A Review on CPU Scheduling Algorithms in Cloud Environment

Sumit Kumar Rohil¹, Jai Bhagwan²

¹, ²Department of Computer Science Engineering, Guru Jambheshwar University of Science & Technology, Hisar

Abstract: Cloud computing is developing area of computing which provides Internet-based services like shared resources, information, software packages and other resources as per client requirements at a specific time. Cloud operations are performed by three types of services: Software as a Service, Platform as a Service and Infrastructure as a Service. Clouds are deployed based on the four types: public, private, hybrid and community. Cloud Computing has two main fields Load Balancing and Scheduling in which there is a vast scope of research. As users are increasing day by day, there is a need to schedule the tasks efficiently as per user’s requirement. In this paper, our focus is on task scheduling. Some of the basic algorithms are FCFS, RR, Priority-based, Min-min, Max-min for scheduling of tasks.

Keywords: Cloud Computing, Scheduling, Round Robin, SRTF, Virtualization

I. INTRODUCTION

Cloud computing is a most recent innovation technique in IT industries. Cloud computing is extraordinary model of advancement in recent time for IT enterprises, research and development. It comes after the presentation of distributed computing and if one match the cloud computing and distributed computing, a multileveled virtualization is occurred. The entire work is in virtual domain which is identified as a cloud computing concept [1][3].

Cloud computing is a technique in which computing is delivered as a service rather than a product, whereby shared resources and software, and information is provided to consumers as a utility over networks. One of the main advantages and motivations behind Cloud Computing is reducing the CAPEX (capital expenditures) of systems from the perspective of cloud users and providers. To get the favour of cloud client needs to just interface with the web and after that client can perform operations without much using the intense computing and storage. Cloud computing administration is given by CSP (cloud service provider) according to client necessities. Keeping in mind the end goal to satisfy the request of diverse clients, they give distinctive nature of services. In order to conclude the term cloud, it is an executable environment having a dynamic behaviour of resources as well as clients giving various services [2][3][4].

II. TYPES OF CLOUD

On the basis of deployment, cloud can be divided into four categories: Private, Public, Hybrid and Community as shown in figure 1 [2][5] and described following:

1) Private Cloud: It is a platform having users own hardware and software on the cloud that means only owner can use it and other persons don’t have access on that cloud. It is also called internal cloud or corporate cloud. It offers hosted services to a partial number of people. When need of higher security to run applications, this cloud is preferred.

2) Public Cloud: It is a cloud that is given to the clients as pay per use. Public Cloud is used by number of clients when required and clients have some level of control over computing infrastructure which is facilitated.

3) Hybrid Cloud: A hybrid cloud is made up with public as well as private cloud. Each node in a hybrid cloud could be separately managed but applications and data would be allowed to move diagonally. Example of hybrid cloud is Amazon web Service.

4) Community Cloud: A cloud that is authorised and accessed by some organisations and used for their requirement of businesses, security and other common use.

Figure 1: Types of Clouds
Cloud service models describe how services are made available to users. Classic Cloud service models can be ordered into three sorts: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [2][4][6] which are described below:

5) **Infrastructure as a Service (IaaS):** Infrastructure is provided by the cloud service provider in the form of virtualization of computing resources. It refers to the provisioning and the delivery of basic resources such as virtual machines, physical servers, network and storage. Rather than putting resources into their own foundations, organizations can lease resources and utilize them on request as opposed to having their resources locally.

6) **Platform as a Service (PaaS):** Development platforms are provided by cloud organization. It also provides the application-hosting facility. PaaS is on a more advanced and more elevated amount of benefits contrast with IaaS. It gives application advancement situations and programming stages to create, convey, and oversee Cloud applications while not giving stress over the innovation and concealing the low-level subtle elements from the client.

7) **Software as a Service (SaaS):** SaaS is the highest level of the Cloud service model. In this scenario, complete applications are provided to users through the internet. Using this platform, finished touch applications are given to clients through the web. SaaS suppliers oversee framework and have finish control of the application programming projects. Clients simply get to their applications as though they were facilitated locally and don't know anything about the Cloud or even know the innovations points of interest.

