TASK SCHEDULING IN CLOUD COMPUTING

SONIA SINDHU[1]

Assistant professor, Govt.College for Women

Jind, Haryana,India

Abstract: **Recently, there has been a dramatic increase in the popularity of cloud computing systems that rent computing resources on-demand, bill on a pay-as-you-go basis, and multiplex many users on the same physical infrastructure. It is a virtual pool of resources which are provided to users via Internet. It gives users virtually unlimited pay-per-use computing resources without the burden of managing the underlying infrastructure. One of the goals is to use the resources efficiently and gain maximum profit. Scheduling is a critical problem in Cloud computing, because a cloud provider has to serve many users in Cloud computing system. So scheduling is the major issue in establishing Cloud computing systems. The scheduling algorithms should order the jobs in a way where balance between improving the performance and quality of service and at the same time maintaining the efficiency and fairness among the jobs. This paper aims at studying various scheduling methods. A good scheduling technique also helps in proper and efficient utilization of the resources. Many scheduling techniques have been developed by the researchers like GA (Genetic Algorithm), PSO (Particle Swarm Optimization), Min-Min, Max-Min, Priority based Job Scheduling Algorithm . This paper reviews certain papers on resource management and job scheduling in cloud computing.**

I. INTRODUCTION

"Cloud computing is a model enabling ubiquitous, ,convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." In Cloud Computing the term Cloud is used for the service provider, which holds all types of resources for storage, computing etc. Mainly three types of services are provided by the cloud. First is Infrastructure as a Service (IaaS), which provides cloud users the infrastructure for various purposes like the storage system and computation resources. Second is Platform as a Service (PaaS), which provides the platform to the clients so that they can make their applications on this platform. Third is Software as a Service (SaaS), which provides the software to the users; so users don't need to install the software on their own machines and they can use the software directly from the cloud. Due to the wide range of facilities provided by the cloud computing, the Cloud Computing is becoming the need of the IT industries. The services of the Cloud are provided through the Internet. The devices that want to access the services of the Cloud should have the Internet accessing capability. Devices need to have very less memory, a very light operating system and browser. Cloud Computing provides many benefits: it results in cost savings because there is no need of initial installation of much resource; it provides scalability and flexibility, the users can increase or decrease the number of services as per requirement; maintenance cost is very less because all the resources are managed by the Cloud providers. Cloud computing is a product from mixing traditional computer techniques and network technologies, such as grid computing, distributed computing, parallel computing, utility computing, network storage, virtualization, load balancing, etc[1].

Fig.1 Evolution of Calculation Mode

The main Purpose is to schedule tasks to the Virtual Machines (VMs) in accordance with adaptable time, which involves finding out a proper sequence in which tasks can be executed under transaction logic constraints [2]. The job scheduling of cloud computing is a challenge. To take up this challenge we review the number of efficiently job scheduling algorithms. It aims at an optimal job scheduling by assigning end user task. The rest of the paper is organized as follows. In next section Literature Survey about different scheduling algorithms of Virtual machine in cloud are discussed. Section 3 describes Existing Scheduling Algorithm in Cloud Computing. Section 4 discusses The Proposed Scheduling Algorithm. Section 5 discusses Experimental Setup and Results are analyzed. Conclusions are discussed in section 6.

ISSN: 2278 – 1323

II. Task Scheduling in Cloud Computing

Job Scheduling of cloud computing refers to dispatch the computing tasks to resource pooling between different resource users according to certain rules of resource use under a given cloud circumstances. At present there is not a uniform standard for job scheduling in cloud computing. Resource management and job scheduling are the key technologies of cloud computing that plays a vital role in an efficient cloud resource management

III. Existing Scheduling Policy

The following task scheduling algorithms are presently established in the cloud environments

A. Ant Colony Optimization (ACO)-inspired:

A new Cloud scheduler based on Ant Colony Optimization is the one presented by Cristian Mateos and et.al [3]. The goal of our scheduler is to minimize the weighted flowtime of a set of PSE jobs, while also minimizing Makespan when using a Cloud. In the ACO algorithm, the load is calculated on each host taking into account the CPU utilization made by all the VMs that are executing on each host. This metric is useful for an ant to choose the least loaded host to allocate its VM.Parameter Sweep Experiments (PSE) is a type of numerical simulation that involves running a large number of independent jobs and typically requires a lot of computing power. These jobs must be efficiently processed in the different computing resources of a distributed environment such as the ones provided by Cloud. Consequently, job scheduling in this context indeed plays a fundamental role. In this algorithm, Makespan and flowtime are evaluated. Evaluation results of this metrics show that ACO performance better than two other (Random and Best effort) algorithms.

