Job Scheduling Algorithms in the Cloud: A review

Somayeh Taherian Dehkordi, Vahid Khatibi Bardsiri

Department of Computer Engineering, Islamic Azad University, Kerman, Iran

E-mail: S.taherian2000@gmail.com

Department of Computer Engineering, Bardsir Branch, Islamic Azad University, Kerman, Iran

Email: Kvahid2@live.utm.my

Abstract

Cloud computing refers to services that are implemented in a distributed network and are available via common internet protocols. Cloud architecture integrates a large number of physical resources and provides them for the user as services under the service level agreements. Therefore, resource management along with task scheduling method has a direct impact on cloud network performance and productivity of its resources. Providing a suitable scheduling method can lead to resources utilization by reducing the response time of tasks and decreasing the costs. This article aims to investigate the current methods of task scheduling and allocation of resources in cloud infrastructure and to assess its advantages and shortcomings. Then, the challenges and open issues in the field of cloud computing scheduling will be discussed as the future research areas.

Keywords: cloud computing, task scheduling, resource allocation.

I. Introduction

Cloud computing is providing any software and hardware services in internet space that follows the rule of pay per usage. The main purpose of using this technology is to minimize cost and to maximize performance and efficacy. Preparation, timing, and failure management are required to implement the management, scheduling, and responding to demands in minimum time. The important issue here is how to manage virtual machines and resources along with tasks and demands scheduling policies. Suitable scheduling will be able to manage numerous requests and certain amount of resources which will lead to resources utilization. In this article, the most common task scheduling algorithms in cloud computing are examined and then the challenges and open issues in this area are introduced as the areas of research in future.

- **Cloud Computing**

Cloud computing refers to a type of distributed and parallel systems that include a set of connected virtual computers. In fact, cloud computing is a computing model which tries to facilitate the users' access based on the type of their request from data and computational resources. The model attempts to supply the users' demands with minimal need to manpower resources and reducing the costs and increasing the rate of access to information. In general, maximum performance and minimal cost are the most important reasons for choosing this technology (Andreadis et al., 2015). Cloud computing is a service model that provides production principles for delivering information technology, infrastructural components, architecture method, and the model based on economic and business principles. It is associated

with concepts such as virtualization, distributed computations, useful computations, hosting, and providing software as service. Virtualization, in general, refers to the simulation of computers and services within a physical computer. In cloud computing, the required services are provided through virtualization.

In this approach, several large servers can be simulated in one group and some groups of servers can be placed in the cloud so that they will be easy to manage. Service-oriented architecture is a new and evolving method for making distributed programs such as cloud computing which provides integrated services for the final user. The relationship between cloud computing and service-based architecture is that the former supplies information technology resources involving both software and hardware so that the user can use them whenever needed, such as the resources that host the data, services, and processes and the latter is a method which is dealing with the true formation of information systems using mechanisms that make them work together properly inside and outside the project. Service-based architecture can cause the decrease of costs, more agility in cloud environment, and the increase of reuse performance and efficiency (Durao et al., 2014).

- **Cloud Computing Characteristics**

Cloud computing is almost taking advantage of the features that other computing and distributed grids own, but the proper use of the features has made this network superior to the other ones.

Cloud computing owns six main characteristics as the following (Mittal and Soni, 2013):

- **Broad access to network**

Access to cloud resources is possible throughout the network and standard methods are used for the users to access the network.

- **Supplying service based on demand**

Users can have access to their required resources and software without having to interact with cloud computing service providers.

- **Calculating service (pay per usage)**

One of the key characteristics of cloud computing is calculating system based on the use of services and resources.

- **Density of resources**

There are massive amounts of resources in cloud computing which are independent of their physical location via virtualization.

- **Multiple users (tenants) (shared resources)**

It makes centralizing, increasing the use of unused resources, and sharing resources possible for the users.

- **Rapid expansion ability**

Resources in the cloud should be able to expand rapidly and should be unlimited and accessible at any time form the users' point of view.

- **Types of Cloud Computing**

Cloud computing is a design of software programs that uses services bases on the available demands through the internet. In fact, this technology is a set of services that appear in the form of capsules and owns an application program interface which is operable in the network and includes storage and processing services. Cloud processing is classified into different types based on service delivery model which supervises the specific type of services that can be provided by the cloud and the arrangement, and also development method that monitors the location and management of cloud infrastructure (Kumari and Srinivarsa, 2014).

