Proceedings of
INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND ENERGY SYSTEMS (ICCCES-16)
In Association with IET, UK & Sponsored by TEQIP-II
29th -30th, Jan. 2016

Workflow Scheduling of Scientific Application in Cloud – A Survey

Priyanka M. Kadam[1]
Priyankakadam222@gmail.com

Prof. S. R.Poojara.[2]
Assistant Professor
shivananda.poojara@ritindia.edu

Prof. N.V.Dharwadkar.[3]
Head Of Deparment
nagraj.dharwadkar@ritindia.edu

Dept. of Computer science and Engineering.[1,2,3]
Rajarambapu Institute of Technology, Sakhrale, India

***Abstract*–In distributed computing environment like clouds and grid workflow is used as popular mechanism for its ability of expressing range of wide applications. With developing the cloud technology and platform, the problem of scheduling workflow in cloud is tough and it is important research topic. Cloud resources and services can be accessed at anytime and anywhere via its users. The science and engineering field consist of large and complex application and take more time for execution and data transmission. So it is necessary to schedule such scientific applications. Some commonly used algorithm for scheduling workflow is ant colony optimization, Particle swarm optimization, Genetic algorithm are focused on a particular problem.**

Keywords—cloud computing, Workflo scheduling, Science workflow scheduling.

1. Introduction

In this report we considered couple of research papers related to the Workflow scheduling and workflow scheduling of scientific applications. Lots of applications consist the number of tasks which that needs more computing power more than single machine capability. The applications include workflows of scientific application, workflows of multi-tier web service, and workflows of big data processing which include MapReduce. These applications can be described using workflow method.

1.1: *Workflow*

Workflow is a series of steps or series of actions are connected together, and input to the next step is an output of previous step. The applications which require more consuming power that are represented using workflow.

1.2: Scientific Workflow

The size of scientific applications is large as compare to the normal applications. It contains millions of tasks. The workflow of scientific applications allows end user to describe multi-step computational task easily. In case of large Workflow the tasks are distributed over the multiple computers and necessary to complete the work in minimum or reasonable time.

Why Workflow Scheduling of Scientific Application?

The problem of workflow scheduling is studied from past years, focusing on distributed environment such as clusters, grids. The field includes science and engineering consist of most applications which are very large in size and are complex. These applications are represented using directed acyclic graph. The example of scientific application includes Montage, Broadband, Bioinformatics, Epigenomics etc. The data size of these applications is continuously increasing. Such a applications takes more time for data transmissions so there is needs of scheduling of workflow of scientific applications [1][2][7].

2. Workflow Model and Definition

2.1: Workflow Model

The workflow application can be represented in the form of directed acyclic graph:

G(T,E), where T is set of tasks {t1,t2…tv} and E is a set of edges$\{e_j^i \mid$(ti,tj) $\in E\}$ represent inter task dependencies, ti represent individual application task with computational workload wli. Each edge indicates that task tj can start after complete execution of task ti. If there is data transmission from *ti* to *tj* then *tj* can start only after all the data from *ti* has been received.d_j^i Associated with edge e_j^i denote amount of data transferred from *ti* to *tj* [1].

2.2: *Definition*

Assume that each resource is associated with a computation property, as million instructions per second (MIPS) for example. The startup time of communication of resource rl is lr , and the bandwidth of data transmission between resources rl and rk is bw(rk, lr) [1]. The Execution time and transmission times can be calculated as:

$$ET_{ti}^{rl} = \frac{wli}{capability(rl)}$$

$$TT_{ti}^{tj} = \{lr(ti) + \frac{d_i^j}{bw_{r(ti)}^{r(tj)}} \quad r(ti) \neq r(tj);\ 0 \text{ otherwise.}$$

Where,

ET_{ti}^{rl} The execution time of task ti on resource ri

TT_{ti}^{tj} The transmission time from task ti to task tj

r(ti) The resource where ti execute on

r(tj) The resource where ti execute on

3. Cloud Workflow System

3.1: Workflow in Cloud

In grid and cloud distributed environment workflow is used as powerful paradigms for its ability of describing applications in wide range. Resources are

Providing in the clouds as a service by which it can be accessed by user at anywhere and when they want (anytime). In cloud environment user want to Access the service on demand and pay per you go mode and access resources in minimum time [1].