![Figure 2: Service Models of Cloud](image)

### III. FEATURES OF CLOUD COMPUTING

Cloud computing provides several features that make it attractive to IT industry, such as [1][2]:

1) **No Up-Front Investment:** The pricing model in cloud computing is based on a pay-per-use principle. Cloud pricing models rent the services and resources which user needs.

2) **Lowering Operating Cost:** Whenever user required or not required the resources, allocation and de-allocation of resources in Cloud done on demand of user. This would help the user to save the cost for resources which are released when their demand are less.

3) **Scalability and Elasticity:** Here, the cloud infrastructure has a large pool of resources and infrastructure. So it can be undoubtedly extend its service to deal with the growing service as indicated by customer request. Rather, Elasticity is the capacity to scale up and down assets or resources when requirement is taken place. Permitting the extraction and dynamic integration of physical resources to the infrastructure take place. That’s mean elasticity enables scalability is present.

4) **Easy access:** The cloud services are provided to users as web-based services. So, they can access the services through any devices supported with Internet connections.

5) **Reducing Business Risks and Maintenance Expenses:** It shifts the business risks such as hardware failures to infrastructure providers because providers have better expertise and resources to manage these risks.

6) **Virtualization:** Virtualization hides a computing platform’s physical characteristics from users. It allows abstraction and isolation of lower level functionalities and underlying hardware.
7) Mobility: Cloud Computing implies mobility since clients can get applications through web effectively anytime.

IV. TASK SCHEDULING

Assigning certain tasks to a particular resource in a particular time is known as task scheduling. Task scheduling is a biggest, most challenging and testing issue in the cloud computing technology. As the industry moved on the cloud day by day the scheduling is needed more appropriate for resource allocation and de-allocation to the user [19]. Subsequently, scheduler ought to be possibly dynamic in nature. Task scheduling for cloud computing is basically focusing to improve the profitable utilization of assets or resources such as bandwidth, memory and reduction in completion time [25]. A successful job scheduling scheme must intend to give output in less response time so the execution of submitted jobs occurs in a possible least time and there will be an occasion of in-time where assets are reallocated. Because of this, extra number of jobs can be submitted to the cloud by clients and less dismissal of jobs happen which finally demonstrate extending realizes quickening the business execution of the cloud [7][20][8]. The facilities, application and resources are generally available dynamically and these dynamic resources must be scheduled properly to achieve maximum benefits. In order to obtain high cloud computing performance the cloud resources must be scheduled optimally otherwise poor results are obtained. Scheduling algorithms are used for dispatching user tasks or jobs to a particular resource or data. Scheduling the resources is not the trivial task due to dynamic nature of resource allocation and de-allocation [9][21]. The scheduler on cloud computing efficiently dispatches the multiple requests from clients such that response time is quick and system performance improves. If resources are scheduled optimally then more and more clients requests can be solved which results in performance and profit for cloud service providers.

Job scheduling for cloud computing has pulled in awesome consideration. Most researches in job scheduling receive a worldview in which a jobs in cloud computing system is characterized by its workload, deadline and the relating utility got by its finishing before the time limit, which are factors considered in contriving an effective scheduling algorithm.

V. BASIC ALGORITHM OF SCHEDULING

There are description of some basic or fundamental algorithm of task scheduling [10][11][12][13][26]:

1) First Come First Serve Scheduling: As the name suggests the tasks come in the queue and these are allocated to VMs as first come first serve basis without considering their execution time. FCFS algorithm or approach is used for parallel processing and waiting queue in this algorithm is smallest as job comes in the queue and it is sent to the server if server is idle. The main disadvantage of it is that it is non-pre-emptive algorithm because the waiting time for the small task is increased as the server is run on long task or waiting queue length is long when long task is at the font of queue. Its turnaround and response time are quite low.

