Analysis of Some Job Scheduling Algorithms In Cloud Computing Environment

Rajendra T. Kaple
BNCOE, Sant Gadge Baba Amravati University, India
Corresponding author: Rajendra T. Kaple

ABSTRACT
Cloud Computing is the new IT term that makes the availability of computing resources (Hardware and Software), applications and data as a service over the internet to its clients. Cloud computing largely expected to offer consistent, vibrant and virtualized services in terms of resources for doing computation, storage, and data sharing. An important requirement in cloud computing is scheduling of jobs to be accomplished within some given metrics or limits. In Cloud computing, finishing of tasks requires various resources which are available to them by filling certain constraints like best performance, minimum execution time, shortest response time, fault-tolerance and quality of expected services. The scheduler should command the jobs in a way where the steadiness between refining the quality of services and at the same time protecting the effectiveness and fairness of the jobs. Thus, in large-scale dispersed systems, the performance assessment of the algorithm is important. Here, our main objective is to learn various job scheduling algorithms.

KEYWORDS: Algorithms, Cloud, Data Centre, IaaS, Job Scheduling, Resource Selection, Virtual Machines**.**

Date of Submission: 12-03-2018 Date of acceptance: 27-03-2018

I. INTRODUCTION

Cloud Computing is a word used to define the use of computing resources (Hardware and Software) that are provided as a service over internet to the users.

In other words - Cloud computing is a platform facilitating universal, useful, on-demand network access to a joint group of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be quickly made accessible and released with minimal handling effort or service provider contact.

Cloud computing is an innovative IT model in which a number of resource allocation is taking place on combined resources from many techniques for parallel computing, distributed computing, and policy virtualization skills. Cloud computing is a type of distributed computing method where a large group of systems is associated in private or public networks providing system for application, data and file storage which can be dynamically available. The resources used for these services can be measured and the users can be charged only for the resources they used. Cloud computing is a setting in which we use the computing resources in distant data Centre's rather than the local computing systems. The cloud environs provide different virtualized platforms that help the user to finish their jobs with tiniest completion time and cost.

Job Scheduling is a decisive factor of the resource management in the cloud environment. Job Scheduling is a procedure of, how jobs should be completed in the environment in terms of actual planning of resources and time. It is mostly accountable for resource sharing at several levels. E.g. A server can be shared amongst many virtual machines, each virtual machine may providing many applications and each application may comprise of many threads. Scheduling method in the cloud can be done in three stages:

a) **Resource determining and filtering:** Datacenter broker know the current position of all the resources that are available in the cloud and also the remaining resources that may be accessible. He frequently collects the status of each resource attached to the cloud.

b) **Resource selection:** Based on data obtained from the resources status about current queued jobs and info on the status of cloud resources, the cloud scheduler makes decisions regarding the establishment or deletion of specific cloud nodes (VMs) in order to best outfit the set of jobs to be completed.

c) Job submission: In this stage, finally the job is submitted to best selected available resource. The goal of this paper is to focus on various task scheduling algorithms. The rest of the paper contains the work on scheduling in cloud environment followed by the conclusion.

II. SCHEDULING

Job scheduling problem in Cloud is defined as, "Jobs and Resources need to be allocated and scheduled in such a way that cloud users can execute their jobs with minimum time, cost and maximize the user satisfaction and throughput of Cloud Resource Provider". In a broader sense, the user is expected to complete the jobs with minimum time and minimum cost.

A. Main Performance Metrics of scheduling Algorithm:

- Execution Cost: It is defined as the total cost of all the resources used at the execution of the job.
- Make span: It is the total length of the Schedule means when all the jobs on schedule get finished.
- Execution Time: It is defined as the time from when a job is submitted to cloud environment till it gets its final execution.
- Job Rejection Ratio: It is defined as the total rejected jobs due to exceeding of execution time than the deadline time or higher cost to the total number of jobs submitted.
- User Satisfaction Level: It is defined as, how far the facilities of the service provider provide satisfaction in terms of resources like storage and computation.

