Job Scheduling for Cloud Computing Using Neural Networks

Mahmoud Maqableh¹, Huda Karajeh², Ra’ed (Moh’d Taisir) Masa’deh¹

¹Management Information Systems, Faculty of Business, The University of Jordan, Amman, Jordan
²Computer Information Systems, King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan

Email: maqableh@ju.edu.jo, h.karajeh@ju.edu.jo, r.masadeh@ju.edu.jo

Received 15 April 2014; revised 22 July 2014; accepted 20 August 2014

Copyright © 2014 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/

Abstract

Cloud computing aims to maximize the benefit of distributed resources and aggregate them to achieve higher throughput to solve large scale computation problems. In this technology, the customers rent the resources and only pay per use. Job scheduling is one of the biggest issues in cloud computing. Scheduling of users’ requests means how to allocate resources to these requests to finish the tasks in minimum time. The main task of job scheduling system is to find the best resources for user’s jobs, taking into consideration some statistics and dynamic parameters restrictions of users’ jobs. In this research, we introduce cloud computing, genetic algorithm and artificial neural networks, and then review the literature of cloud job scheduling. Many researchers in the literature tried to solve the cloud job scheduling using different techniques. Most of them use artificial intelligence techniques such as genetic algorithm and ant colony to solve the problem of job scheduling and to find the optimal distribution of resources. Unfortunately, there are still some problems in this research area. Therefore, we propose implementing artificial neural networks to optimize the job scheduling results in cloud as it can find new set of classifications not only search within the available set.

Keywords

Cloud Computing, Job Scheduling, Artificial Intelligence, Artificial Neural Networks

1. Introduction

Cloud computing is an emerging paradigm that accesses network and shares computing resources with convenient and minimal management efforts, see Figure 1. It is one of the smart technologies that will reshape the world and shifts Information Technology infrastructure to third party to be available to the customers as com-
The computing environment of cloud computing can be outsourced to another party to use the computing power or resources via Internet. Emerging of this technology moves the computing power and data from personal computer and portable devices into large data centers. End-users access and use all the services without knowing the physical location and the configuration of the system at the providers’ sides [3] [4].

This paper is organized as follows. First, the researchers will introduce and discuss cloud computing deployment, characteristics, models and advantages. Then, the researchers will briefly discuss and introduce genetic algorithm and artificial neural networks. After that, the cloud computing job scheduling techniques are explained in details. Subsequently, the researchers will provide a literature review of job scheduling in cloud computing. Finally, conclusion and future works are discussed.

2. Cloud Computing Deployment, Characteristics, Models and Advantages

Cloud computing has wide acceptance due to its characteristics such as fulfilled customization, portability, availability on demand and isolation [6]. Moreover, it attracts the users due to the reduction of the cost of the provided services and at the same time improving the outcome [7]. Companies that use cloud computing do not need to invest in new infrastructure and training your employee. Using cloud computing, Small and Medium Businesses (SMB) can access to the best applications and resources at very low cost [8]. In the Information Technology industry, cloud computing is growing very fast at the same time many concerns are growing about the environment safety [8]-[10].

There are four types of cloud computing deployment: public cloud, private cloud, hybrid cloud and community cloud. In the public cloud, the users access the cloud via interfaces using the web browsers. Thus, the user needs to pay only for the time duration of service usage. This will reduce the operation costs. On the other hand, public clouds are less secure compared to other clouds models, as all the software and data on this model are more vulnerable to various attacks [11]. In the private cloud, all the operations of this model are within an organization’s data centers. This model is similar to the Intranet. The main advantage is that it is easy to manage the security and the maintenance and upgrades are more controlled. Compared to the public cloud where all the services and the applications are located outside the organization, in private model these services and applications are available at the organization level [4].

The hybrid model is a combination of both public cloud and private cloud. In this model, a private cloud is linked to one or more external cloud services. It enables the organization to meet its need in the private cloud, if some occasional needs occur; it asks the public cloud for intensive computing resources. Finally, the community cloud occurs when many organization jointly construct and share the cloud infrastructure, the requirements and policies [4] [11].

