Priority Based Job Scheduling Techniques In Cloud Computing: A Systematic Review

Swachil Patel, Upendra Bhoi

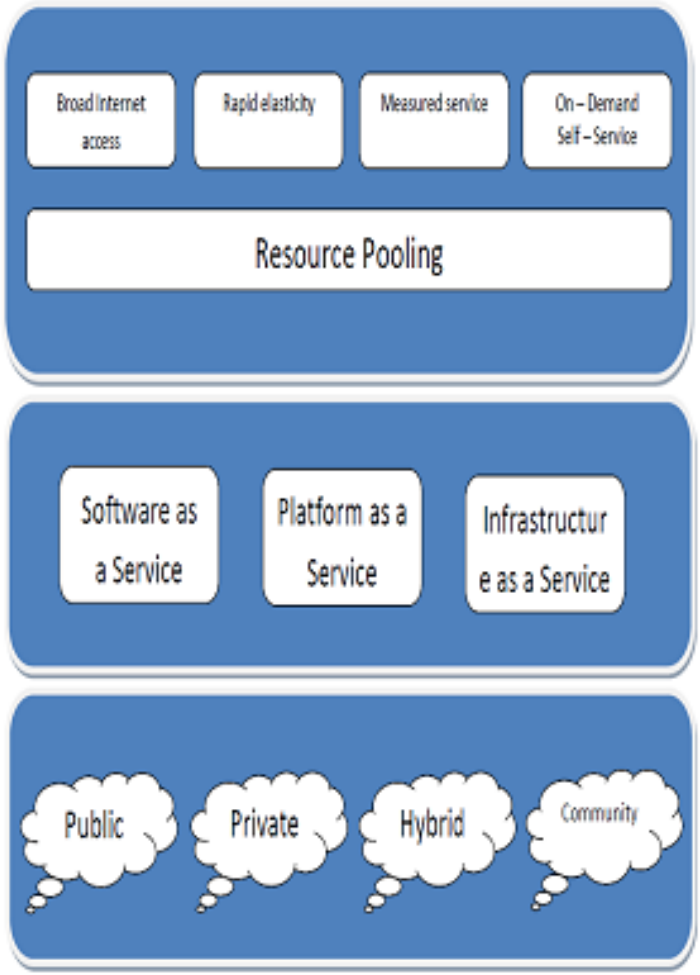
Abstract: Cloud Computing is the use of computing resources (Hardware and Software) that are delivered as a service over a network (typically the internet) to the customers. It intends to share large scale equipments and resources for computation, storage, information and knowledge for scientific researches [1]. In cloud computing, there are many jobs requires to be executed by the available resources to achieve best performance, minimal total time for completion, shortest response time, utilization of resource usage and etc. Because of these different objectives and high performance of computing environment, we need to design, develop, propose a scheduling algorithm to outperform appropriate allocation map of jobs due to different factors. In job scheduling priority is the biggest issue because some jobs need to scheduled first then the other jobs which can wait for a long time. In this paper, a systematic review of various priority based job scheduling algorithms is presented. These algorithms have different perspective, working principles etc. This study concludes that all the existing techniques mainly focus on priority of jobs and reduces service response time and improving performance etc. There are many parameters that can be mentioned as factor of scheduling problem to be considered such as load balancing, system throughput, service reliability, service cost, service utilization and so forth.

Index Terms: Cloud Computing, Job Scheduling, Priority, Quality of service, Service response time, Task Completion Time, Makespan

1 INTRODUCTION

Cloud computing is Distributed Computing paradigm which provides services to the customers. Cloud Providers provides services to their customers and charges as per usage by particular customer. That is, use as much or less you want to use, use services when you want to use and pay for only what you have used. Cloud computing is a construct that allows you to use applications that actually reside on a location different from your machine location. The cloud environment provides a different virtualized platform that helps user to accomplish their jobs with minimum completion time and minimum costs. Figure 1 shows the framework of cloud. In the cloud computing model, computing power, software, storage services, and platforms are delivered on demand to external customers over the internet. Cloud makes it possible for users to use services provided by cloud providers from anywhere at any time. The high growth in virtualization and cloud computing technologies reflect the number of jobs that are increasing nowadays, require the services of the virtual machine. Different types of job scheduling algorithms have been applied on different types of data workloads. And results are measured with different performance parameters to evaluate the performance.