Workflow scheduling uses cloud provider such as Platform as a service (Paas) Infrastructure as a service (Iaas). Virtual pool of unlimited resources

Provides by the Iaas clouds, by which resources can be accessed on demand. In Paas clouds logical scheduling is important instead of resource scheduling. The

Combination of Platform as a service and Infrastructure as a service clouds built complete structure ofworkflow scheduling [1].

3.2: Cloud Workflow System Architecture

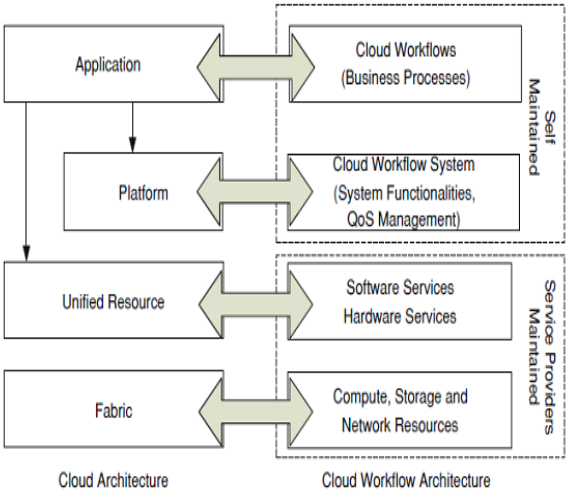

Fig 3.1: Cloud workflow system architecture

Figure 3.1 shows the architecture of cloud workflow system. There are four layers. The real-world business processes workflow applications are contained in the application layer of cloud. The platform provides implementation and running platform for cloud workflows. The business organization self-maintains the application layer and platform layer. The unified resource layer consists of software and hardware services which necessary to run the cloud workflow. SaaS provide software for developing business tasks and PaaS provide resources on demand to process business tasks [6].

4. Literature Survey

Representation of workflow applications are generally in the form of directed acyclic graph. The Assigning job or task to the compute resources is the problem called *NP-complete* problem. There are two cases: (1) with the weights to number of processors, (2) Jobs are scheduled with assigning weights is equal to one or two units to two processors. So, previous work has proposed using many heuristics approaches to schedule applications workflow.

Deelman [10] have done considerable work on mapping, planning, and data-reuse in the workflow scheduling area. They proposed Pegasus; it is a framework that maps onto distributed resources such as the Grid, complex scientific workflows. Pegasus with DAGMan, tasks are scheduled on Condor system. The Taverna project [4], has implement composition and enactment tool for the bioinformatics workflows for the life science community.

As per we know scheduling of task is a NP-Complete problem, and Genetic Algorithm (GA) is used for workflows scheduling. But GA may not be the best heuristic. [4] Shown that the PSO

performed faster than the GA in solving task assignment problem in distributed environment.

In [7] propose the algorithm for scientific workflow. It compares genetic algorithm and HEFT (Heterogeneous Earliest Finish Time) algorithm. It shows HEFT performs better than the GA in ASKALON Grid Environment.

In [11], the ant colony optimization algorithm is the algorithm which is used to solve computational problems. The problems which are complex are needs to reduce into simpler ones to find out the good optimized path through the graphs. The base of the algorithm is how ants search a path between the source of food and colony. This algorithm is also called Meta-heuristic. The algorithm implementation consists of a no of steps. First initialized the heuristic information, which following's generation. Next step need to map the ants with the path and evaluated the objective.

In [12], Genetic algorithm produces solution which is optimal within polynomial time involved different techniques from a large search space. Genetic algorithms use the operations mutation, selection, and recombination. A problem which necessary an optimized solution and implemented within the deadline depends on genetic algorithms.

In [4], the author describes working of Particle swarm Optimization and how it is used to schedule workflows. This algorithm is based on population developed by Dr. Eberhart and Dr. Kennedy in 1995. This algorithm is influenced by social behavior of animals or birds of fish flying through the search space. [13] Proved algorithm PSO provides 57 out of 90 better solution than the algorithm which are known for the problems.