According to M. Malathi, 2011, cloud computing system has many attractive characteristics. Cloud computing users can access the resources via the Internet regardless of the users’ location or the machine type at the minimum cost [4]. Its implementation and configuration are required the minimum skills. Moreover, using cloud
computing is reliable due to multiple sites service delivery. In addition, cloud computing resources utilization is efficient due to its sharing and scheduling between several customers. Cloud computing users do not need to concern about resources and system maintenance, which is being performed at the provided side. Finally, the security issues in cloud computing can be solved easier than the issues in the traditional systems that is being solved by specialized people and resources at the provider side using several traditional security methods such as encryption methods and Hash functions [10] [12]-[16].

Furthermore, cloud computing system consists of two main parts that are connected to each other via the Internet: front-end and back-end. The front-end is the part that the user see and has on the machine with the required applications to connect to the cloud computing. The back-end part is the cloud system with all resources and services such as software, servers and data storages. Three different cloud computing models have been proposed to gain its benefits: Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS), see Figure 2 [6]. Software as a Service (SaaS) model offers finished application to the end-users via the Internet. Thus, end-users do not need to install the programs and applications on their machines that are controlled and managed by centralized authority. Platform as a Service (PaaS) provides an operating system, programming languages and software development through the cloud infrastructure. Infrastructure as a Service (IaaS) provides the required infrastructure as a service such as processing, data centers and network resource.

Cloud computing aims to maximum the benefit of distributed resources, and aggregate them to be able solve large scale computation problems [4]. It provides the computing services for users as public utility, which is available to organizations and individual [9]. In this technology the customers do not have the physical infrastructure, but they use the resources as a service and only pay when they need to use a resource [4]. Service providers provide the services to the subscribers on contractual basis. They charge the subscribers according to the provided services. Users can pay for the provided service with the payment system “pay as you go” [18]. Thus, cost reduction is considered one of the main advantages of using cloud computing. Moreover, service providers guarantee the quality of the provided services such as data processing, data storage and data access.

Another advantage is the ease of management as the maintenance of the infrastructure (software or hardware) is simplified. Furthermore, the applications that needs huge amount of storage are easier to use in the cloud.
computing environment. At the user level, the user just only needs a web browser with Internet connection to use the cloud computing system [4]. Furthermore, the uninterrupted service of the reliability of the provided service is another advantage. Finally, in case of disaster, an offsite backup is always helpful. In the cloud computing system frequently backed up the data in case of disaster occurred [3].

3. Genetic Algorithm and Artificial Neural Networks

Genetic Algorithms (GAs) are search algorithms that mimic the processes of natural selection and natural genetics that use to find estimated solutions to difficult problems, see Figure 3 [19] [20]. The principles of genetic algorithm were introduced in 1962 by Holland [21]. Genetic algorithm population is competing with each other to evolving the best candidate for the problem solution, which will be selected based on the fitness function [20]. The main genetic algorithms steps are initial population, fitness function, selection, crossover and mutation [22] [23]. The initial population is composed of all the individuals that are used in the genetic algorithm to find out the optimal solution. Every single solution in the population is called as an individual. And every individual is called as a chromosome to make it suitable for the genetic algorithm. The individuals are selected from the initial population and some operations are applied on those to generate the next generation. The selection operation of mating chromosomes is based on some specific criteria.

A fitness function is used to measure the quality of the selected individuals from the population according to a specific optimization objective. The fitness function can be different, in some cases the fitness function can be based on maximization some factors while in other cases it can be based on minimization other factors. The mutation means that the values of some gene that is located in the chromosome code were replaced by the other gene values in order to generate a new individual in the population. In every generation, the population individuals are evaluated scouring to some defined quality measures. The chromosomes are selected based on the fitness value. Two parents’ genes are allowed to be exchanged to generate new children generation, which will replace their parents. The current population becomes the new population and the old generation is removed. Then, the current population is examined to find the solution suitability. This operation will be iterated a number of times or until the desired result is obtained. GAs has been adopted in many different disciplines to optimize problem the solution such as scheduling algorithms, game playing, cognitive modeling, and salesman problems. Genetic algorithms have been used in some aspects of neural networks design as it can find the optimal solution. GAs can be view as a way of job scheduling that is based on the biological concept of population generation [22].