FIGURE 1
CLOUD COMPUTING



Job-scheduling algorithms are developed to accomplish several goals like expected outcome, efficient use of resources, low makespan, high throughput, better quality of service, maintaining efficiency. In job scheduling algorithms, priority of jobs is a challenging issue because some jobs need to be serviced first than those other jobs which can stay for a

- *Swachil Patel is currently pursuing masters degree program in Computer science and engineering in Gujarat Technical University, India, PH-09586961585. E-mail: pswachil@gmail.com*
- *Upendra Bhoi is currently working as Assistant Professor in Computer Science & Engineering Department in Parul Institute of Technology,Gujarat,India. PH9427221244. E-mail: upendra.r.bhoi@gmail.com*

long time. Suitable job scheduling algorithm must consider the priority of a job [1]. In Fig 1, Cloud computing architecture is presented. Cloud services are divided into three types namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) respectively. Fig 1 shows the essential characteristics of cloud computing such as resource pooling, broad network access, elasticity, on-demand services, physical cloud resources (System Level) and middleware capabilities form the basis provider of delivering IaaS and PaaS in the form of a collection of transparently data centres and runtime environment and composition tools which ease the creation, deployment and execution process of application in the cloud. Finally, to provide the above mentioned services, deployment models such as Public Cloud, Private Cloud, Hybrid Cloud and Community Cloud are used by the cloud providers. The infrastructure of the cloud is provided publicly to all the general public by the organization in public cloud. Anyone can access services from anywhere publicly. Where, private cloud is used for a single organization only. Community Cloud is formed by several organizations and supports a specific community that has shared concerns for their future use. It might be managed by the any one of the shared organization or a third party organization. Last type is Hybrid Cloud, is a cloud formed by the composition of two or more clouds that is private, community, or public. Hybrid computing is bound together by standardized technology which enables data and application portability [10].

2 GUIDELINES OF SCHEDULING

Job Scheduling is used to allocate certain jobs to particular resources in particular time. In cloud computing, job-scheduling problem is a biggest and challenging issue. Hence the job scheduler should be dynamic. Job scheduling in cloud computing is mainly focuses to improve the efficient utilization of resource that is bandwidth, memory and reduction in completion time. An efficient job scheduling strategy must aim to yield less response time so that the execution of submitted jobs takes place within a possible minimum time and there will be an occurrence of in-time where resources are reallocated. Because of this, less rejection of jobs takes place and more number of jobs can be submitted to the cloud by the clients which ultimately show increasing results in accelerating the business performance of the cloud. There are different types of scheduling based on different criteria, such as static vs. Dynamic, centralized vs. Distributed, offline vs. Online etc are defined below:

1) ***Static Scheduling*:** Pre-Schedule jobs, all information are known about available resources and tasks and a task is assigned once to a resource, so it's easier to adapt based on scheduler's perspective [8].

2) ***Dynamic Scheduling*:** Jobs are dynamically available for scheduling over time by the scheduler. It is more flexible than static scheduling, to be able of determining run time in advance. It is more critical to include load balance as a main factor to obtain stable, accurate and efficient scheduler algorithm [8].

3) **Centralized Scheduling:** As mentioned in dynamic scheduling, it's a responsibility of centralized / distributed scheduler to make global decision. The main benefits of centralized scheduling are ease of implementation; efficiency and more control and monitoring on resources. On the other hand; such scheduler lacks scalability, fault tolerance and efficient performance. Because of this disadvantage it's not recommended for large-scale grids [9].