In paper [3], some scientific applications are tested using the Pegasus workflow management system the applications includes Montage, Broadband and Bioinformatics. Pegasus contains the set series of technology that help to execute the applications which are based on workflow on various environments including cloud gird clusters etc. Pegasus WFMS make plan to execute application using DAGManto track dependencies and release tasks as they become ready,

5. Particle Swarm Optimization Heuristic

5.1: Introduction

Large amount of data and intensive activities are necessary to processed by the scientific workflow. A scientific workflow management system is used to manage scientific experiments. In Cloud computing environments, scheduling application has, policies differ according to the function: execution time should be minimize, minimize total cost of execution, while meeting the deadline constraints, balance the load on resources used, and so on.Particle Swarm Optimization (PSO) is a global search optimization approach which is introduced by Kennedy and Eberhart. PSO algorithm is influenced by the behavior of animals. For example the flock of bird searches for a food in an area. All birds know there is only one food resource in that area but they don't know actually where the food is. Then they use strategy to find food is following the bird that is nearest to the food in the area. A particle in PSO is similar to a bird flying through a search space. Each particle movement is co-ordinate by a velocity has direction and magnitude. Position of each particle at is influenced by its best position and the best particle position in a search space [4]. Velocity and position of particle is calculated using the following equations:

$$v_i^{k+1} = w v_i^k + c_{1rand_1} * (pbest_i - x_i^k) + c_2 rand_2 * (gbest - x_i^k) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

5.2: PSO Algorithm:

1. Set dimension of particles to d.
2. Initialize the population of particles with random position and velocity.
3. For each particle, calculate its fitness value
   - 3.1. Compare the particles fitness value with the particles *pbest*.If the current value is better than pbest then set pbest to current value and location.
   - 3.2. Compare the particles fitness value with the global best *gbest.* If the particles current value is better than *gbest* then set *gbest* to the current value and location.
   - 3.3. Update the position and velocity of the particle according to equations 1 and 2
4. Repeat from step 3 until the stopping criterion is met.
6. Genetic Algorithm

Genetic algorithm is heuristic approach for finding global minima from a search space. The basis of GA is to encode the possible solution into population of chromosomes and then transform population using the operations selection, crossover and mutation [7].

Algorithm:

- Create the initial population of chromosomes;

While convergence criteria is false do

- Perform crossover and mutation;
- Calculate fitness values for the population;
- Create a new population, based on actual fitness

Values;
End

Selection operation is performed to form the fitness function which computes the chromosomes in the form of accuracy. The operation crossover and mutation is analogous to the biological operations which mutually exchange body part within the couple of chromosomes.

The final condition of GA is basically the convergence criteria which check how many best fitness values found changes within the iterations. GA can establish maximum number of iterations.

7. HEFT Algorithm

HEFT is a Heterogeneous Earliest Finish Time algorithm schedule the workflow based on making orders list of task of workflow and assigns the task to the resources in appropriate manner [7].

The HEFT algorithm that we applied consists of 3 phases:

1. Weighting assigns the weights to the nodes and edge sin the workflow;
2. Ranking creates a sorted list of tasks, organized in the order how they should be executed;
3. Mapping assigns the tasks to the resources.

Assign the weight to the each node is based on the predicted time of execution and assign the weight to the each edge based on the predicted data transfer time between the resources.

The ranking phase traverses the graph of workflow upward and assigns the value of rank to each task. Rank value is the weight of node plus execution time of successor. Execution time of successor is estimated for each node which is immediate node of successor. Add its weight into value of rank of successor node and choose the summation which is maximum.

The last phase is mapping in which for all task the resources provide the earliest time to finish the execution is taken.

8. Case Study: Scientific Workflows

8.1: Applications Tested

Three scientific applications are chosen for this study. The applications include Montage, Broadband, and Bioinformatics.

Montage [9] creates science-grade astronomical image mosaics from data collected using telescopes. The Montage workflow size (Figure 8.1) depends upon the sky area (in square degrees) covered by the output mosaic.