2) Round Robin Scheduling: As the name suggests Round Robin schedules the tasks or processes having same waiting time in queue, RR server is distributed between the tasks in a fair manner. RR algorithm uses queue in a ring fashion for storage of incoming job. Each job is executed with same time and in a turn process in the queue. Whether, a job isn’t being completed during a time slot of its turn, it will still wait for the time slot in next turn. The advantage of RR scheduling is each job is executed in the queue and none has to wait for another job completion. If many jobs are in the queue then here, completion time of all jobs is more.

3) Priority scheduling algorithm: The basic idea is straightforward as each process is assigned a priority for execution. In this case, Equal-Priority processes are scheduled in FCFS order.

4) Min–Min algorithm: Here at first, finding the minimum expected time of all tasks in meta-task happens. The task having the minimum expected completion time is selected and assigned to the corresponding resource. This step is iterated until meta-task is not empty. Here, a big task has to wait for the completion of smaller ones.

5) Max–Min algorithm: The Max–Min algorithm completion time of each task as per the available resource is calculated. A task which has overall maximum completion time is scheduled over a resource according to overall minimum execution time. This step is repeated until meta-task is not empty. Here, the waiting time of a larger task is reduced.

VI. LITERATURE REVIEW

Numerous scientists have proposed distinctive scheduling algorithms that keep running under cloud computing circumstances. A Maximum number of scheduling algorithms have been proposed to accomplish two principal goals to be specific, to run the client assignment within time limit and to manage capability (load balancing) and reasonableness for all tasks. Here, we have inspected the most applicable research works done in the literature for job scheduling in cloud computing.
A. Parsa Saeed et al. (2009) proposed task scheduling algorithm named Resource-Aware-Scheduling Algorithm (RASA) based on composed of two algorithms Max-min and Min-min. RASA cover disadvantages and use the advantages of both algorithms. Experimental results show that RASA is much better than both algorithms in large scale scheduling [14].

B. Gharooni-farda Golnar et al. (2010) present a novel chaos-genetic based algorithm that uses chaotic sequences rather than random processes in genetic algorithms. Evaluation of an algorithm was done by considering it into both balanced and unbalanced workflow structures. The chaos-genetic algorithm gave better results when compared to the Traditional Genetic Algorithm in both cases balanced and unbalanced [15].

C. Pandey Suraj et al. (2010) propose a particle swarm optimization (PSO) based heuristic algorithm for cloud resources to schedule application that minimizes the cost of application workflow in cloud computing. The PSO algorithm performed well in term of at least three times cost saving of application workflow when compared with Best Resource Selection (BRS) algorithm. PSO balanced the load on resources by dividing tasks to available resources [16].

D. Paul and Sanyal (2011) described the issue how to utilize cloud computing asset capably and increment benefits with the job scheduling system. Thus, they designed a credit based scheduling algorithm to figure the entire group of tasks in task queue and find the minimum completion time of all tasks. The proposed scheduling method considered the scheduling issue as a task issue in science where cost of a task was appointed to a resource given by the cost matrix. Anyway, the algorithm did not recognize the handling time of a job, but rather distinct issues were viewed as the probability of an asset to be free not long after executing a task with the goal that it will be accessible for next waiting job [17].

E. Choudhary Monika et al. (2012) presented a scheduling algorithm define crucial challenges in cloud computing. Incoming tasks were categorized on the basis of the task requirement like minimum execution time or minimum cost and prioritized. Resources were selected by using a greedy approach on the basis of task constraints. Results validated the correctness of the framework and shown a significant enhancement over sequential scheduling [18].

F. Vignesh V, Sendhil Kumar (2013) proposed an approach for task oriented resource scheduling for cloud. The accessible resource and client preference were scheduled for the job process by resource allocation, and the resources that required for job process were given to the client by the job rank. This paper built the analysis of resource scheduling algorithms. When compared with the previously designed approaches (RR, Pre-emptive priority) to Shortest Remaining Time First on the basis of time parameter the results shown that SRTF has the least time parameters in all situation and is the proficient algorithm for resource scheduling [19].