- **Types of clouds based on service**

Service models in cloud computing are as layers which are linked together in such a way that each one can be developed independently. Cloud computing provides services based on three middle layers, i.e. infrastructure layer (structure), platform layer, and service-application layer which are known as cloud computing services. Figure (1) displays a variety of cloud computing based on using hardware and software resources. Three services presented in cloud computing are known as software, platform, and infrastructure

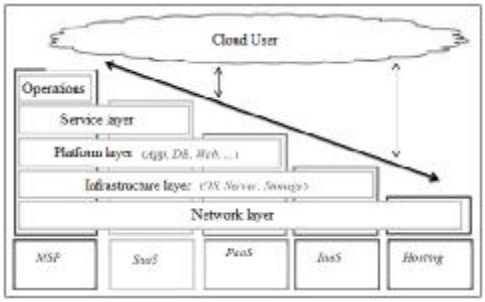

Figure. 1. layers of cloud computing and their relationship.

As shown in Figure (1), the cloud user just uses the services available in the cloud without any specific communication with different levels. In operating layer, the cycle of sending, managing, and responding to request is formed. Each operation is followed on the layer of existing services. On the other hand, each service is realized on a specific software bed. The required facilities and resources are provided for the implementation of the desired service in the infrastructure layer. All the needs and resources are realized in the network layer as hardware. Infrastructure approach as a service (IaaS) in cloud computing merely involves switches, servers, virtual storage space, and other kinds of applicable hardware. This approach is more restricted than the layer of platform as a service (PaaS), which is assisted by the layer of software bed in providing cloud services and which is the supplier of operating systems, frameworks, programmers, interactions development, and monitoring structures. Software as a service (SaaS) fully benefits from the existing facilities in service layer as well as the three layers of network, infrastructure, and software bed. In fact, software as the service provides the users with all kinds of software that they need through the internet.

- **Types of clouds based on arrangement**

Cloud computing is divided into four categories based on arrangement and interaction (Sultan, 2014):

A. Public Cloud

Public cloud describes cloud computing in its traditional and original sense. In fact, it is an open state for public access that is protected by firewall and is entirely hosted and managed by the providing company. Therefore, is has low security. This cloud is prepared for public use without much control over the infrastructure like services on the internet that are replacing a large industrial group whose owner is an organization selling cloud services. Public clouds are generally inexpensive and might be offered to test and develop new products of a company and private users and companies can have access to it.

B. Private Cloud

Private cloud is applied for exclusive use of an organization therefore it works within the specific range of an organization and owns all facilities of public cloud except that the organization itself

controls and manages it. Moreover, the maintenance method and location of infrastructural hardware of the cloud are different from public cloud so it is more secure. In fact, private cloud is a cloud computing infrastructure which is created by an organization for its internal usage. The main factors that separate private clouds from public clouds in trade are the location and maintenance method of infrastructural hardware of the cloud. Private cloud allows more control over all implementation levels of cloud. Another advantage of private clouds is more security which results from the placement of equipment within the boundaries of the organization and lack of communication with the outside world.

C. Hybrid Cloud

Hybrid cloud is a combination of public and private mode. Although public and private clouds keep their identities in this model, they act together as a unit.

D. Community Cloud

Community or group cloud is generated wherever several organizations have similar needs and seek to benefit from cloud computing advantages by sharing the infrastructure. Since the costs are divided among fewer users than public clouds, this option is more expensive than public cloud but it provides higher levels of privacy, security, and compatibility with policies. In other words, it is a community cloud. In fact, this type of cloud is prepared to serve a public action and several organizations with nearly common needs share their resources and services and constitute a group cloud.

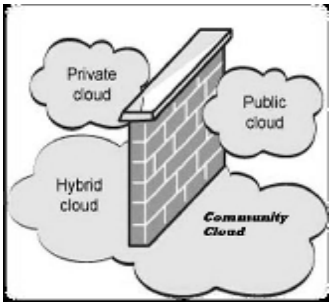

Figure. 2. Existing layers in cloud computing.

