A HYBRID COST EFFECTIVE ALGORITHM FOR WORKFLOW SCHEDULING IN CLOUD COMPUTING

Ishadeep Kaur[1], P S Mann[2]

Research Scholar, Department of Computer Science and Engineering, DAVIET, Jalandhar, India[1]

Assistant Professor, Department of Information Technology, DAVIET, Jalandhar, India[2]

Abstract : Cloud Computing is emerging as a new paradigm that provides high quality and low-cost information services by pay-per-use model, but still it has several issues that need to be focused to execute scientific workflows. Workflow scheduling is one of the critical issue in cloud computing which aims to find a feasible schedule that minimizes the finish time and execution cost while meeting the quality of service requirements .Efficient scheduling can have significant impact on the performance of the system. In this paper, we propose a hybrid cost effective Genetic and Firefly Algorithm for workflow scheduling in cloud computing. In existing approach the number of iterations are very large which increases the Total Execution Cost and Total Execution Time which we will optimize in proposed algorithm. The initial population is generated by the PEFT algorithm which makes the use of OCT table is fed to hybrid algorithm of Firefly algorithm and Genetic Algorithm. The best optimized schedule is produced by the Hybrid of Genetic and Firefly Algorithm. The performance evaluation on well-known scientific workflows such as Montage, LIGO and Epigenomics of different size shows that our proposed hybrid algorithm performs better than the existing algorithm.

IndexTerms **- Cloud Computing, scientific workflows, Workflow scheduling, quality of service (QoS)**

I. Introduction

Workflow is the set of interconnected activities or steps which need to be completed in a sequence for the successful execution of the application. A workflow is represented in the form of Directed Acyclic Graph(DAG), with a tuple G = (T, E), where T depicts the number of tasks (vertices) and E are the edges that shows the dependencies among the tasks. Highly complex data intensive scientific and business applications are represented with the help of workflows these days. Workflows have made our computation easy by decomposing them into the smaller indivisible units called tasks that are executed in a defined sequence to get the better results. Workflow scheduling in cloud can be defined as to map the various tasks to the available best VMs according to their functional and non-functional requirements. Due to the large number of tasks and available resources, workflow scheduling is a NP-Hard problem as there is no algorithm which could provide the optimal solution in polynomial time. Therefore, a number of heuristics and meta-heuristics algorithms have been presented in past to achieve a near optimal solution in finite time. Cloud Computing has several advantages but still it has several challenges that needs to be addressed like performance of virtual machines, acquisition delay, termination delay and VM failure. So there is a need to develop a Cost Effective algorithm that will overcome these drawbacks. Our research work is based on Hybrid Cost Effective Genetic Algorithm that will minimize the execution cost, finish time and termination delay with high degree of security and provides QoS. To achieve this a Hybrid of Firefly and Genetic algorithm is proposed that minimize the total execution cost while meeting the deadline constraints. The initial population for the Firefly algorithm is generated using the PEFT algorithm that minimizes the number of iterations. The rest of the paper is organised as follows: Section II shows the related work. Section III describes the Genetic Algorithm. The proposed algorithm is discussed in section IV. Section V describes the simulation Results and Discussions and section concludes the paper.

II. Related work

Workflow scheduling is one of the key issue in the management of workflow execution. The Workflow scheduling on cloud and grid are studied over the years and it is a NP-Hard problem. Therefore, it is very difficult to find an optimal solution within a polynomial time. Initially the workflows are implemented in Grid but due to reduced performance now workflows are implemented on cloud. Cloud provides various services like scalability, QOS on demand resource provisioning. However, various heuristic and meta-heuristic algorithms have been proposed for multitask workflow scheduling.