Artificial Neural Network (ANN) is an information processing paradigm that simulates the human brain neural, see Figure 4 [25]. It designed to mimic the way that the human brain execute a specific task or function [20]. The adaptive nature of this network is consider one of the most important feature, where “learning by example” is used to solve the problems [19]. Thus, this model is used to solve complex or ambiguous systems problems, pattern classification and recognition. These problems would be very difficult to be extracted using many other computer techniques. ANN can give very great result when it used with complex systems that has not fully understandable relationships or chaotic properties [26]. ANN model has three main issues: network topology, transfer function, and training algorithm. ANN consists of processing units, weighted connections, activation rule, and learning rules. Neural network consist of three or more layers and each layer has number of processing unit that called neurons [27]. It has input layer, output layer and hidden layers. ANN link the input layers with the output layers using hidden layers with nonlinear transformation function and weighted connections. Artificial neural networks can have different number of layers and different number of nodes. The nature of the problem and the degree of complexity are controlled the number of hidden layers and their neurons. The nonlinear transformation functions give an advantage over the predictable functions.

ANN are trained using different learning rates, parameters and different propagation methods such as feedforward and back propagation [25]. It learns by changing the connections between the input and output layers. The ANN output accuracy is depending on the parameters that used to train the system. Networks performance is affected by number of layers, number of nodes and training algorithms. ANNs are trained by iterating the recombination, mutation and fitness selection until development of chromosomes with accurate ANN. After neural network has been trained in certain information collection, it can be used to predict new situation and to model various non-linear applications [20]. The suitable output is generated at the output layer at the end of the learning or training process. Better results can be achieved by using neural network architecture with proper selection of input variable and training set [26].
4. Cloud Computing Job Scheduling

As mentioned earlier, cloud computing is a new technology, and it becomes so popular because of its great characteristics. By using this technology, everything such as software, hardware and platform are provided as a service. The users of these services pay for every use of them. The cloud provider in cloud computing provides services based on the clients’ requests [22]. One of the biggest issues in cloud computing is job scheduling. It is a hot research area in cloud and grid computing. It plays the same role in cloud and grid computing. Job scheduling of users’ requests means how to allocate resources to these requests. Therefore, the required tasks can be finished in minimum time according to time defined in user request. The main task of job scheduling system is to find the best resources in a cloud for the cloud computing user’s jobs, taking into consideration some statistics and dynamic parameters restrictions of users’ jobs. Most researches that used in grid computing can be used in cloud computing environment [29].

However, scheduling in cloud computing can be divided into two main views: from the cloud computing users and from the cloud computing provider. From the user’s view, the scheduling algorithm should minimize both the execution time and user’s budget. On the other hand, from the cloud provider view, the scheduling algorithm should improve the resource utilization and reduce the cost of maintenance and energy consumption [30]. Job scheduling is a combinational problem. It cannot be considered as linear programming and it is impossible to find a global optimal solution by using a simple algorithm or rule. It is well known as NP-complete problem. In order to solve this problem, some kind of branch and bound and other approximation method are proposed, but the result is unpredictable and needs a lot of time that is not practical in cloud environment. Moreover, the goal of cloud computing maybe too complex and depends on the business orientation of cloud environment, by which it is impossible to be solved in linear time by using traditional scheduling algorithm [31]. Indeed, many researches in the literature tried to solve the issue of job scheduling in cloud computing. All of them share the same goal in mapping the user jobs onto a computing resource to achieve the maximum benefit, and satisfying the various quality of service (QoS) of user’s jobs is the main goal of cloud provider. In the following section, the researchers will discuss the literature review of job scheduling in cloud computing.
5. Literature Review of Job Scheduling in Cloud Computing