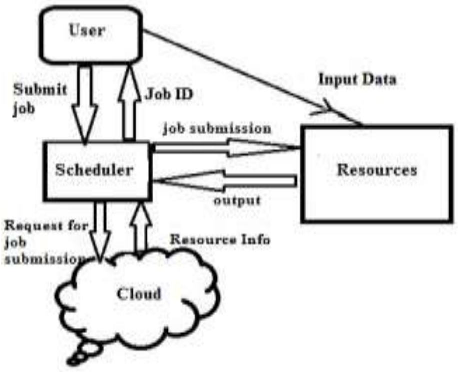

Fig.1 Job scheduling outline

B. Types of scheduling:

(a) Static and Dynamic Scheduling: In Static Scheduling [4], all the information about the status of all available resources in the cloud as well as all the needs of jobs knowing in advance and after that the job is mapped to suitable resource. No job failure and resources are assumed available all the time.

In Dynamic Scheduling [4], the task allocation is done on the go as the application executes, where it is not possible to find the execution time. The jobs are entering dynamically and the scheduler has to heavy efforts in decision making to assign resources. The benefit of the dynamic over the static scheduling is that the system need not have the runtime behavior of the application before it runs.

(b) Centralized, Decentralized and Hierarchical Scheduling: In Centralized Scheduling [4], there is a central scheduler or a bunch of many spread schedulers that have the accountability to make the global scheduling decisions. In this, there is more control on resources: the scheduler continuously monitors all available resources status and therefore it is easier to obtain efficient schedulers. The advantage is ease of implementation, but the disadvantage is lack of scalability, fault-tolerance, and performance.

In Decentralized Scheduling [5], there is no central thing controlling the resources. In this type, the lower schedulers known as local resources machine (LRM) manage and retain the job queue. It is less efficient than centralized scheduling.

In Hierarchical Scheduling [5], the computation is at three levels. The top level is meta-level, the tasks are not planned directly, but reconfiguring the scheduler according to the features of the input jobs. The middle level is called the group level, where the manager in each group cooperates with each other and assigns jobs for the workers in the group. The bottom level is called the within-group level, where the workers in each group perform self-scheduling.

(c) Pre-emptive and non pre-emptive scheduling: In Pre-emptive Scheduling [6], pre-emption is allowed; that is, the current execution of the job can be interrupted and the job is migrated to another resource. In Non pre-emptive Scheduling [7], a job should entirely be accomplished in the resource (the resource cannot be taken away from the task, job or application).

(d) Immediate and Batch Scheduling: In Immediate Scheduling [9], as soon as the job is received, it is scheduled as there will be no waiting for the next time interval. In Batch Scheduling [9], Jobs are first grouped into batches, and then they are assigned to the resources by the scheduler.

C. **Analysis of some algorithms**:

(i) Adaptive resource allocation for preventable jobs in Cloud Systems: Jiayin Li et.al [9] proposed an algorithm based on adaptive resource allocation for the cloud system in which tasks can be preventive. This algorithm adapts the updated status of the actual task executions and adjusts its resource allocation scheme accordingly. In this paper, infrastructure as a service (IaaS) is taken into consideration. The cloud services are provided through the data center in the form of Virtual machines (VM). In Cloud Computing, three different modes of renting the computing resources are available, such as Advance Reservation (AR), Best-effort and Immediate from a cloud provider. If we have problems in resource utilization, AR and best-effort can be combined. In this paper, some of the submitted jobs are in AR mode while others are in Best-effort. Jobs of AR mode have the higher priority than best-effort. So they can prevent the jobs of best-effort. In this paper, two algorithms for the task scheduling: adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) [9] are projected. Once a job is submitted to a scheduler, it will be divided into tasks in the form of Directed Acyclic Graph (DAG). Both ALS and AMMS include a static task scheduling for resource allocation. Then the scheduler will repeatedly re-evaluate the remaining static allocation with a pre-defined frequency, based on the latest information of task execution. To generate the static allocation two greedy algorithms, the Cloud List Scheduling (CLS) and the Cloud Min-Min Scheduling (CMMS) are used. . The final execution time of a task may be different as we expect since there may be chase condition for resources or due to high network traffic. The experimental results show that these algorithms work significantly in extreme resource controversy situation.