B. Min-Min Algorithm

Min-Min begins with a set of tasks which are all unassigned. First, it computes minimum completion time for all tasks on all resources. Then among these minimum times the minimum value is selected which is the minimum time among all the tasks on any resources. Then that task is scheduled on the resource on which it takes the minimum time and the available time of that resource is updated for all the other tasks. It is updated in this manner; suppose a task is assigned to a machine and it takes 20 seconds on the assigned machine, then the execution times of all the other tasks on this assigned machine will be increased by 20 seconds. After this the assigned task is not considered and the same process is repeated until all the tasks are assigned resources. We have implemented the logic of Min-Min and Max-Min algorithms on the execution time values as given in the TABLE I below. [4]

C. Max-Min Algorithm

Max-Min is almost same as the min-min algorithm except the following: in this after finding out the completion time, the minimum execution times are found out for each and every task. Then among these minimum times the maximum value is selected which is the maximum time among all the tasks on any resources. Then that task is scheduled on the resource on which it takes the minimum time and the available time of that resource is updated for all the other tasks. The updating is done in the same manner as for the Min-Min. All the tasks are assigned resources by this procedure.

TABLE I
EXECUTION TIMES

	M0	M1	M2	M3
T0	200	250	220	300
T1	150	170	190	160
T2	300	320	180	360
T3	400	380	350	310
T4	100	120	140	160
T5	220	250	280	200

We have assumed four machines and six tasks. By using different scheduling techniques, the tasks are assigned in a different sequence to different machines for execution. When we apply the different scheduling techniques Min-Min and Max-Min, then the tasks will be assigned to the machines as given in the following figures:

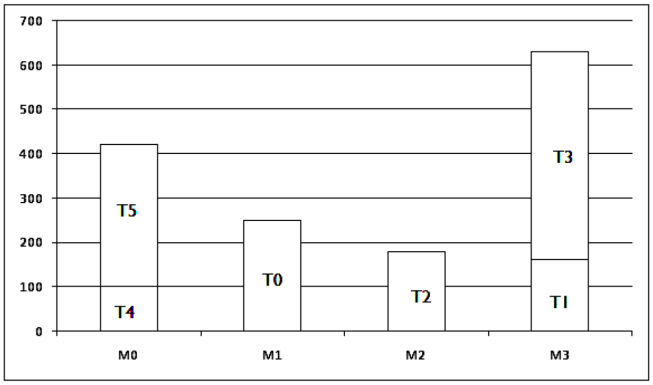

Fig. 2 Task assignment by Min-Min algorithm

There is a term "makespan" in Min-Min and Max-Min scheduling techniques, which is the maximum execution time on any machine among the machines on which the tasks are scheduled. For example, in Fig. 2, "630" is the makespan because it is the maximum execution time among the four machines. In Fig. 2 and Fig. 3, the x-axis represents the different machines and y-axis represents the execution times. We have got the following different values of makespans by the two techniques: Method used Makespan :

Min-Min 630
Max-Min 590

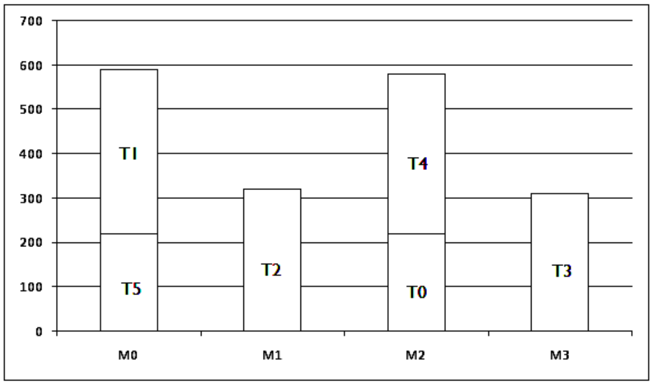

Fig. 3 Task assignment by Max-Min algorithm

Based on the different execution times of tasks on resources, one technique can outperforms the other and the assignment of resources to the tasks can change i.e. if any task is assigned to a machine if we use one technique; the same task can be assigned to another machine if we use other technique.