Authors of **[1]** proposed a genetic algorithm approach for scheduling workflow applications by minimizing the cost while meeting user's deadline constraint or minimizing the execution time while meeting user's budget. The proposed algorithm evaluates fitness function into two parts cost fitness and time fitness. It solves the budget and deadline constrained optimization problems. The result shows that the genetic algorithm is better for handling complex workflows structure. Authors of **[3]**proposed a new heuristic algorithm for task scheduling which embeds a new fast technique named Elitism Stepping into Genetic Algorithm with the objective to reduce the schedule length within an acceptable computational time. The author compared the proposed algorithm with BGA and obtains better schedule length or finish time. The result shows significant improvement in computation time of the new algorithm. Authors of **[6]** surveyed the various existing workflow scheduling algorithm in cloud computing and tabulated their various parameters along with tools. The author concluded that Existing workflow scheduling algorithms does not consider reliability and availability. So there is a need to implement a workflow scheduling algorithm that can improve the availability and reliability in cloud environment. Authors of **[13]** proposed a new method for initial population so that the algorithm gives the results faster. HSGA is a hybrid algorithm for heterogeneous distributed system like cloud computing which works on genetic algorithm by combing the advantages of two heuristic techniques. It combines the Best Fit and Round Robin method for assigning the resources to the tasks. The aim is to decrease the completion time and the failure rate of the application. Authors of **[10]** proposed a new Modified Genetic Algorithm for scheduling the tasks in private cloud for minimizing the make span and cost. In MGA, initial population is generating using SCFP(Smallest cloudlet to Fastest Processor), LCFP(Longest cloudlet to Fastest Processor) and 8 random schedules. Two point crossover and simple swap are used. This gives the good performance under the heavy loads. Authors of **[14]** has characterized workflows from six diverse scientific applications including astronomy, bioinformatics, earthquake science and gravitational wave physics, characterizing their runtimes, I/O requirements, memory usage and CPU utilization based on tasks. The author found that same data is re-read multiple times in some workflows. Authors of **[18]** proposed a MPQGA which a generates various priority queues using a heuristic based cross over and heuristic based mutation operator in order to minimize the makespan. It uses the advantages of HEFT heuristic algorithm to find a better result in which the highest priority task calculated by the upward rank is mapped on to the processor which gives the less EFT. This algorithm covers a large search space than the deterministic algorithm without much cost. Authors of **[19]** presented a scheduling technique based on relatively new swarm based approach known as Cat Swarm Optimization. This technique shows improvement over PSO in terms of speed of Convergence. By using the Seeking mode and Tracing mode ,the algorithm reduces the wastage of energy and obtain solution in much lesser number of iterations. The author has targeted at minimization of total cost, minimum number of iterations, fair distribution of workload. The authors had proved that CSO give better results than PSO in terms of exaction time and computation time. Authors of **[15]** proposed a novel hybrid algorithm named ACO–FA, which integrate ant colony optimization (ACO) with firefly algorithm (FA) to solve unconstrained optimization problems. The proposed algorithm integrates the merits of both ACO and FA, where the algorithm is initialized by a set of random ants that is roaming through the search space. The proposed algorithm to handle complex problems of realistic dimensions has been approved due to procedure simplicity. It can efficiently overcome the drawback of classical ant colony algorithm which is not suitable for continuous optimizations. Authors of **[20]** proposed a new list based scheduling technique named PEFT for heterogeneous distributed computing which gives the better results than HEFT in terms of makespan. It has the same time complexity as that of HEFT. It consists of task prioritizing phase and processor selection phase. It makes the use of an optimistic cost table (OCT) which is the basis for finding the task priority and selecting a resource for it. OCT indicates the minimum time required for processing all the tasks which lies on the longest path from the current task to the end task. In task prioritization, task priority is calculated by cumulative OCT. Optimistic EFT is calculated to assign a processor for a task. The aim is to ensure that the children tasks of the current task are finished earlier. Authors of **[25]** presented a hybrid approach which combines the positive benefits of heuristic algorithm and a meta-heuristics algorithm by modifying its genetic operators. It uses the HEFT heuristic which is better than the other list based heuristic in terms of its robust nature and make span for initial seed. Elitism helps in maintaining the quality by copying the best chromosomes from one iteration to next iteration. The two fold genetic operators namely crossover and mutation are used which helped in optimizing the fundamental objective (to minimize makespan) in less amount of time. It also optimized the load balancing during the execution. Authors of **[26]** proposed a meta heuristic cost effective genetic algorithm that minimizes the execution cost of the workflow while meeting the deadline in cloud computing environment. Its main goal is to develop a novel scheme for encoding, population initialization, crossover and mutation operators of genetic algorithm. The simulation experiments are conducted on four scientific workflows such as Montage, LIGO, Cybershake and Epigenomics. It shows that the algorithm CEGA has lower execution time and cost than IC-PCP,RTC,RCT and PSO. Authors of **[21]** presented Deadline Constrained Heuristic based Genetic Algorithm for scheduling applications on cloud that minimize the execution cost while meeting the deadline. Each task is assigned priority using bottom-level and top-level. The algorithm is compared with SGA under same deadline constraint and pricing model. The simulation results show that the proposed algorithm have a promising performance as compared to SGA.