(ii) Priority-based consolidation of Parallel Workloads in the Cloud: Xiao cheng Liu et.al [10] proposed an algorithm on improving resource utilization for data centers on which jobs are executed in parallel, particularly when purpose is to make use of the remaining computing capacity of data centers nodes that run parallel processes with low resource utilization which affects the performance of parallel job scheduling and gradually improve it. Basic Algorithms used CMBF (Conservative Migration supported Backfilling) continuously searches for backfilling jobs for each job in the queue when making pre-emption decisions. AMBF (Aggressive Migration Supported Backfilling) tracks for all those backfilling jobs for the job that is at the top of the job queue. It will allow the head-of-queue job to prevent / override the other jobs. The author uses virtualization technologies to divide the computing capacity of each node into two tiers, the foreground virtual machine (VM) tier (with high CPU priority) and the background VM tier (with low CPU priority). By using them, there is an efficient use of two-tier VMs to improve the responsiveness for parallel jobs scheduling algorithm.

(iii) Improved cost-based Algorithm for Task scheduling in Cloud computing: S. Selvarani et.al [11] proposed a method based on the user task schedule, cost of each task differs. Cloud does not schedule the user task as it is done in traditional ways. In this paper, an improved cost-based scheduling algorithm [11] is proposed in which their actual resources are mapping for jobs are done. This algorithm mainly calculates the cost of resources and computation performance and works on gradually improves the computation/communication ratio by grouping the user tasks by taking consideration of processing capability of a resource and sends the grouped jobs to the resource. In this paper, main focus is on devising a scheduling strategy in which the independent jobs are grouped having small processing requirements and according to network conditions, Scheduler schedules them. In traditional ways, we take user tasks as overhead application but problem is that in doing so, there is no relationship between overhead application and the way tasks create the overhead cost of the resources in cloud system. If complexity of tasks is less but size is large, then cost is high else complexity of the task is high and size is short, then cost is low. Activity-based costing measures both the cost of the resources and the computation performance. The cost of every individual resource is different. Tasks are sorted according to their priority, and they are placed in three different lists based on three levels of priority namely high priority, medium priority, and low priority. For computation of tasks, the system can take from high priority list first, then medium and then low. The Improved Activity Based Costing method selects a set of resources to be used for computing. It groups tasks according to the processing capability of resources available. The coarse-grained tasks are processed in the selected resources so that the Computation-Communication ratio is reduced. The experimental results using a simulator show that the time taken to complete tasks after grouping the tasks is very less when compared with time taken to complete the tasks without grouping the tasks.

(iv) A priority based Job Scheduling Algorithm in Cloud Computing: Shamsollah Ghanbari et.al [12] proposed a new job scheduling algorithm in cloud computing by using mathematical statistics. This algorithm made its establishment on the priority property that's why it is known as Priority-Based Algorithm. It is based on multiple criteria decision-making model. In 1980 Thomas Saaty was first developed a model that build pair wise comparison based on multiple criteria and multiple attributes and named it as Analytical Hierarchy Process (AHP)[16][17]. AHP is purely based on steady Comparison Matrix, so by making the use of AHP, comparison matrices are computed according to the attributes and criteria's accessibilities. In this algorithm, each job requests a resource which has a pre-determined precedence. So according to resources accessibilities, comparison matrices of each job are computed. The author also computes the comparison matrix of resources which will help later for jobs selection. Then author computes priority vectors (vector of weights) for each comparison matrix and finally a normal matrix of all jobs is computed named as Δ. Similarly, a normal matrix of all resources is computed and marks this matrix as γ. The next step of the algorithm is to compute Priority Vector of S (PVS), where S is set of jobs. PVS is calculated by multiplying matrix Δ with matrix γ. At the final step, the algorithm chooses the job with highest calculated priority on basis of that appropriate resource is allocated to that job. Now the list of jobs is updated and the scheduling process continues till all the jobs are assigned to a suitable resource. Experimental results indicate that the algorithm has realistic complexity. But there are some issues such as complexity, consistency and end time.