D. Particle Swarm Optimization (PSO) Algorithm:

Particle Swarm Optimization (PSO) as a meta-heuristics method is a self-adaptive global search based optimization technique introduced by Kennedy and Eberhart [9]. The PSO algorithm is alike to other population-based algorithms like Genetic algorithms (GA) but, there is no direct recombination of individuals of the population . The PSO algorithm focuses on minimizing the total cost of computation of an application workflow. As a measure of performance, Authors used cost for complete execution of application as a metric. The objective is to minimize the total cost of execution of application workflows on Cloud computing environments. Results show that PSO based task-resource mapping can achieve at least three times cost savings as compared to Best Resource Selection (BRS) based mapping for our application workflow. In addition, PSO balances the load on compute resources by distributing tasks to available resources.[5],[10]

E. Round Robin Algorithm:

The Round Robin algorithm mainly focuses on distributing the load equally to all the resources [11]. Using this algorithm, the broker allocates one VM to a node in a cyclic manner. The round robin scheduling in the cloud computing is very similar to the round robin scheduling used in the process scheduling. The scheduler starts with a node and moves on to the next node, after a VM is assigned to that node. This is repeated until all the nodes have been allocated at least one VM and then the scheduler returns to the first node again. Hence, in this case, the scheduler does not wait for the exhaustion of the resources of a node before moving on to the next. Although round robin algorithms are based on simple rule, more load is conceived on servers and thus unbalancing the traffic. Result of Round Robin algorithm shows better response time and load balancing as compared to the other algorithm.[7]

F. GENETIC ALGORITHM

Genetic algorithm is a method of scheduling in which the tasks are assigned resources according to individual solutions (which are called schedules in context of scheduling), which tells about which resource is to be assigned to which task. Genetic Algorithm is based on the biological concept of population generation. The main terms used in genetic algorithm are

a. Initial Population

Initial population is the set of all the individuals that are used in the genetic algorithm to find out the optimal solution. Every solution in the population is called as an individual. And every individual is represented as a chromosome for making it suitable for the genetic operations. From the initial population the individuals are selected and some operations are applied on those to form the next generation. The mating chromosomes are selected based on some specific criteria.[6]

b. Fitness Function

A fitness function is used to measure the quality of the individuals in the population according to the given optimization objective. The fitness function can be different for different cases. In some cases the fitness function can be based on deadline, while in cases it can be based on budget constraints.

c. Selection

We use the proportion selection operator to determine the probability of various individuals genetic 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.

d. Crossover

We use single-point crossover operator. Single-point crossover means only one intersection was set up in the individual code, at that point part of the pair of individual chromosomes is exchanged.

e. Mutation

Mutation means that the values of some gene locus in the chromosome coding series were replaced by the other gene values in order to generate a new individual. Mutation is that negates the value at the mutate points with regard to binary coded individuals. Genetic Algorithm works in the following manner:

```
1. Begin
2. Initialize population with random candidate solutions
3. Evaluate each candidate
4. Repeat Until (termination condition is satisfied)
       a. Select parents
       b. Recombine pairs of parents
       c. Mutate the resulting offsprings
       d. Evaluate new candidate
       e. Select individuals for the next generation;
5.  End.
```

G. *Priority based Job Scheduling Algorithm*

Shamsollah Ghanbari and Mohamed Othman [6] proposed a priority based job scheduling algorithm named it "PJSC" which can be applied in cloud environments. This Scheduling algorithm is consisted of three levels of priorities including: scheduling level (objective level), resources level (attribute level) and job level (alternative level), as shown in Fig 5.