III. Cost Effective Genetic Algorithm

GA is the meta-heuristic approach which takes the inspiration from the evolution techniques of nature like mutation, crossover, selection and fitness evaluation to achieve a near optimal solution. It is based on the survival of the fittest principle which is applied to the consecutive generations in order to find the best solution among the large number of feasible solutions. The chromosome which consists of genes represents the possible feasible solution. On the randomly selected initial population (chromosome), fitness function is applied which determines the most powerful chromosomes for the next generation. Fitness function is based on the optimizing criteria which may be reducing the makespan, cost or increasing the reliability.

The Cost Effective Genetic Algorithm includes a novel scheme for encoding ,population initialization, crossover and mutation. Firstly a Task Order method for assigning an order to tasks is created. Initially ordered list is empty, a task is selected an put at the end of the ordered list and then the predecessors of that task are also added. The process continuous until all the tasks are added into the ordered list. After completion of this method it generates 'n' number of ordered lists. Then it used Cheapest task VM-map method to map each task of the Ordered List to cheapest virtual machine. So, the initial population is generated by this method Then Fitness Evaluation method is used to schedule each task of the workflow. In this the execution time and transfer time matrix of all the tasks are computed. Then the Start time and finish time of each task is estimated. Then the LST and LET of the VM leased from the cloud to execute tasks are estimated. So at the end the complete information to define a schedule S is generated. After the fitness evaluation two point Crossover is used to generate a new child schedule from the existing parent schedules. The crossover rate used is 100 per cent. Initially a random value is generated between 0 and 1.Then this random value is compared with crossover rate. If it is less then crossover rate then crossover is performed otherwise it ends. At the end Mutation is performed using topological level of each task to maintain the dependency constraint of the workflow. For mutation swapping is performed. If new schedule is better in terms of minimization of execution cost while meeting the deadline constraint then this schedule is kept otherwise it is discarded.

IV. Proposed Hybrid Cost Effective Genetic Firefly Algorithm (CEGF) for Workflow Scheduling

A Hybrid Cost Effective Genetic and Firefly Algorithm for Workflow scheduling is proposed. In this Algorithm firstly the initial population is generated using PEFT algorithm. PEFT algorithm makes the use of an optimistic cost table (OCT) which is the basis for finding the task priority and selecting a resource for it. This initial population generated by PEFT algorithm is fed to the Firefly algorithm. Population of firefly is initialized using the prioritize solution of PEFT algorithm. Then the fitness of all the fireflies using the objective function i.e. execution time is evaluated. The light intensity of the fireflies is updated. Then the fireflies are ranked and then the position of fireflies are updated. Optimized solution of firefly algorithm is then fed to Genetic algorithm. Before applying the Genetic Algorithm, the chromosomes for applying genetic operations are selected by using the binary tournament selection technique. On the selected chromosomes the crossover and mutation genetic operators are applied to produce the new children. Then the new children are validated by the fitness function and the good quality off-springs are added into the new population. At the last the best optimized schedule is produced by the Genetic Algorithm. The performance parameters including execution time, execution cost and termination delay are evaluated. The various steps of this algorithm are discussed as below.