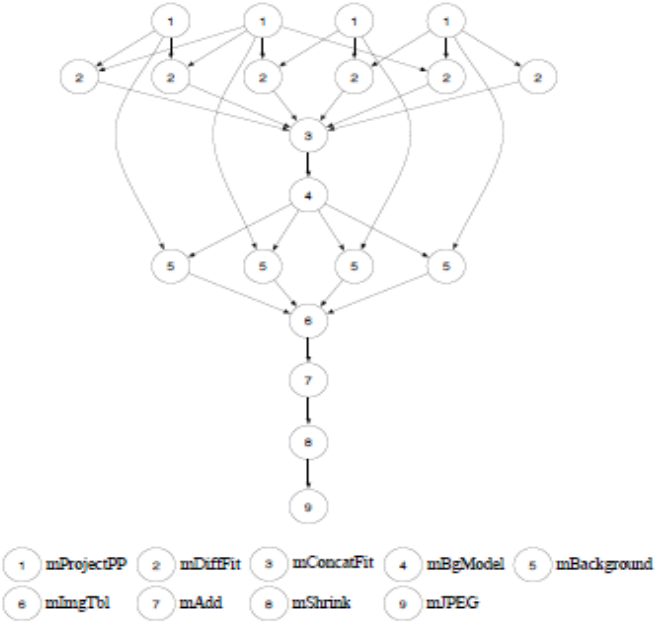

Fig 8.1: Montage workflow.

Broadband [9] forms and compares seismograms from several high- and low frequency earthquake simulation codes. Each Broadband workflow (Figure 8.2) generates seismograms for several sources (scenario earthquakes) and sites (geographic locations). For each (source, site) combination the workflow runs several high- and low-frequency earthquake simulations and computes intensity measures of the resulting seismograms.

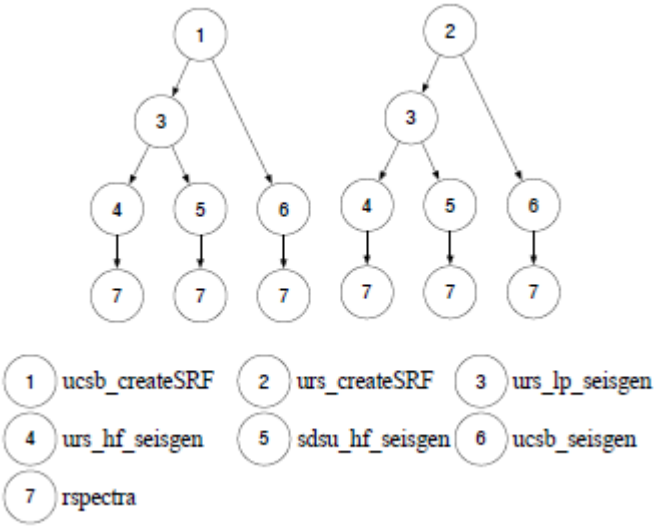

Fig 8.2: Broadband Workflow.

Epigenome [9] maps short DNA segments collected using high-throughput gene sequencing machines to a previously constructed reference genome using the MAQ software. The workflow (Figure 8.3) splits several input segment files into small chunks, reformats and converts the chunks, maps the chunks to a reference genome, merges the mapped sequences into a single output map, and computes the sequence density for each location of interest in the reference genome.

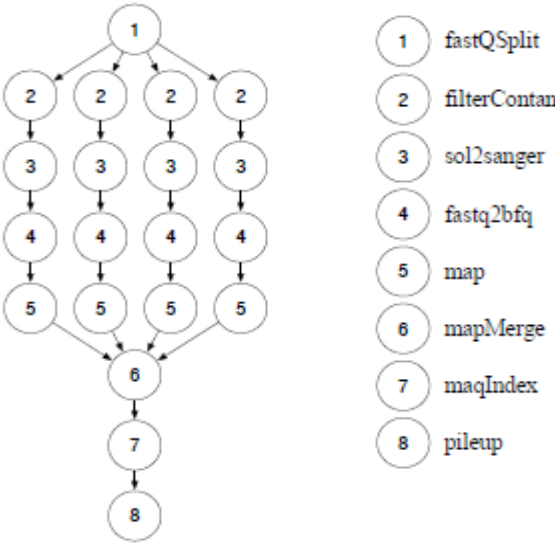

Fig 8.3: Epigenome Workflow.