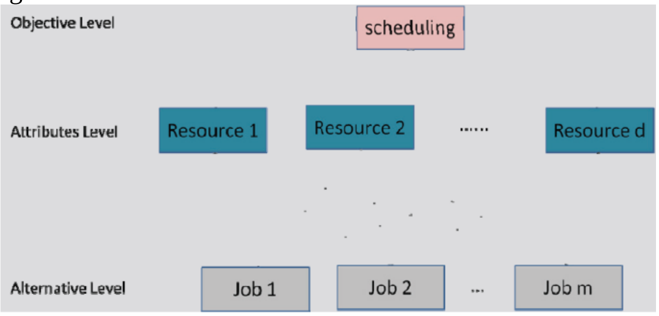

Fig. 4 Priority level of jobs and services in cloud computing

The detailed steps of Priority job scheduling are illustrated in Fig. 4. The algorithm take a set of jobs J= {j1, j2,.., jm} that request resources in a cloud environment and a set of resources C={c1,c2,…,cd} available in cloud environment as input where (d<<m). Each job requests a resource with a determined priority. The priority of each job is compared with other jobs separately. For example, suppose the ratio of priority of Ji to Jj for gaining a particular resource such as Cg is 7,

in this case

Pgij=7 and Pgji=1/7;

In above Equation Pg denotes a matrix with m rows and m columns known as comparison matrix. And for each resource such comparison matrixes of jobs are created according to priority of resource accessibilities say Q1, Q2 ... Qd. Now for each of comparison matrixes a priority vector (vector of weights) Aw is computed by following equation:

$$Aw= \lambda max * w$$

where is denoted the principal eigen value of Matrix A and is denoted the corresponding eigenvector. Suppose W1, W2 ... Wd are corresponding priority vectors of Q1, Q2 ... Qd respectively. And using these priority vectors normal matrix of jobs level are defined as Δ= [W1, W2 ... Wd] The next step of the algorithm is to make a d×d comparison matrix for resources based on their priorities. This matrix helps to determine which resource has higher priority than others based on decision maker(s). Say M is comparison matrix for resources level and β will be defined as priority vector of M. Then PVS is calculated which is denoted as priority vector of scheduling jobs. PVS can be calculated by following equation: PVS= Δ .β Finally, the maximum element of PVS is selected and then select corresponding job of J in order to allocate a suitable resource and after that list of jobs is updated.[9]

H. *Dynamic Scheduling Algorithm Based on Threshold*

To get the real-time feedback of the state of virtual machine, there are two ways. One of them is to construct a set of feedback mechanism between dispatcher and virtual machines to get the real-time feedback of the tasks load on virtual machine, and then make a real-time adjustment on job allocation upon the fact of virtual machines. The other one is to use the dynamic scheduling among virtual machines themselves to get the real-time state of the load of virtual machines. If overload or idleness occurred, tasks could be readjusted and redistributed among virtual machines. By dynamic dispatch in virtual machines, dynamic scheduling algorithm based on threshold can allocate jobs and resources flexibly and reduce the efficiency impact caused by the synchronization among virtual machines.There are some virtual machines overload and some idle at a certain time, dynamic job adjustment is conducted to shorten the total cost time, thereby enhancing efficiency. However, task allocation between virtual machines refers to synchronization problems, which is also the biggest problem of the dynamic scheduling algorithm based on threshold. Since each virtual machine is independent to each other, in the other word they are non-interfering. They can perform tasks in parallel. If virtual machines are synchronized, they inevitably bring effects to their performance. Therefore, the synchronization operation should be kept to minimum range. In order to reduce the impact of synchronization, two measurements are taken. First, set the threshold. Synchronization is executed only when the virtual machine reaches a threshold. The larger the threshold is, the smaller the impact of synchronization will be. Second, limit the synchronization down to two virtual machines. The smaller the number of virtual machines for synchronization is, the weaker the impact it brings Task assignment involves in setting task classification according to PRI. A task that has a higher execution priority has higher PRI. Reaching the uptime or task threshold means that the time threshold of task running or the number of tasks that are waiting in the line is reached. It includes two conditions. There are two types of task threshold. One is the number of tasks waiting to be done in the queue on one virtual machine. The other one is the number of tasks that have been finished on another virtual machine. If both numbers were larger than the threshold value at the same time, these two virtual machines would be synchronized. And their tasks will be balanced and will continue working.Task equilibrium means that if there is at least one idle virtual machine and at least one overload virtual machine, other virtual machines will execute tasks independently. Xian’s paper [23] compares between dynamic scheduling algorithm based on threshold and virtual machines with the static independent job scheduling algorithm on CloudSim platform. The result suggests that when there are a fairly large number of tasks, the former can complete task allocation efficiently and reduce the running time greatly. It shows an obvious advantage over the latter.[8]