Recently, many researchers studied job scheduling in cloud computing [22] [29]-[41]. In [22], the authors discussed three scheduling algorithms: Min-Min, Max-Min and genetic algorithm. Further, they propose a new scheduling algorithm in which Min-Min and Max-Min can be combined in genetic algorithm. The Min-Min algorithm starts with a set of all unassigned tasks. Firstly, the minimum completion time for all tasks is calculated. Then, among these calculated minimum times, the minimum value is selected. After that, the task is scheduled on the corresponding machine. Then, the execution time for all other tasks is added to the execution time of the assigned task, and the assigned task is removed from the list. Then, again and again the same operation is repeated until all tasks are assigned on the resources. The Max-Min algorithm is approximated the same as Min-Min algorithm except for the following: after computing the minimum execution times, the maximum value is selected that is the maximum time between all the tasks on any resource. After that and according to maximum time, the task is scheduled on the corresponding machine. Then, the execution time of the assigned task is added to the execution time all other tasks on that machine, and the assigned task is popped out from the list. Then the same operation is repeated until all tasks are assigned on the resources.

[22] used the proportion selection operator to determine the probability of various individuals genetic to be chosen to the next generation in population. The proportional selection operator means the probability which is selected and genetic to next generation groups is proportional to the size of the individual’s fitness. They also used a single-point crossover operator. Single-point crossover means only one position was chosen in the individual code, at that point part of the pair of individual chromosomes is exchanged. The mutation means that the values of some gene that is located in the chromosome code were replaced by the other gene values in order to generate a new individual in the population. The authors in [22] proposed a new technique that is based on genetic algorithm which generates the initial population by using Min-Min and Max-Min can provide better initial population than if they choose the initial population randomly. The experimental results show that the improved genetic algorithm maximizes the utilization of the resources effectively than the original genetic algorithm. They used the makespan as a fitness function for checking the fitness of the scheduling results. The idea can be further extended in which they can use the execution cost of the resource as fitness criteria. This method can be modified in existing cloud computing systems for decreasing makespan and better resource utilization.

In [34] the authors proposed a cloud task scheduling policy based on Load Balancing Ant Colony Optimization (LBACO) algorithm. The main contribution of this algorithm is to balance between the entire system load while trying to minimizing the makespan of a given tasks set. The authors used the Cloud Sim toolkit package in order to simulate the new scheduling algorithm. The experimental results show that the proposed LBACO algorithm outperformed FCFS (First Come First Serve) and the basic ACO (Ant Colony Optimization).

In [31] the authors presented a genetic algorithm approach to cost based multi QoS job scheduling. The authors also proposed a model for cloud computing environment and some popular genetic cross over operators, like PMX, OX, CX and mutation operators, swap and insertion mutation are used to produce a better schedule. The algorithm guarantees the optimal solution in finite time. The experimental results show that this approach for job scheduling guarantees the QoS requirement of customer job, and also make best profit of cloud providers.

The authors in [35] presented private cloud characteristics that are used for e-Learning purposes along with a genetic algorithm that being used to optimize the scheduling of the e-Learning workloads according to a set of factors that are imposed by the underlying virtualization technology such as memory over-commitment and IOPS rate distribution. The experimental results show that the genetic algorithm is an efficient technique for enabling co-existence of Planned Scheduling Requests and One-Off Scheduling Requests, by enabling a high and uniform utilization of the cloud resources. Also, the solutions generated by the genetic algorithm generate the optimal co-scheduling of workloads based on the workload profile.