(v) A new class of Priority-based Weighted Fair Scheduling Algorithm: Li Yang et.al [13] proposed a kind of weighted fair scheduling algorithm. It uses the strict rob priority class which uses an absolute priority queue that is based on the class weighted fair scheduling algorithm (CBWFQ). This algorithm removes the drawbacks of the traditional weighted fair scheduling algorithm. In traditional Weighted Fair Scheduling algorithm, based on the weight of each business flow, the services of all active queue is measured and treat accordingly. The job of the classifier is to classify the job when a new job arrives. Then buffer is checked for each group and if the buffer is not overloaded then the job is stored in the buffer otherwise job is dropped. Each job enters a different implicit queue. The four main features of this algorithm are Weight, Dispatch, Discard, and Rob. Introduction of rob rule jointly with dropping rule makes it more efficient. This new algorithm pooled, buffer management and queue scheduling. When the author works on the real-time applications, only then it may cause some delays. It also promises surety of fairness and better utilization of buffers. Main advantages are bandwidth allocation and delay without loss of throughput.

(vi) Balanced Reduce Algorithm (BRA) an efficient data locality driven Task Scheduling Algorithm for Cloud Computing: Jiahui Jin et.al [14] proposed a heuristic task scheduling algorithm known as Balance-Reduce (BR) in which first the initial task allocation is proposed and then the job completion time is gradually reduced by initial task allocation. By collecting the entire global view of all the resources state, the algorithm dynamically allocates the data locality. In this system, first a set of independent tasks is taken on a homogeneous platform with *m* tasks and *n* servers, where each task processes an input block on a server. A job is not completed until all tasks are finished. Cloud computer Clusters workload is mainly designed to have an allocation strategy that minimizes the makespan of tasks. BAR called Balance-Reduce is a data locality driven task scheduling algorithm, having the best-case complexity of $O(\max\{m+n, n \log n\}\cdot m)$. BAR is dividing into two phases, balance and reduce: In Balance phase, a balanced total allocation is produced where all tasks are allocated to their preferred servers uniformly. In Reduce phase, we create a sequence of total allocations and reduce the makespan iteratively. In a poor network environment, BAR tries its best to improve data locality. When the cluster is overloaded BAR decreases the data locality to make tasks commence early. The simulation results show that BAR shows an enhancement and can deal with a large problem case in a minimal time.

(vii) Agent based priority Heuristic for Job Scheduling on Computational Grids: Shah and Syed Nasir Mehmmod et. al. [15] proposed a job scheduling method based on the agent used for effective and efficient execution of user jobs. This algorithm mainly accounts on Quality of Service (QoS) parameters like waiting time, turnaround time, response time, total completion time, etc. based on the classification priority are assigned to each of the job. Agent based Heuristic Scheduling (AHS) uses task agent that makes job allocation more effective to achieve an optimal solution. Firstly, the task agent receive jobs from users and distributes according to different user levels to different prioritized global queues to obtain the optimal job distribution based on user levels, AHS uses agent based job distribution plan at the global level. It also handles job priorities at local levels that provide efficient and effective execution of jobs. For different global queues, for assigning jobs to the global queue, the priorities are proposed as threshold value. If at any instant, jobs have same priorities then execution of job having minimum run time will takes place first, otherwise First Come First Serve (FCFS) algorithm is used. AHS has optimal performance with respect to QoS parameters.

III. CONCLUSION

Scheduling is one of the major issues in the management of job execution in a cloud environment. Based on different parameters like time, cost, and makespan, speed, scalability and resource allocation, analysis and synopsis of various diverse job scheduling algorithms are carried out. Balanced Reduce (BRA) algorithm is a data locality focused task scheduling algorithm, which provides a good response in time O(max{m+n, n log n}•m. Adaptive Resource Allocation for active Jobs in Cloud Systems algorithm adjust the resource allocation adaptively based on the updated status of the actual task accomplishments. It works fine in strong resource controversy situation. The interim results of the Improved cost-based algorithm for task scheduling in Cloud computing shows that the time taken to complete tasks after grouping the tasks is very less as compared with the time taken to complete the tasks without grouping. Priority-based Weighted Fair Scheduling Algorithm considers certain restrictions of jobs to schedule, whereas Priority based Job Scheduling Algorithm studies all the parameters of jobs to perform scheduling. Improving one of the main parameters like makespan, throughput, and uniformity of the existing algorithm is the main development of the forthcoming work.

