Computers and Electrical Engineering 67 (2018) 551–565

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

An Integer Linear Programming model and Adaptive Genetic Algorithm approach to minimize energy consumption of Cloud computing data centers☆

Huda Ibrahim, Raafat O. Aburukba*, Khaled El-Fakih

Department of Computer Science and Engineering, American University of Sharjah, Sharjah 26666, UAE

ARTICLE INFO

Article history:
Received 30 June 2017
Revised 17 February 2018
Accepted 19 February 2018
Available online 1 March 2018

Keywords:
Cloud computing
Task scheduling
Optimization
Integer Linear Programming
Energy consumption
Genetic algorithm
Cloud data centers

ABSTRACT

Cloud computing infrastructures are designed to support the accessibility and availability of various services to consumers over the Internet. Data centers hosting Cloud applications consume massive amount of power, contributing to high carbon footprints to the environment. Hence, solutions are needed to minimize the energy consumption. This paper focuses on the development of a dynamic task scheduling algorithm by proposing an Integer Linear Programming (ILP) model that minimizes the energy consumption in a Cloud data center. Furthermore, an Adaptive Genetic Algorithm (GA) is proposed to reflect the dynamic nature of the Cloud environment and to provide a near optimal scheduling solution that minimizes the energy consumption. The proposed adaptive GA is validated by simulating the Cloud infrastructure and conducting a set of performance and quality evaluation study in this environment. The results demonstrate that the proposed solution offers performance gains with regards to response time and in reducing energy consumption.

1. Introduction

Cloud computing has been emerging as a successful paradigm for providing computing services to consumers. The availability of low cost computers, servers, storage devices, and high capacity networks motivated Cloud providers to expose underutilized resources as a utility to consumers over the Internet in a pay as you go manner. Those Cloud providers ensure the ultimate use of the delivered services while guaranteeing high quality of service and customer satisfaction. The work in [1] showed that up to 20% savings can be achieved on the energy consumptions of data centers. These savings lead to an additional 30% saving on cooling energy requirements.

Cloud computing makes use of the virtualization technology to achieve better resource utilization and gives the ability to dynamically consolidate Virtual Machines (VM) and live VM migration over the compute resources. Some techniques such as demand projection, heat management and temperature-aware allocation, dynamic power management by shutting servers down when they are not in use, Dynamic Voltage and Frequency Scaling (DVFS) to minimize the power level of physical compute resources, load balancing and task scheduling were reported in the literature to minimize the energy consumption in Cloud environments [2,3].

This work focuses on the allocation of Cloud consumers' requests to the capable resources within the Cloud data center at a specific time, where the consumers' requirements are met and the overall power consumption over time is mini-

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Hong Shen.

* Corresponding author.
E-mail address: raburukba@aus.edu (R.O. Aburukba).

https://doi.org/10.1016/j.compeleceng.2018.02.028
0045-7906/

mized. Effective scheduling is a key challenge for performance and quality driven requirements of Cloud computing requests. Scheduling tasks to physical/virtual resources in the Cloud data center involves considering the incoming task workload with their requirements. The total capacity of data centers consists of several tasks, and each task includes several VMs. Those VMs are to be allocated to the physical/virtual resources at specific time. In addition to the power consumption of compute and storage, networking resources have an impact for both performance and power consumption when allocating the tasks to the physical resources. Network communication increases the task execution latency and power consumption. One way to alleviate the network power usage within a Cloud data center is by applying a traffic aware VM placement methods [4,5].

This work aims at developing a model and an algorithm for minimizing the energy consumption in a Cloud computing Infrastructure by scheduling the incoming tasks to resources in such a way that consumer demands are met and the Cloud data centers energy consumption is minimized. The environment setting focus of this work is on a single Cloud data center. The main contributions of this paper are as follows:

- Formulation of an Integer Linear Programming (ILP) model for minimizing the energy consumption for a Cloud computing data center infrastructure.
- Proposing an Adaptive Genetic Algorithm (GA) for dynamic task scheduling in the Cloud data center.
- Simulation of Cloud computing infrastructure, testing, and analysing the quality and performance of the proposed algorithm.

This paper is organized as follows: Section 2 presents the literature review on task scheduling and optimization of energy consumption in Cloud computing, Section 3 presents the formulation of the ILP model, Section 4 illustrates the proposed Adaptive Genetic Algorithm, Section 5 presents the Genetic Algorithm Parameters. Section 6 illustrates the experimentation and results, and Section 7 presents the conclusion and future work.

2. Related work

Many researchers proposed algorithms for task scheduling in Cloud computing with different parameters and characteristics to achieve a specific goal. In order to minimize the makespan, the work in [6] proposed a task scheduling strategy with the assumption that multiple independent tasks form a single job. In this strategy, the data center is composed of clusters of heterogeneous computation servers. In [7], an Artificial Bee Colony algorithm to minimize the makespan and waiting time is used. Each task has a number of operations and a known sequence among them and can be processed on one machine. The Cloud nodes process one operation at a time. A random solution is generated for each bee in the initialization phase and then the bees search for a better solution around the randomly selected one till they find the best solution based on the fitness value.

Multi objective task scheduling algorithms were investigated in [8,9] to improve data center throughput and reduce cost without violating the Service Level Agreement (SLA). In both work the Cloud is made of multiple service providers and each service provider manages multiple data centers. A Cloud broker is responsible for scheduling the received tasks to the allocated VMs in order to minimize the execution time. Each VM has an ID and Millions of Instruction Per Second (MIPS) value while each task has an ID, a Quality of Service (QoS) value and a size. High priority tasks are be assigned to high QoS VMs and this result in better execution time and throughput in comparison to the First Come First Serve (FCFS) and the priority algorithm.

Genetic Algorithm (GA) has been extensively used for finding the optimal solution to the task scheduling problem in Cloud computing. The work in [10] used Tournament Selection Genetic Algorithm to minimize the completion time and maximize resource utilization. Tournament selection is used to choose the parents for the generation of new children. If there is a good solution but it is not selected in the crossover process it will not be removed from the list but it will be reserved to the next selection iteration.

For energy consumption in Cloud environment, researchers proposed different ways to reduce the amount of energy consumed in data centers and to minimize carbon emission levels without compromising the quality of service. In [11], the work proposed a most-efficient-server-first algorithm with a central scheduler to allocate tasks to servers in order to minimize servers processing time which in turn will minimize the total energy. The authors formulated the problem as an ILP for minimizing data center power consumption with respect to task response time and the number of servers required for processing a task in a specific time frame.

The work in [12,13], used the DVFS techniques to reduce energy consumption using different algorithms. This technique sets applicable supplies of voltages and frequencies to servers according to their resource requirements. The work in [14], proposed an algorithm that uses DVFS to consolidate the load to maximize resource utilization, and minimize energy consumption. The algorithm starts by calculating the instruction length for each task and high computing tasks are scheduled to resources with maximum frequency, followed by medium computing tasks to resources with medium frequency. The scheduling within the resource is based on FCFS and if one of the tasks in medium or low frequency resources fails it will be rescheduled to the high frequency resource. A multi-objective task scheduling algorithm that uses the DVFS called Green Task Scheduling (GTS) was presented in [15] to minimize the makespan, cost, and energy consumption by minimizing the number of active servers. The work proposed a *cloud controller* that received service requests from consumers and sends the information about these requests to the GTS algorithm. The algorithm finds the proper schedule and sends it to the *green*