II. Scheduling

Cloud scheduling is done in three steps. First, the existing resources should be identified and data should be collected about them. The next step is selection. In this step, the desired resources are selected based on a series of main parameters of resources and tasks and finally the selected task is designed to the selected resource. Figure (3) displays scheduling steps in cloudy environment.

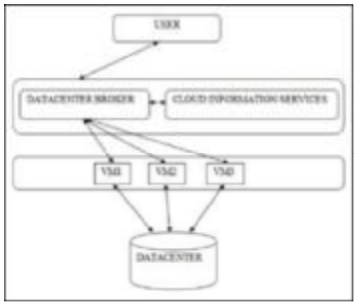

Figure. 2. Scheduling steps in cloudy environment.

- **Scheduling Levels**

Scheduling is broadly divided into two levels (Agnetis et al., 2014):

ü **User-Level Scheduling**

This step is dealing with providing service by servers for customers. In general, it is related to economic issues such as the balance between supply and demand, competition among customers, and relative minimization of customer cost.

ü **System-Level Scheduling**

This level is dealing with solving resource management problem within the data center. From a customer's point of view, a data center is an integrated system which offers services to the users. However, data centers are in fact a combination of many physical machines which provide integrated services. After receiving a large number of tasks from different users and allocating them to physical machines, the efficiency of a data center will be affected. In order to control and improve the system productivity several points such as implementing processes simultaneously, sharing resources, debugging, etc. should be considered. System level scheduling is done in two static and dynamic methods. Static methods are used when the characteristics of the set of tasks which should be scheduled such as tasks processing time of tasks, communications, data dependency, and synchronization requirements such as implementation are identified. Dynamic methods are used when there are very few data before running and the tasks are set on line.

- **Available Scheduling Algorithms**

Nowadays, increasing cloud efficiency is an important issue. Proper and optimal scheduling increases the performance and efficiency of cloud computing. Dynamic and changeable natures of cloud computing and different demands by the users have made the scheduling in cloud environment very complex. Optimal scheduling for the available tasks is one of the main challenges in heterogeneous computing systems particularly cloud computing. Task scheduling in cloud means to do N tasks on M sources. In this problem, the most efficient mode for doing N work on M machine should be considered, so that the final time of doing the whole task is minimized. The purpose of scheduling in cloud computing is giving turns to standby processes and are supposed to selected for implementation. The selecting method creates scheduling algorithms. Each algorithm has certain criteria to be selected. The purpose of most scheduling methods is to minimize the time of running workflow (Kaur and Kinger, 2014). Tasks scheduling algorithms in cloud computing can generally be classified in the following groups:

ü **Heuristic Scheduling Algorithms**

According to a simple classification, task scheduling algorithms in cloud environments can be classified into two main groups: batch heuristic scheduling algorithm and instant online heuristic scheduling algorithm (Paul and Francis, 2014). Batch heuristic scheduler is defined before tasks scheduling on a set of resources before it starts to run like heuristic scheduler of Min-min and Max-min. In instant heuristic scheduler the tasks are mapped in cloud resources as soon as they enter scheduling like First In, First Out (FIFO) method of service algorithm (Liu, 2013).

Some of the heuristic algorithms can be referred to as the following:

A. Min-min Algorithm

The algorithm was suggested by Braun et al. (2001) to be used in distributed grid systems and cloud computing system in order to minimize the resource allocated time to the task fulfillment. The algorithm aims to reduce the time spent to fulfil tasks with early allocation of resources to small tasks. At first, a set of minimum times is calculated for each task on the resource; then the task which requires less time to be fulfilled is selected in the time set and is allocated to the related machine. The selected task is removed from the set of tasks and by adding its running time to other tasks on the selected resource, the running time of the other tasks will be updated.

B. Max-min Algorithm

The algorithm was suggested by Ibarra in 1977. In this algorithm, at first, a set of minimum times is selected for each task on the resource. Then the task which requires maximum time to be fulfilled is selected in the time set and is allocated to the related machine. The selected task is removed from the set of tasks and by adding its running time to other tasks on the selected resource, the running time of the other tasks will be updated.

C. Random Algorithm

The idea of random algorithm is the random allocation of selected task to the available resource. In this algorithm no attention is paid to the resource situation to see whether its load is light or heavy. Therefore, a resource under heavy burden might be the result of selection. In this case the tasks must be waited for a long time before service allocation. The complexity of this algorithm is low and tasks do not have overhead or pre-processing.

