IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

A Review on Resource Scheduling Algorithms for Big Data Workflows in Cloud

Arunkumar Panneerselvam[1], Bhuvaneswari Subbaraman[2]

[1]*Research Scholar, Dept. Of Computer Science, Pondicherry University, Karaikal Campus*
[2]*Registrar, Central University of Tamil Nadu, Thiruvarur, Tamil Nadu*

***Abstract*— Today Big Data Processing has find its importance in various sectors of business. Big Data Analytics strengthens the businesses worldwide by providing valuable insights about the products, services, market trends and customers of a business. This gives way for the investigation of technological platforms for big data processing. Being scalable and elastic in nature with abundant virtual resources cloud forms naturally the right choice for big data processing. This Paper aims to explore the resource scheduling algorithms for big data workflows in cloud. The taxonomies of scheduling algorithms for scientific workflows are broadly classified as traditional rule based, Heuristic Scheduling Algorithms, Meta Heuristic Algorithms, Hybrid Algorithms and Mathematical Models for Scheduling.**

***Keywords*— Big data workflows, Cloud Resource Provisioning, Scheduling algorithms, Scientific workflow, optimization techniques.**

I. INTRODUCTION

The workflows are ordered set of tasks that is represented in Directed Acyclic Graph (DAG). The workflows represented as DAG will contain a start task and end task with many non iterative intermediate tasks. In this model one task depends on another task for its execution which means that the output of the predecessor task is used as input for the successor task for execution. A task cannot be executed until its previous task completes to execution. It is always easy and convenient to represent the flow of complex and scientific applications in terms of workflows. Big data workflows are class of scientific workflows since the nature of the tasks involved in the workflows are data intensive and compute intensive. Scheduling represents the allocation of cloud resources required for the workflow task to complete its execution. The scientific workflow management systems are used to create, manage and execute the workflows and many scientific workflow management systems are now integrated with cloud for better execution of the big data applications. The rest of the paper is organized as follows. Section II details about the nature of the big data workflows and challenges involved in scheduling big data workflows in cloud. Section III explores the important objectives, constraints and gap of the scheduling algorithms. Section IV quickly reviews the traditional algorithms for resource scheduling in cloud. Section V elaborates on the literature of Heuristic and Meta Heuristic approaches for resource scheduling for scientific workflows. Section VI explores the hybrid and mathematical models used for scheduling cloud resources. Section VII gives conclusion and future directions for big data workflow scheduling.

II. BIG DATA WORKFLOW CHALLENGES

As opposed to normal workflows the big data workflows posses certain special characteristics. The big data workloads are highly data intensive and compute intensive. The tasks in the workflow receive massive data input from various data sources and also it uses several servers to process and store data. It is very difficult to predict the amount of data incoming forehand and how much resources would be used by the workflows to execute the big data task. The format of data is unpredictable and thus the number and type of virtual data sources needed are highly dynamic. Next the big data analytics requires parallel processing in many servers and how much virtual servers needed, its type and when it is needed also dynamic. The decision of the resource allocation cannot be static and it should be decided dynamically. The execution of scientific workflows is challenging in terms of data scaling, computational complexity, dynamic resource allocation and issues with collaboration in heterogeneous environment [1]. Further the cloud provider should provide robust workflow application integration with their cloud environment to gain the user satisfaction and cost benefit.

III. OBJECTIVES, CONSTRAINTS AND GAPS

The two important objective of Scientific workflow scheduling in cloud is the time and cost. The submitted workflows must be completed before the deadline by utilizing less cost for execution. There should always a trade off between time and cost.

Fast execution is possible with utilization of high speed virtual resources but unfortunately leasing high speed virtual resources will increase cost for the cloud user who is doing business under budget constraints. This trade off is applicable both for cloud service user and the cloud resource provider. The important constraints that control scheduling algorithms are deadline, SLAs, budget, execution time and cost. These constraints should not be violated while designing a scheduling algorithm for resource provisioning. All the algorithms experimented so far to schedule workflows in cloud environment are basically experimented for Grid environment and the same is applied to cloud. Unfortunately these algorithms do not take into account of the characteristics of cloud such as scalability, elasticity, dynamic nature of the resources etc. Also very few research work addresses the scientific workflows in multi cloud environment and Big Data Workflows execution in cloud is in very infant stage.