An Improved Differential Evolution Algorithm (IDEA) is proposed by Tsai et al. [41] to optimize task scheduling and resource allocation on cloud computing organization. The proposed algorithm improves the Differential Evolution Algorithm (DEA) by using Taguchi method to generate improved offspring. Two models are developed to minimize the total cost and the time in task scheduling. The processing and receiving cost are included in the cost model, while time model takes into account the receiving, processing, and waiting time. The effectiveness of the proposed algorithm is tested using two scenarios of the cloud environment which are; five-task five-resource scenario and ten-task ten-resources scenario. In both scenarios, the proposed algorithms
(IDEA) outperform the other scheduling algorithms in the literature (DEA/NSGA). Moreover, Gantt chart is used to show the efficiency of the proposed algorithm in task scheduling in term of having smaller cost and time. In addition, this approach can help the decision makers to choose the correct decision in case of object conflicting.

A new fully distributed scheduling framework for uncoordinated federated cloud environment is proposed by Palmerieri et al. [36]. This scheduling schema based on independent and self-organized agents which do not depend on any kind of centralized control that coverage towards Nash equilibrium solution with taking into account the possible contradiction between the client and the service provider interests in the cloud environment. An implicit coordination is forced by applying a marginal cost on agent behaviour. The effectiveness of the proposed schema is tested on the simulator for the cloud environment that emulates the service provider, agents, and the used protocols. The experimental results show that the proposed approach provides a good solution in terms of scalability and quality. In addition, it gains a high performance in according to the completion time. Due to the efficient partitioning strategy of the complex task into smaller one, this approach had great benefits in a very large cloud organization that have a lot of nodes with large number of tasks to be served.

Scheduling algorithms for highly available applications on cloud computing is proposed by Marc Frîncu [37]. This algorithm ensures the applications functionality despite the number of node failure. Two algorithms are proposed to achieve a highly available applications; optimal and sub-optimal algorithms. The optimal algorithm is proposed when the load of each component type is known while the suboptimal algorithm is proposed in case the load is unknown. By taking advantage of the component based architecture and the application scaling property, a highly available applications is build. A solution for determining the best number of component types on each node is presented. In addition, each node has a threshold of component load that cannot exceed it and the application running cost need to be minimized. The performance of the suboptimal algorithm is tested accordance to the node load, closeness to the optimal solution and the success rate.

A priority based job scheduling algorithm in cloud computing (PJSC) is proposed in [32] that is based on multiple criteria decision making model. This algorithm is based on the theory of Analytical Hierarchy Process (AHP) which is considered as a suitable method for priority based problem such scheduling a task with multi-attributes. The efficiency of the proposed algorithm is tested intern of consistency, complexity, and makespan. The experimental results show that the proposed algorithm has an acceptable complexity while needs more improvement to gain less makespan.

A taxonomy for cloud computing research is revealed based on an intensive literature survey for 205 journal article in cloud computing field [42]. These articles are classified into four main categories: technological issue, business issues, domain and application issues, and conceptualizing cloud computing. This study shows that the current state of cloud computing is skewed on technological issues. On the other hand, a new research issue is emerging that focus on social and organizational implications. This descriptive review is considered as a good reference to guide the practitioner and researchers on cloud computing for future research.

Scheduling algorithm is considered from one of the important issues on the cloud computing environment that enhance the workflow of the job tasks and improve the user satisfaction from the service provider. Because of that, a comprehensive survey on the different types of scheduling algorithm that is used on a cloud computing environment is presented by Vijindra and Shenai [38]. This study provides a detailed survey on the existing scheduling algorithms in cloud, grid, and workflows environment. Increasing number of parameters for scheduling algorithms may improve the framework for resource allocation and scheduling in cloud computing environment. Execution time, deadline, energy efficiency, transmission cost, performance issues, and makespan can be taken as an input for scheduling algorithm. The nature of the job such as size, availability of the resources and the environment decide which of the mentioned parameter will consider in the scheduling algorithm since considering all of them in one algorithm it will enter into the complexity problem.