D. Most Fit Task Scheduling Algorithm

This algorithm is in charge of selecting the best task in relation to the current server. The selection is made from the beginning of the tasks queue. This algorithm is one of the most famous methods in the second classification.

E. FIFO Algorithm

The idea of maintenance algorithm in order of arrival to the distributed networks was first proposed by Brent in 1989 and was used in 1998 by Schvigelshon in grid networks and parallel tasks scheduling in cloud computing. This algorithm assigns tasks in order of their arrival to the resource. Selecting criteria in this method is the entry time. In fact, the tasks receive services from the cloud server in the same order they enter the system.

F. Round Robin Algorithm

This algorithm is an unexclusive scheduler. In this algorithm, the scheduler allocates a fixed unit of time to each process which is called interrupt and then circulates among them. The combination of this algorithm and service algorithm is used in order of arrival in task scheduling in cloud environment. In fact, by achieving to the desired interval, the running task enters the ready queue and the next task is selected based on service algorithm in order of arrival.

G. Resource Aware Scheduling Algorithm (RASA)

It is a hybrid algorithm proposed by Saeed Parsa and Reza Maleki in 2009 and examined the distribution and scalability of distributed systems including grid and cloud computing. This algorithm is analyzed by the analysis of two Min-min and Max-min algorithms and has considered the advantages of Min-min and Max-min algorithms and had covered their disadvantages. This algorithm takes into consideration a set of tasks. If the amount of tasks is an even number it uses Max-min method in order to allocate resources to tasks and if the amount of tasks is an odd number it uses Min-min method.

H. Reliable Scheduling Distributed in Cloud Computing algorithm (RSDC)

Mehdi Javanmard et al. (2012) presented an algorithm of reliable scheduling distributed in cloud computing. In this algorithm, a new algorithm is developed with a new technique and by classifying and considering the sweep time of tasks in competence function. Through careful examination and evaluation of previous algorithms, the tasks have been scheduled by parameters which are associated with a failure rate. Therefore, in the proposed algorithm in addition to previous applied parameters some other remarkable parameters are used according to which different tasks scheduling can be obtained. The task is associated with a mechanism. The great task is divided into small tasks. In order to balance the tasks, their sweep time has been calculated separately. Then all tasks scheduling has been fulfilled in the form of a common task by considering the sweep time and finally the system efficiency and productivity have increased and also its real time has improved compared with the previous algorithm.

ü **Scheduling based on Economic models**

Economic scheduling is presented based on the tasks repetition. The main strategy of this algorithm is to repeat tasks on appropriate machines by minimum cost and the best time and includes the following groups (Nagadevi et al.,2013):

A. Economically Enhanced Resource Management Scheduling (EERM)

Economically enhanced resource management model was raised by Elmroth et al. (2005). This model is considered as a resource interface which establishes bilateral communication between resources and business layers in order to support appropriate decision making in the management of resources.

B. Cloud Bus Scheduling

This model was introduced by Boya according to market-based resource management strategies. It provides indirect access to virtual physical resources and distribution.

C. Open Provisioning and Execution Scheduling (OpenPEX)

This model was developed by Venugopal in 2009 based on resource reservation. This model is one of the supply systems which allocate virtual resources using an advanced reservation approach. In this system each user has a chance of booking any amount of resources without any time limit.

ü **Scheduling based on Evolutionary Algorithm**

In classic models, obtaining entirely optimal solution is quite time consuming and in many cases the random implementation of tasks requires more time. Conventional task scheduling algorithms in cloud computing do not consider all the desired elements of the user. Therefore, the use of evolutionary algorithms in cloud computing environment can relatively create more efficient and better response by considering the user satisfaction and service quality (Sajedi et al.,2014).

A. Genetic Algorithm (GA)

The idea of implementing genetic algorithm in cloud computing was first proposed by Braun in 2001. This algorithm aims to provide a solution close to the optimal solution in cloud computing space. It is used for optimal scheduling of tasks in cloud computing.

B. Particle Swarm Optimization Algorithm (PSO)

This algorithm is random group optimization which is inspired by simulating the social behavior of birds. The work is based on the principle that each particle in every moment adjusts its location in the search space according to the best place where it has been so far and the best place which exist in its all neighborhood.