Maria Alejandra Rodriguez and Rajkumar Buyya has elaborated the taxonomy of scheduling algorithms for scientific workflows in IaaS cloud [2] and are listed in the table below

TABLE I
SCHEDULING ALGORITHM TAXONOMIES EXPLORED BY MARIA AND BUYYA

Taxonomy	Strategy Classifications	Strategy types
Application model taxonomy	Workflow Multiplicity	Single workflow, Multiple workflows, Workflow Ensembles (grouping of inter related workflows)
Scheduling model taxonomy	Task-VM Mapping Dynamicity	Static, Dynamic, Hybrid
	Resource provisioning strategy	Static VM pool, Elastic VM pool
	Scheduling Objectives	Cost (Budget, minimization of cost) Makespan (Deadline, Makespan minimization)
	Workload maximization	
	VM Utilization Maximization	
	Energy Consumption Minimization	
	Reliability Awareness	
	Security awareness	
	Optimization Strategy	Optimal, Sub Optimal (Heuristic, Meta Heuristic, Hybrid)
Resource model taxonomy	Provider	VM leasing model, VM type uniformity, Deployment model
	Storage and Network	Intermediate Data Sharing model, Data transfer cost awareness, storage cost awareness
	Virtual Machine	VM Pricing Model, VM delays, VM Core Count

While designing the algorithms for scientific workflows for cloud environment the above taxonomies can be considered. For example, one can think of whether the designed algorithm is taking into account of static virtual machines pool or virtual machine pool with dynamic VMs. Likewise when designing algorithm with a goal of minimizing the Makespan (total execution time of the workflow) whether the algorithm is taking account of VM delays or not. And thus various research gaps can be addressed satisfactorily.

IV. Traditional Algorithms

The basic resource scheduling algorithms used for allocating suitable resources to workflow tasks are Round Robin (RR), First Come First Serve (FCFS) algorithms.

A. Round Robin (RR) Scheduling

The Round Robin scheduling algorithm is a simple algorithm used to allocate CPU cycles of a virtual machine with static time quantum for each task. The drawback of the algorithm is it is not dynamic in nature. Recently Smarter Round Robin Scheduling Algorithm for cloud computing and big data has been proposed by incorporating Dynamic time quantum to the tasks depending on the context [3].

B. Firs come First Serve (FCFS) Scheduling

The First come First Serve scheduling allocates resource for the task that arrives first. This algorithm is a basic algorithm in many scheduling applications. The other variation of FCFS is Shortest Job First FCFS which allocates the resource to the First Shortest job that is having less burst time in the queue. The FCFS scheduling is not fit for workflows with Parallel computation and complex tasks execution.

V. Heuristic And Meta Heuristic Approach

The Resource scheduling problem is classified as NP-Hard and hence it is difficult to find exact optimal solution in a polynomial time. Hence the heuristic and Meta heuristic approaches are more popular in cloud resource scheduling for big data workflows.

A. Heuristic Algorithms

Heuristic algorithms are mostly problem dependent and produce good optimal solutions for larger problems in considerable amount of time. However the algorithm constrained to local optimum solution. The following heuristic based algorithms are used for workflow scheduling in cloud environment.

1) *Min-Min heuristic:* This algorithm iteratively schedules a set of task by computing ECT (Early Completion Time) of each task on its every available resource and thus obtains the MCT (Minimum Estimated Completion Time). The task with minimum MCT is selected to schedule first. Thus the selected task is assigned on the resource which is expected to finish it at first [4].

2) *Max-Min heuristic:* This heuristic approach is similar to Min-Max, but it sets high scheduling priority to task which have long execution time [4].

3) *Heterogeneous Earliest Finish Time (HEFT):* In this heuristic method average execution time for each task and average communication time between two resources for two dependent tasks are calculated. Then each task is assigned with a rank value which is calculated recursively based on the rank value of the following dependent task The exit task in the graph will have the smallest rank value which being the average execution time. The predecessor task of the exit task will have their average execution time and the maximum time. The task with the highest priority will be executed first [5]. The Multi-objective Heterogeneous Earliest Finish time (MOHEFT) is the extension of HEFT which computes pareto-based solution and produces several intermediate workflow schedules as opposed to HEFT which propose single schedule [6].

A. Meta Heuristic Algorithms

We can say Meta Heuristics as heuristics of heuristics. These algorithms are independent of problems and are generic in nature. The Meta heuristics approach can be applied to wide variety of problems. The Meta heuristics method provides global optimal solutions. The Meta heuristics algorithms based on bio inspirations and swarm intelligence are very widely explored now a days. The following are some of the explored Bio inspired and swarm intelligence algorithms.

- Particle Swarm Optimization (PSO)
- Ant Colony Optimization (ACO)
- Genetic Algorithm (GA)
- Artificial Bee Colony Algorithm (ABC)
- Cuckoo Search Algorithm (CSA)
- Fish Swarm Optimization (FSO)
- Cat Swarm Optimization (CSO)
- Bat Algorithm (BA)
- Intelligence Water Drop Algorithm (IWD)
- Harmony Search (HS)
- League Championship Algorithm (LCA)
- Lion Optimization Algorithm (LOA)
- Simulated Annealing (SA)

Although there are many algorithms the research community extensively experiments Ant Colony Optimization (ACO), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) and their variants for resource scheduling in cloud computing environment. Hence one who experimenting with resource provisioning algorithm with any new Meta heuristic algorithm can compare their findings with ACO, GA and PSO to check their performance of their algorithms.