4) **Distributed / Decentralized Scheduling:** This type of scheduling is more realistic for real cloud despite of its weak efficiency compared to centralized scheduling. There is no central control entity, so local schedulers' requests to manage and maintain state of jobs' queue [12].

5) **Pre-Emptive Scheduling:** This type of scheduling allows each job to be interrupted during execution and a job can be migrated to another resource leaving its originally allocated resource, available for other jobs. If constraints such as priority are considered, this type of scheduling is more helpful [11].

6) **Non Pre-Emptive Scheduling:** It is a scheduling process, in which resources are not being allowed to be re-allocated until the running and scheduled job finished its execution [11].

7) **Co-operative scheduling:** In this type of scheduling, system have already many schedulers, each one is responsible for performing certain activity in scheduling process towards common system wide range based on the cooperation of procedures, given rules and current system users [9].

8) **Immediate / Online Mode:** In this type of scheduling, scheduler schedules any recently arriving job as soon as it arrives with no waiting for next time interval on available resources at that moment [3].

9) **Batch / Offline Mode:** The scheduler stores arriving jobs as group of problems to be solved over successive time intervals, so that it is better to map a job for suitable resources depending on its characteristics [3].

3 RESEARCH METHOD

This review aims at summarizing the current state of the art of various priority based job scheduling techniques in cloud computing.

4 SOURCE OF INFORMATION

The Search was widely conducted in the following electronic sources to gain a broad perspective:

- ScienceDirect (www.sciencedirect.com)
- IEEE eXplore (ieeexplore.ieee.org)
- Springer LNCS (www.springer.com/lncs)

These sources cover the most relevant journals, conferences and workshop proceedings. The searches in the selected sources resulted in overlap among the papers, where the duplicates were excluded by manual filtering.

5 SEARCH CRITERIA

The initial search criteria included the titles (dynamic task scheduling in grid computing using prioritized round robin algorithm), (a new class of priority based weighted fair scheduling algorithm), (job scheduling in clouds) and (job scheduling in datacentres). The start year set to 2002, and the end year was 2013. Only papers written in English were included.

6 VARIOUS PRIORITY BASED JOB SCHEDULING TECHNIQUES IN CLOUD

Following Priority based Job scheduling techniques are currently prevalent in clouds:

1) **Dynamic task scheduling in grid computing using prioritized round robin algorithm:** Sunita Bansal, Bhavik Kothari, Chitranjan Hoda [4] proposed a novel grid scheduling heuristic that adaptively and dynamically schedules task without requiring any prior information on the workload of incoming tasks. This models the grid system in the form of a state – transition diagram with job replication to optimally schedule jobs. This algorithm uses prediction information on processor utilization. In this algorithm they uses concept of job replication that is, a job can be replicated to other resource if that resource completes execution of current job than the resource it is currently allocated. This algorithm uses two types of queue namely, Waiting Queue and Execution Queue. This approach is based on exploiting information on processing capability of individual grid resources and applying replication on tasks assigned to the slowest processors. The approach facilitates replication of tasks, and also assigned to execute on slower machines, on machines with higher processing capacity. In this approach the communication cost are ignored. Experimental results show the better performance of this approach compared to traditional round robin algorithm.

2) **A new Class of Priority-based Weighted Fair Scheduling Algorithm:** Li Yang, ChengSheng Pan, ErHan Zhang, HaiYan Liu [5] proposed one kind of weighted fair scheduling algorithm. It is based on strict rob priority class which adds an absolute priority queue based on the foundation of based class weighted fair scheduling algorithm (CBWFQ). This algorithm covers the disadvantage of traditional weighted fair scheduling algorithm. Weighted Fair Scheduling algorithm differentiates the services of all active queues on the basis of weight of each business flow. When a new job arrives the classifier classifies the jobs into categories. Then buffer is checked for each category and if buffer is not overloaded then job is stored in the buffer otherwise job is dropped. Each job enters a different virtual queue. Weight, Dispatch, Discard and Rob are four main rules of this algorithm. The main advantage of this algorithm is that it has introduced the rob rule together with dropping rule. Experiments are done on NS-2 software to simulate SRPQ-CBWFQ algorithm. This new algorithm combined buffer management and queue scheduling and only guarantees low delay of real time applications. It also gave consideration to fairness and better utilization of buffers. This algorithm has two great advantages of bandwidth allocation and delay without throughput reducibility.

