involves more than one criteria function, the task is to find one or more optimal solutions regarding each criterion. Here, a solution which is good with respect to one criterion can be worse for another, and vice versa [60]. Therefore, the goal of multi-criteria optimization is to find a set of solutions that are optimal with respect to all other criteria. Noticeable, most real-world problems are multi-criteria. Nowadays, there exist optimization techniques that search for solutions by using Meta-heuristic and heuristic based search techniques. Stochastic and deterministic search principles are applied in these techniques. If an algorithm successfully solves all instances of problem (P), then we can say that it is capable of solving that problem. Usually, we are interested in which technique solves the problem more efficiently. Normally, the term efficiency is connected with the resources of the computer (space and time) that are occupied by running a technique [61, 62]. Generally, the most efficient technique is the one that finds the solution to the problem in the fastest way. In practice, the time complexity of an algorithm is not measured by the effective time necessary for solving the problem on a concrete computer because this measurement suffers from a lack of criteria. The same algorithm could be run on