G. Karthick AV. et al. (2014) presented multi-queue scheduling algorithm to reduce estimated cost of the reservation and on-demand plans using the global scheduler. This proposed methodology was a concept of a grouping the burst time and solve the starvation problem that occurred at the scheduling time in traditional approach i.e. FCFS, SJF, Combinational Backfill and Improved Backfill. The Multi-Queue Scheduling gave more importance to select job dynamically with respect to achieving the optimum cloud scheduling problem and using free space in an economic way [20].

H. LIU Chun-Yan et al. (2014) focused on a task scheduling algorithm based on Genetic-Ant Colony optimization aiming to solve slow convergence problem that occurred in Ant Colony Optimization due to lack of initial pheromone. GA-ACO used global search ability of Genetic Algorithm (GA) for a better solution, and then; converted it to the initial incense of Ant Colony Optimization (ACO) and get optimal scheduling result through the positive feedback of ACO. The result shown that GA and ACO are beneficial to use cloud computing for solving of scheduling problem and improved the searching efficiency of algorithms [21].

I. Cai Zhiming et al. (2014) proposed Demand Driven Scheduling Model and defined a method to complete time of tasks. An improved algorithm using Demand Driven chromosome was defined to handle multi-objective scheduling. The algorithm results shown high accuracy of prediction when compared with Markov model and IGA. Authors got better performance.
results for task scheduling, smaller makespan and lower deviation of the load. Change in the weight of fitness function adjusted the priority objective [22].

J. Singh Shekhar et al. (2014) focused on a genetic algorithm based scheduling approach which was presented using initial population that was defined by Enhanced Max-Min. The proposed method gave outcomes in a better way in term of makespan. When compared proposed algorithm with Enhanced Max-Min, GA with LCFP and IGA, results were got better in terms of makespan in scheduling [23].

K. Hong Liang et al. (2015) proposed a heuristic approach based on an improved Ant Colony Algorithm (ACA) to remove virtual machine problem (VMP) defined as the Genetic Ant Colony Algorithm-Virtual Machine Problem (GACA-VMP). The goal of this technique was to efficiently obtain servers that minimized total resources simultaneously. The performance evaluation of this technique compared with ACA and a first-fit decreasing algorithm publicized that the proposed algorithm was competitive with a large number of virtual machines (VMs) [24].

L. S.Sujan, R.Kanniga Devi (2015) proposed an approach named as dynamic scheduling scheme for cloud computing. It was considered the resource provisioning as the primary problem was need to be addressed and the scheduling was considered as the unique circumstance. Here, it was utilized a dynamic scheduling scheme considering the makespan as simulation parameter. When compared with the min-min and round robin the proposed approach gave better results. By considering and assessing all of the given parameters and approaches, a makespan based dynamic scheduling approach was introduced for cloud computing [25].

M. Sushil Kumar Saroj, Aravendra Kumar Sharma (2016) proposed a novel scheduling algorithm using variable time quantum based on mean difference of burst time. They discussed in their paper that CPU scheduling plays an important role in utilization of assets or resources and incremented the system execution by exchanging the CPU among different processes. Despite that, it additionally presented few issues, for example, starvation, waiting time, turnaround time and its practical implementation. Previously scheduling approaches determined to solve these issues but most of them solved one issue or two issues. For removing these problems, the authors proposed an algorithm that utilizes the both average and variable time quantum. A few processes were presented along with average time quantum and rest of along with variable time quantum. The proposed algorithm not just reduced the minimum average waiting time and turnaround time but additionally attempted to keep the starvation issue [26].

N. Lin Ruonan et al. (2016) proposed a Pre-allocation Ant Colony Optimization (PACO) for PTS implementation. A pre-allocation strategy for task scheduling was performed in cloud computing and defined new algorithm PACO based on ACO. Algorithm run on the simulation in Cloudsim tool and performance matrices presented that the algorithm in this paper gave better results than a conventional min-min algorithm and genetic algorithm in terms of completion time of jobs [27].