3) **A Priority based Job Scheduling Algorithm in Cloud Computing:** Shamsollah Ghanbari, Mohamed Othman [1] presented a novel approach of job scheduling in cloud computing by using mathematical statistics. This algorithm considers the priority of jobs for scheduling and named as priority based job scheduling algorithm. It is based on multiple criteria decision making model. A pairwise comparison based on multiple criteria and multiple attributes method was first developed by Thomas Saaty [13] in 1980 and named as Analytical Hierarchy Process (AHP). Consistent Comparison Matrix is the foundation of AHP, so to use the concept of AHP comparison matrices are computed according to the attributes and criteria's accessibilities. In this algorithm, each job requests a resource with determined priority. So comparison matrices of each jobs according to resources accessibilities is computed and also comparison matrix of resources is computed. For each of the comparison matrices priority vectors (vector of weights) are computed and finally a normal matrix of all jobs is computed named as Δ. Likewise, normal matrix of all resources is also computed and name of that matrix is 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. The final step of the algorithm is to choose the job with maximum calculated priority, so a suitable resource is allocated to that job. The list of jobs is updated and the scheduling process continues till all the jobs are scheduled to suitable resource. Experimental results indicate that the algorithm has reasonable complexity. Also there are several issues related to this algorithm such as complexity, consistency and finish time.

4) **Agent based Priority Heuristic for Job Scheduling on Computational Grids [14]:** This algorithm presents an agent based job scheduling for effective and efficient execution of user jobs. This considers QoS parameters like waiting time, turnaround time, response time, total completion time, etc. Priorities are assigned to the jobs under different classifications. Agent based Heuristic Scheduling (AHS) uses task agent for job distribution to achieve optimum solution. Task agent receives jobs from users and distributes them among different prioritize global queues based on user levels. AHS uses agent based job distribution strategy at global level for optimal job distribution based on user levels and job priorities at local levels for efficient and effective execution of jobs. For different global queues, priorities are defined as threshold levels for assigning jobs to global queues. If jobs have same priorities then jobs having minimum run time executes first otherwise First Come First Serve (FCFS) algorithm is used. AHS has optimal performance with respect to QoS parameters.

5) **Design, Development and performance analysis of Deadline based Priority Heuristic for Job scheduling in a Grid [15]:** A modified prioritized deadline based scheduling algorithm (MPDSA) is

proposed using project management algorithm for efficient job execution with deadline constraint of user's jobs. MPDSA executes jobs with closest deadline time delay in cyclic manner using dynamic time quantum. It assumes each job to be described by its process_id, burst_time, arrival_time and deadline. Time quantum is assigned by computing LCM of all burst times. Then the jobs having minimum time delay is selected for execution. If jobs have same time delay then First Come First Serve (FCFS) algorithm is used for scheduling. Jobs are pre-empted based on time quantum and if a job completes its execution before time quantum, that job is deleted from queue. This algorithm satisfies system requirements and supports scalability under heavy workloads.

6) **A two-stage-priority-rule-based algorithm for robust resource-constrained project scheduling [16]:** The algorithm solves the resource-constrained project scheduling problem (RCPSP). This algorithm presents a two-stage algorithm for robust resource-constrained project scheduling. First stage solves the RCPSP for minimizing makespan by using a priority-rule-based heuristic. Second stage is intended to find most robust schedule with a makespan not larger than threshold value found in first stage. Both the stages are referred as two phases. In phase I, each iteration has three steps:

   1. Priority values issued from selected priority rule.
   2. Random-biased selection of eligible activities according to their selection probabilities.
   3. Selected activities are scheduled to the resources.