Mala Kalra and Sarbjeet Singh has compared the ACO, GA and PSO techniques [7] and their findings are presented below in table

TABLE II
COMPARISON OF ACO, GA AND PSO ALGORITHMS

Algorithm	**Performance Metrics**	**Nature of Tasks**
ACO and its improvement algorithms	Makespan, Load Balancing, Reliability, Execution cost, Deadline Constraint, Resource Utilization Ratio, Energy Conservation and Consumption, SLA	Independent, Workflows, Virtual Machine Placement
GA and its improvement algorithms	Makespan, Flow time , Load Balancing, Reliability, Availability, Execution cost, Deadline Constraint, Budget Constraint, Resource Utilization Ratio, Energy Conservation and Consumption,	Independent, Workflows, Virtual Machine Placement
PSO and its improvement algorithms	Makespan, Flow time , Communication time, Load Balancing, Communication cost, Execution cost, Deadline Constraint, Budget Constraint, Resource Utilization Ratio, Energy Conservation and Consumption	Independent, Workflows, Virtual Machine Placement

In the above table Independent task represents the tasks which do not depend on completion of other task for execution. These tasks can be scheduled in parallel to the suitable virtual machines whereas workflow indicates the task dependency for execution. Virtual Machine (VM) Placement refers to placing of VMs in available computing servers.

VI. HYPER HEURISTIC ALGORITHMS AND MATHEMATICAL MODELS

Hyper Heuristic algorithms are combination of two or more Meta heuristics algorithms.

The hyper heuristic algorithms are hybrid in nature and reap the benefits of individual heuristic algorithms in combination to solve complex problems. A fewer studies has been made in hyper heuristics algorithms with respect to resource scheduling for scientific workflows in cloud. ACO and PSO are combined to form a hyper heuristic algorithm [8]. PSO and GA has been hybrid for cloud resource scheduling [9]. ACO and GA are used to form hyper heuristic algorithms for task allocation [10]. The Mathematical Models are mainly used for cost optimization in resource scheduling. Algebraic Mathematical models such as linear and non linear programming model, Integer programming model and Mixed Integer linear Programming models have been proposed in scientific workflow scheduling.

VII. CONCLUSION AND FUTURE DIRECTIONS

This paper explores the various algorithms like rule based heuristic, Meta heuristic and Hyper heuristic approaches that have been adapted by the researchers for resource scheduling in cloud computing. However all the works considers only normal workflows execution and more concentration is needed on Big Data workflow resource scheduling in cloud. Also the work done so far by the research community addressed only few taxonomies and strategies of scientific workflow scheduling in cloud. Our future research directions will concentrate on designing powerful scheduling algorithms which take into account of characteristics of cloud environment and the nature of big data applications. Prediction and Machine learning approaches in combination with Meta heuristics algorithms will yield more advantages to the Business community.

REFERENCES

[1] Yong Zhao, Youfu Li, Siyiong Lu, Ioan Raicu and Cui Lin, 2014. Devising a cloud scientific Workflow Platform for Big Data, IEEE World Congress on Services.

[2] Maria Alejandra Rodriguez, Rajkumar Buyya, 2017. A taxonomy and survey on scheduling algorithms for scientific workflows in IaaS, Concurrency and Computation: Practice and Experience, volume 29, issue 8, John Wiley & Sons, Ltd.

[3] Hicham Gibet Tani, Chaker El Amrani., 2017. Smarter Round Robin Scheduling Algorithm for Cloud Computing and Big Data, https://hal.archives-ouvertes.fr/hal-01443713.

[4] Mihaela-Catalina, Mihaela Vasile, Florin Pop and Valentin Cristea, 2016. Workflow Scheduling Techniques for Big Data Platforms, Chapter 2, Resource management for Big data Platforms, pg. 35-53, Computer Communications and networks, Springer International Publishing.

[5] Topcuoglu H., Hariri S., Wu M.Y., 2002. Performance-effective and low-complexity task scheduling for heterogeneous computing, IEEE transactions on Parallel Distributed Systems, volume 13(3), pg. 260-274.

[6] Durillo J.J, Nae V, Prodan R., 2014. Multi-Objective energy-efficient workflow scheduling using list-based heuristics, Future generation computing system, volume 36, pg. 221-236.

[7] Mala Kalra, Sarbjeet Singh, 2015. A review of Met heuristics scheduling techniques in cloud, Egyptian international Journal, volume 16, pg. 275-295.

[8] Junliang Lu, Wei Hu, Yonghao Wang, Lin Li, Peng Ke and Kai Zhang, 2016. A Hybrid Algorithm based on Particle Swarm Optimization and Ant Colony Optimization, International conference on Smart computing and Communication, pg. 22-31.

[9] Sridhar M, 2015. Hybrid Genetic Swarm Scheduling for Cloud computing, Global Journal of Computer Science and Technology: Cloud and Distributed, Volume 15, Issue 3, Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[10] Sawsan Yousef Abu Shuqeir and Tamara Amjad Al Qublan, 2014. Hybrid Algorithm based on ANT and Genetic Algorithms for Task Allocation on a Network of Homogeneous Processors, International Journal of Computer Networks & Communications (IJCNC) Vol.6, No.1, pg. 191-202.