O. Indukuri et al. (2016) described that cloud computing was utility based environment as pay per use model achieved by Parallel, Distributed and Cluster computing accessed through the Internet. A key advantage of cloud computing is on-demand self-service, scalability, and elasticity. In on-demand self-service, the cloud user can request, deploy their own software, customize and pay for their own services. Scalability is achieved through virtualization. Being elastic in nature, cloud service gives the infinite computing resources (CPU, memory and storage). In cloud environment, to achieve the quality of service many scheduling algorithms are available, but the scalability of task execution increases, scheduling becomes more complex. So there is a need for better scheduling. This paper dealt with the survey of dynamic scheduling, different classification and scheduling algorithms currently used in cloud providers [28].

P. Shridhar Domanal et al. (2016) introduced a novel hybrid Bio-inspired algorithm for job scheduling and asset or resource manage, by reason of it assumed a critical part in the cloud computing. Many scheduling approaches, for example, RR, FCFS, ACO and so forth have been generally utilized as a part of several cloud computing systems. In the proposed approach, distribution of the jobs to the VM in a proficient way utilizing Modified Particle Swarm Optimization (MPSO) approach and afterward allocation/management of assets or resources, as requested by the jobs, was taken care of by proposed HYBRID
Bio-inspired calculation (Modified PSO + Modified CSO). Experimental results came to show that the proposed HYBRID approach was better than benchmark algorithms in terms of proficient usage of the cloud assets, enhancing reliability and reducing average response time [29].

The Table 1 shows the comparison of various algorithms in tabular form.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Simulation Parameters</th>
<th>Findings</th>
<th>Tools Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min-Min [13]</td>
<td>Execution Time</td>
<td>Better Than RR, FCFS</td>
<td>Workflowsim</td>
</tr>
<tr>
<td>Resource-Aware Scheduling Algorithm (RASA) [14]</td>
<td>Minimum completion time, Execution Time</td>
<td>Max-min, min-min</td>
<td>Cloudsim, workflowsim</td>
</tr>
<tr>
<td>Particle Swarm Optimization based Heuristic (PSO) [16]</td>
<td>Makespan, waiting time, minimize cost</td>
<td>Best Resource Selection (BRS)</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Credit Based Scheduling Algorithm [17]</td>
<td>Minimum completion time</td>
<td>Better than FCFS</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Shortest Remaining Time First [19]</td>
<td>Makespan, waiting time</td>
<td>Perform better than RR, pre-emptive priority</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Multi-Queue Scheduling Algorithm [20]</td>
<td>Processing time, Burst time</td>
<td>FCFS, SJF, Combinational Backfill and Improved Backfill</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Genetic-Ant Colony algorithm (GA-ACO) [21]</td>
<td>Execution time</td>
<td>GA, ACO</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Modified Genetic Algorithm (MGA) [23]</td>
<td>Makespan</td>
<td>Improved Max Min, GA with LCFP and IGA</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Dynamic Scheduling Scheme [25]</td>
<td>Makespan</td>
<td>Perform better than RR, min-min</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Novel CPU Scheduling with Variable Time Quantum [26]</td>
<td>waiting time and turnaround time</td>
<td>Better than FCFS</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Pre-allocation Ant Colony Optimization (PACO) [27]</td>
<td>Completion time</td>
<td>Min-min and genetic algorithm</td>
<td>Cloudsim</td>
</tr>
<tr>
<td>Novel Hybrid Bio-Inspired Algorithm [29]</td>
<td>average response time, efficient utilization of the cloud resources</td>
<td>Better than RR, FCFS, ACO</td>
<td>Cloudsim</td>
</tr>
</tbody>
</table>

VII. CONCLUSIONS

Cloud computing is used to provide massive computing or storage resources to the user as they need. In this paper, we discuss about task scheduling that is a major concern in the cloud environment and various scheduling approach or algorithm which are proposed by the various researchers. We compared RASA, PSO, GA-ACO, PACO, Dynamic Scheduling Scheme and others as well. These schemes are centered on minimum completion time, execution time, minimizing makespan, waiting time, efficient use of resources etc. PACO, Novel Hybrid Bio-Inspired, GA-ACO, Modified GA, Multi-Queue Scheduling performed better than traditional
algorithm as per survey. In the future, our goal is to develop an scheduling algorithm that would be able to solve the issues of execution time, makespan and resource utilization.

REFERENCES