REFERENCES

[1]. The NIST definition of cloud computing, NIST special publication 800-145.

[2]. Soma Sundaram Thamari Selvi and Kannan Govindarajan “CLOUDRB: A framework For Scheduling and managing High-Performance Computing (HPC) applications in science Cloud.” Future Generation Computer Systems 34(2014): 47-65

[3]. Thomas A. Henzinger, Anmol V.Singh, Vasu Singh. Thomas Wies, “Static Scheduling in Clouds”.

[4]. M.Arora, S.K.Das, R.Biswas “A Decentralized Scheduling and Load Balancing Algorithm For heterogeneous Grid Environments”, “Proc. Of International Conference on Parallel Processing Workshop (ICPPW’02)”,Vancouver, British Columbia Canada, August 2002, pp.400-505

[5]. T.Casavant and J.Kuhl,“A Taxonomy of Scheduling in General Purpose Distributed Computing Systems”,“IEEE Trans. On Software Engineering”, vol.14, no.3, February 1988,pp.141-154.

[6]. Fatos Xhafa, Ajith Abraham, “Computational models and heuristic methods for Grid scheduling problems”, “Future Generation Computer Systems 26”, 2010, pp.608-621.

[7]. Amalarethinam, D. I., and Palaniandy Muthulakshmi. "An Overview of the Scheduling Policies and Algorithms in Grid Computing." International Journal of Research & Reviews in Computer Science 2.2 (2011).

[8]. Yun-Han Lee et al, Improving Job Scheduling Algorithms in a Grid Environment, Future Generation Computer Systems, 27 (2011) 991–998.

[9]. Jiayin Li, Meikang Qiu, Jian-Wei Niu, Yu Chen, Zhong Ming Adaptive Resource Allocation for Pre-emptable Jobs in Cloud Systems‖ 2010 IEEE.

[10]. Liu, Xiaocheng, et al. "Priority-Based Consolidation of Parallel Workloads in the Cloud." (2012): 1-1.

[11]. Mrs S .Selvarani Dr G Sudha Sadhasivm “Improved Cost-Based Algorithm For Task Scheduling In Cloud Computing”, 978-1-4244-5967-4/10/$26.00©2010 IEEE.

[12]. Ghanbari Shamsollah and Mohamed Othman, ”A Priority based Job Scheduling Algorithm in Cloud Computing” Procedia Engineering 50(2012): 778-785.

[13]. Li Yang et al, A new Class of Priority-based Weighted Fair Scheduling Algorithm, Physics Procedia, 33 (2012) 942 – 948.

[14]. Jiahui Jin, Junzhou Luo, Aibo Song, Fang Dong and Runqun Xiong, BAR: An Efficient Data Locality Driven Task Scheduling Algorithm for Cloud Computing‖, 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing.

[15]. Shah, Syed Nasir Mehmmod, et.al “Agent Based Priority Heuristic for Job Scheduling on Computational Grids” Procedia Computer Science 9 (2012):479-488.

[16]. T L Satty, How to Handle Dependence With the Analytic Hierarchy Process, Math Modeling, 9(3-5) (1987) 369-376.

[17]. T.L.Saaty, Decision Making for Leaders; The Analytical Hierarchy Process for Decisions in a Complex World, Pittsburgh: RWS Publications, (1982,2000).

Rajendra T. Kaple " Analysis of Some Job Scheduling Algorithms In Cloud Computing Environment.” International Journal of Computational Engineering Research (IJCER), vol. 08, no. 03, 2018, pp. 17-21.