IV. Our Contributions

A scheduling framework can be implemented by using different parameters. Good scheduling framework should include the following specifications. It must focus on:

- Load balancing and Energy efficiency of the data centers and virtual machines (VMs)
- Quality of Service parameters calculated by the user which contain execution time, cost and so on
- It should satisfy the security features.
- Fairness resource allocation places a vital role in scheduling

Considering all the metrics in a single scheduling framework is not a possible solution, hence it increases the complexity of the design; the scheduling framework can be implemented.

V. Conclusion

Cloud computing is one of the user oriented technology in which user faces a pool of virtualized computer resources. In this paper we survey various existing scheduling algorithms in cloud computing. Since cloud computing is in infancy state, a scheduling framework should be implemented to improve the user acquiescence along with the service providers. The scheduling metrics can be coupled to prepare a framework for recourse allocation and scheduling in cloud computing. The scheduling framework should consider the user input limitations (deadlines, performance issues, execution cost, transmission cost, energy efficiency, Load Balancing, and Makespan) and so on.

REFERENCES

[1] Mohammad Hamdaqa and Ladan Tahvildari , “Cloud Computing Uncovered: A Research Landscape”. Elsevier Press. pp. 41–85. ISBN 0-12-396535-7.

[2] Huang Q.Y., Huang T.L.,“An Optimistic Job Scheduling Strat egy based on QoS for Cloud Comput ing”, IEEE Int ernat ional Conference on Intelligent Comput ing and Integrated Systems (ICISS), 2010, Guilin, pp. 673-675, 2010

[3] Cristian Mateos, Elina Pacini & Carlos Garc Garino, (2013), An ACO-inspired algorithm for minimizing weighted flowtime in cloud-based parameter sweep experiments.

[4] K. Etminani, and M. Naghibzadeh, "A Min-min Max-min Selective Algorithm for Grid Task Scheduling,"The Third IEEE/IFIP International Conference on Internet, Uzbekistan, 2007.

[5] Rajkumar Buyya, A Particle Swarm Optimization-basedHeuristic for Scheduling Workflow Applications in Cloud Computing Environments, Cloud Computing and Distributed Systems Laboratory, Department of Computer

[6] Yin H., Wu H., Zhou J., “An Improved Genet ic Algorithm with Limited It erat ion for Grid Scheduling”, IEEE Sixt h Internat ional Conference on Grid and Cooperat ive Comput ing, 2007. GCC 2007, Los Alamitos, CA, pp. 221-227, 2007

[7] Pooja Samal and Pranati Mishra, (2013), "Analysis ofvariants in Round Robin Algorithms for load balancing inCloud Computing", International Journal of Computer Science and Information Technologies, pp. 416-419, Vol. 4(3)

[8] Wang Yonggui, Han Ruilian. Study on cloud computing task schedule strategy based on MACO algorithm[J]. Computer Measurement & Control, 2011, 19 (5): 1203~1204, 1211

[9] Shamsollah Ghanbaria & Mohamed Othmana, (2012), A Priority based Job Scheduling Algorithm in Cloud Computing, Procedia Engineering 50, and PP. 778 – 785.).

[10] J. Kennedy and R. Eberhart, (1995), Particle swarms optimization In IEEE International Conference on Neural Networks, volume 4, pages 1942–1948.