8.2: Software

All workflows were planned and executed using the Pegasus WMS. Pegasus is used to transform abstract, resource-independent workflow descriptions into concrete, platform-specific execution plans. These plans are executed using DAGMan to track dependencies and release tasks as they become ready [9].

8.3: Experimental Results

In [7] experiments, comparison of HEFT algorithm with a genetic algorithm. The results show execution times of the scheduled workflow applications (execution times), and the times spent in preparing the schedules (scheduling times).

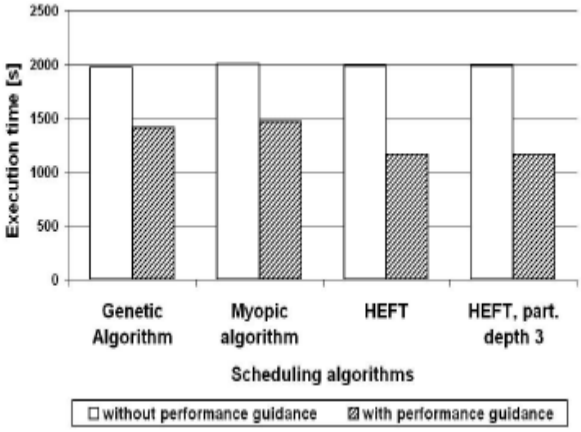

Fig 8.4: WIEN2k executed in heterogeneous environment, execution time.

From the figure 8.4, the HEFT algorithm gives the better result than the Genetic algorithm. In other word the execution time of HEFT algorithm is less than the Genetic algorithm. Execution time of the workflow is 17% shorter than for the genetic algorithm.

9. Conclusion

Cloud computing provide the services as hardware infrastructure and software applications. We saw the DAG model and definitions for workflow scheduling. We focused on study of Workflow of scientific applications in cloud environment. We have studied various paper related to workflow scheduling. In this study we consider the three algorithms that include PSO, GA, and HEFT to minimize the execution time of applications. From [4], we conclude that PSO algorithm minimize the total execution cost of workflows of an application in Cloud. From [7], HEFT algorithm performs better than the Genetic algorithm. It minimum total execution time and execution cost of workflow application in cloud environment.

References

[1] Fuhui Wu QingboWu,Yusong Tan “Workflow scheduling in cloud: a survey”, © Springer Science+Business Media New York 2015

[2] Chiaki Sato,Luke M. Leslie,YoungChoon Lee "Running Data-Intensive Scientific Workflows in the Cloud", International Conference on Parallel and Distributed Computing, Applications and Technologies

[3] Gideon Juve and EwaDeelman, “ScientificWorkflows in the Cloud”, University of Southern California, Marina del Rey, CA e-mail: juve@usc.edu

[4] L. Wu, S. Panday, R. Buyya, “A Particle Swarm Optimization-Based Heuristic for Scheduling

Workflow Applications in Cloud Computing Environments", proceedings of the IEEE International Conference on Advanced Information Networking and Applications.

[5] "A Survey on Deadline Constrained workflow scheduling in cloud environment", Nallakumar. R1, SruthiPriya. K. S2 .

[6] X. Liu et al., The Design of Cloud Workflow Systems,SpringerBriefs in Computer Science,"Cloud Workflow System Architecture" .

[7] MarekWieczorek, RaduProdan and Thomas Fahringer, "Scheduling of Scientific

[8] Gideon Juve and EwaDeelman, "ScientificWorkflows in the Cloud".

[9] Arabnejad H, Barbosa JG (2014) A budget constrained scheduling algorithm for workflow applications.J Grid Comput, pp 1–15.

[10] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Sci. Program.*, 13(3):219–237, 2005.

[11] A. Salman. Particle swarm optimization for task assignmentproblem. *Microprocessors and Microsystems*, 26(8):363–371, November 2002.

[12] W.N. Chen, J. Zhang and Yang. Y,"Workflow Scheduling in Grids: An Ant Colony Optimization Approach." Evolutionary Computation. IEEE Conference, Pg. 3308-3315, September 2007.

[13] Gencyilmaz. A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem. European Journal of Operational Research, 177(3):1930–1947, March 2007.