A novel job scheduling algorithm in cloud computing environment based on Berger model of distributive justice is proposed by Xu et al. [33] through the expansion of cloudSim platform. In this algorithm, the scheduling is performed based on the fairness in resource allocation. Therefore, two fairness constraints are proposed to classify the user tasks. In the first constraint, the user’s tasks are classified based on QoS to create an expectation function that control the fairness in resource allocation. Then, in the second constraint, the fairness justice function is defined to assess the fairness of the resource allocation. The effectiveness of the proposed algorithm is tested on the extended simulation platform. The experimental results show that the proposed algorithm is effective in achieving the user task with better fairness.
A route scheduling algorithm for a cloud database is proposed by Yan-Hua et al. [39] based on a combination of genetic and ant colony algorithm. This is performed by taking the initial value of genetic algorithm as an input to the ant colony algorithm (after transforming it into the pheromone initial value) to find the optimal solution from ant colony algorithm. The experimental results show that the proposed algorithm improves the efficiency of cloud computing by finding the suitable application’s database quickly and effectively. A new scheduling algorithm is proposed by Mezmaz et al. [40] that based on a parallel bi-objective hybrid genetic algorithm that takes into account the makespan and the energy consumption. To minimize the energy consumption this algorithm uses the dynamic voltage scaling (DVS). The experimental results show that the proposed algorithm outperforms the other scheduling algorithm in term of time completion and energy consumption.

6. Discussion

In the literature, many researches tried to solve the problem of job scheduling in cloud computing using artificial intelligence techniques such as genetic algorithm and ant colony. Unfortunately, the proposed techniques have some problems. Neural Networks designed to mimic the way that the human brain executes a specific task or function. Its most important feature is the adaptive nature, where “learning by example” is used to solve complex or ambiguous systems problems, pattern classification and recognition. ANNs are trained using different learning rates, parameters and propagation methods. It learn by changing the connections between the input and output layers. Networks performance is affected by number of layers, number of nodes and training algorithms. ANNs are trained by iterating the recombination, mutation and fitness selection until developing chromosomes with accurate ANN. After neural network has been trained in certain information collection, it can be used to predict new situation. The suitable output is generated at the output layer at the end of the learning or training process. Better results can be achieved by using neural network architecture with proper selection of input variable and training set. Neural networks are widely used for identifications, classification, and prediction when a vast amount of information is available. By examining hundreds, neural network detects important relationships and patterns in information. The advantages of using neural networks are: learn and adjust to new cases on their own, lend them to massive parallel processing, function without complete or well-structured information and cope with huge volume of information with many dependant variables. Finally, neural network can learn to classify new input instantly that has not been seen before while the genetic algorithm finds acceptable solution within the solution space. Thus, the job scheduling result would be optimized using neural network by finding new set of classifications based on the provided tasks. Therefore, solving and optimizing the scheduling problems in cloud computing environment can be achieved using artificial neural networks.

7. Conclusions and Future Works

Could computing is considered one of the most important research areas that helps to get the maximum benefit of distributed resources, and aggregates them to achieve higher throughput and be able to solve large scale computation problems. Job scheduling is considered one of the main issues and hottest research topic in cloud computing. The main task of job scheduling system is to find the best resources in a cloud for the cloud computing user’s jobs. In this research, the researchers review the literature of cloud computing job scheduling. In the literature, many researches tried to solve the problem of job scheduling in cloud computing. Most of them use artificial intelligence techniques such as genetic algorithm and ant colony to solve the problem of job scheduling and to find the optimal distribution of resources. However, as there are still problems in this research area, the researchers propose a new technique that solves the issue of job scheduling in cloud computing environment. This technique is based on using neural network to classify the job queues that exist on any resource and to give priorities to different jobs. The artificial neural network is an artificial intelligence system that is able of finding and differentiating pattern. It can learn by example and can adjust to new concepts and knowledge. Using artificial neural networks will be highly potential to solve and optimize the scheduling problems in cloud computing environment.

References


http://dx.doi.org/10.1109/ICCEET.2012.6203873


http://dx.doi.org/10.1016/j.clay.2006.11.005


http://www.engineering.lancs.ac.uk/lureg/group_research/wave_energy_research/Collector_Shape_Design.php