C. Bee Algorithm (BA)

This algorithm simulates the behavior of groups of bees searching for food which was presented by Sung Chung et al. (2006) in order to be used in distributed networks and cloud computing.

D. Ant Colony Algorithm

This algorithm aims to allocate the presented tasks to resources based on their processing power. It is inspired with the social behavior of ants.

III. Conclusions

The main challenge in this cloud computing system is scheduling and providing better service to the applicants. In this paper, first of all, cloud computing technology and its various services were introduced. Then, with regard to scheduling nature and its importance in cloud computing efficiency, different classification of scheduling algorithm and common techniques of scheduling in cloud environment were reviewed. The results of the research show that batch statistic and

dynamic scheduling algorithms have the fastest processing time. Comparison of heuristic and deterministic algorithms showed that time complexity of heuristic algorithms was lower than unexclusive algorithms. In this regard, service algorithm in the order of entrance is very suitable for tasks with high preference. By examining task scheduling algorithms in cloud computing system it can be understood that heterogeneity, dynamicity, calculations, security and isolation of data are the primary challenges which have been widely investigated. It can also be concluded that by considering management of confidence and integration of policies, secure service management, authentication and identity management, hardware capabilities enhancement and evolution of cloud computing infrastructure, i.e. supporting complex application models, resources information, and task displacement principles, an opportunity will be provided to implement more complicated and more complete scheduling algorithms. In order to improve popular and classic scheduling in cloud computing, new methods like economic models and heuristic algorithms along with the algorithms inspired with nature such as ant colony algorithm, particle swarm optimization (PSO), and genetic algorithm can be used. By combining different approaches and considering input parameters such as running costs and deadline it is possible to provide a powerful approach for future tasks. According to the conducted studies, combining evolutionary algorithms and conventional scheduling algorithms is necessary in the field of cloud computing. Meta-heuristic methods and evolutionary algorithms besides the conventional scheduling algorithms can be used in cloud computing environment in order to obtain relatively optimal solution in a time interval and to achieve better results. Using evolutionary algorithms and noticing the advantages and disadvantages of conventional scheduling techniques will guarantee the balanced distribution of tasks between cloud computing resources and will produce less average response time.

References

i. Andreadis G, Fourtounis G, and Bouzakis KD. (2015). Collaborative design in the era of cloud computing. Advances in Engineering Software. Vol(81). pp 66-72.

ii. Durao F, Carvalho SFJ, Fonseka A, and Garcia CV. (2014). A systematic review on cloud computing . The Journal of Supercomputing, Springer US. Vol(68). pp 1321-1346.

iii. Mittal R, and Soni k. (2013). Analysis of Cloud Computing Architectures. International Journal of Advanced Research in Computer and Communication Engineering. Vol(2). pp 2087-2091.

iv. Kumari RC, and Srinivarsa KR. (2014). Services of Cloud Computing. International Journal of Advance Research in Computer Science and Management Studies. Vol(2). pp 343-347.

v. Sultan NA. (2014). Making use of cloud computing for healthcare provision: Opportunities and challenges. International Journal of Information Management. Vol(34). pp 177-184.

vi. Agnetis A, Billaut ChJ, Gawiejnowicz S, Pacciarelli D, Soukhal A. (2014). Multiagent Scheduling, Models and Algorithms, Springer Us.

vii. Kaur R, and Kinger S . (2014) . Analysis of Job Scheduling Algorithms in Cloud Computing . International Journal of Computer Trends and Technology (IJCTT) .Vol(9) . pp 379-386.

viii. Paul RA, Francis FS. (2014). Dynamic Scheduling of Requests Based on Impacting Parameters in Cloud Based Architectures. ICT and Critical Infrastructure journal : Advances in Intelligent Systems and Computing. pp 513-521.

ix. Liu J. (2013). Job Scheduling Model for Cloud Computing Based on Multi-Objective Genetic Algorithm, International Journal of Computer Science Issues.

x. Nagadevi S, Satyapriya K, Malathy D. (2013). A Survey on Economic cloud schedulers for optimized task scheduling. International Journal of Advanced Engineering Technology. Vol(5). pp 58-62.

xi. Sajedi H, Rabiee MA. (2014). Metaheuristic Algorithm for Job Scheduling in Grid Computing, Modern Education and Computer Science. pp 52-59.