In phase II, same number of iterations is executed as in phase I. Each iteration starts by execution of forward recursion which allows determining project makespan. Backward recursion is carried out to obtain latest completion time of each job. The above step is performed only if makespan is not larger than threshold value calculated in phase I.

7 Comparison of Priority Based Job Scheduling Algorithms [7]

The comparison table of priority based job scheduling algorithms based on different factors is shown in Table 1.

8 Conclusion and Future Work

Priority based scheduling is one of the key issues in the management of jobs execution in cloud environment. In this paper, I have surveyed the various existing priority based job scheduling algorithm in cloud computing and tabulated their various parameters such as Cost, Makespan, Resource utilization, Scalability and so on. In addition to these algorithms, there are many studies related to and based on them that are searching for improvements, job scheduling, optimization and etc. There are certain aspects that should be considered as topics of research to introduce more accurate and improved algorithms rather that those introduced here such as the arriving rate of the jobs, cost of the job execution on each of the resource, cost of the communication, less makespan etc. Prioritized Round Robin Algorithm and Priority based Weighted Fair Scheduling Algorithm considers certain (especially one of the) parameters of jobs to schedule whereas Priority based Job Scheduling Algorithm considers all the parameters of jobs to perform scheduling. Improving one of the main parameter like makespan, throughput and consistency of the existing algorithm is considered as a future work.

TABLE 1
EXISTING PRIORITY BASED JOB SCHEDULING ALGORITHM

Algorithm			Complexity		Focus On			Scalability		Recommended in
Name	Method	Factor	Cost	Time	Makespan	Resource utilization	Speed	Dynamic	Static	
Prioritizes Round Robin Algorithm	Batch Mode	Meta tasks	x	√	x	√	x	x	√	Grid
Priority based Weighted Fair Scheduling Algorithm	Batch Mode	Meta tasks	x	√	x	√	√	x	√	Cloud
Priority based Job Scheduling Algorithm	Batch Mode	Meta tasks	√	√	√	√	x	x	√	Cloud
Agent based Priority Heuristic	Batch Mode	Meta tasks	√	√	√	x	x	√	x	Grid
Deadline based Priority Heuristic	Batch Mode	Meta tasks	x	√	√	√	x	√	x	Grid
Two stage priority rule based algorithm	Batch Mode	Meta tasks	x	√	x	√	x	x	√	Grid

ACKNOWLEDGEMENT

Swachil Patel would like to thank to my thesis guide Asst.Prof. Upendra Bhoi for his great effort and instructive comments in this paper work. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this paper.

REFERENCES

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

[2]. Isam Azawi Mohialdeen, Comparative Study of Scheduling Al-gorithms in Cloud Computing Environment, Journal of Computer Science, 9 (2): 252-263, 2013.

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

[4]. Sunita Bansal et al, Dynamic Task-Scheduling in Grid Computing Using Prioritized Round Robin Algorithm, IJCSI International Journal of Computer Science Issues, 8(2)(2011) 472-477.

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

[6]. Ioannis A. Moschakis et al, Performance and cost evaluation of Gang Scheduling in a Cloud Computing System with Job Migrations and Starvation Handling, 978-1-4577-0681-3/11/$26.00 ©2011 IEEE, pp. 418 – 423.

[7]. O. M. Elzeki, M. Z. Rashad, M. A. Elsoud, “Overview of Scheduling Tasks in Distributed Computing Systems”, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

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

[9]. 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.

[10]. R.Madhubala, "An Illustrative Study on cloud computing", "International Journal of Soft Computing and Engineering", Vol.1, issue.6, January 2012, pp. 286-290.

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

[12]. 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.

[13]. T.L. Saaty, The Analytic Hierarchy Process,(New York 1980).