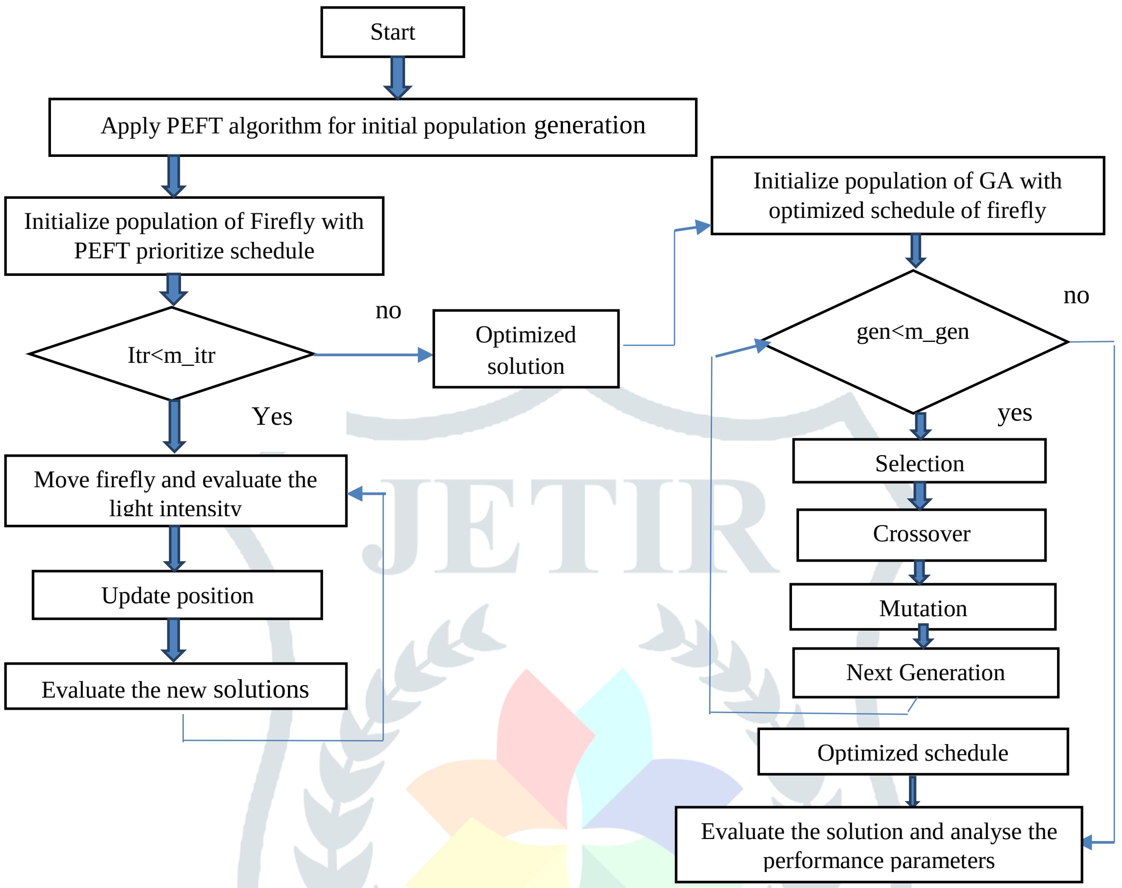

Figure 1. Flowchart of Proposed work scheme

In the Figure 1, flowchart of proposed approach is given. First step is to generate a good seed using the PEFT algorithm. Compute the OCT table. The OCT is in the form of a matrix whose rows are the tasks and columns are the VMs. The OCT value is calculated according to the equation below recursively using the backward approach. It will give the cost to execute all the children tasks of a current task until it reaches the last point.

$$O(t_i, p_k) = \max_{t_j \in succ(t_i)} [\min_{p_w \in P} \{OCT(t_j, p_w) + w(t_j, p_w) + \overline{c_{i,j}}\}] \quad (1)$$

$$\text{where } \overline{c_{i,j}} = 0 \text{ if } p_w = p_k$$

Find cumulative OCT of every node and cumulative OCT defines the rank of every node or task($rank_{oct}$) as in Eq 3. Tasks are ordered in the list on the basis of decreasing order of the $rank_{OCT}$.

$$rank_{oct}(t_i) = \Sigma_{k=1} OCT(t_i, p_j) / P \quad (2)$$

If the list is not empty, else return the best schedule in terms of makespan. Optimistic EFT is calculated according to the below equation to allocate a task for the processor using the insertion-based policy. Optimistic EFT is given by:

$$O_{EFT}(t_i, p_j) = EFT(t_i, p_j) + OCT(t_i, p_j) \quad (3)$$

Task is assigned to the processor which will give the minimum $O_{EFT.}$ A high-quality schedule thus generated is seeded it into the Firefly algorithm. Population of firefly is initialized using the prioritize solution of PEFT algorithm. If the termination condition is met, return the optimal solution else repeat steps second to last till the results. Next step is to evaluate the fitness of all the fireflies using the objective function i.e. (execution time). Then updates the light intensity of the fireflies i.e. (update the fitness of solutions)and final step is to rank the fireflies and update the positions till the optimized solution will be generated. Optimized solution of firefly algorithm is then fed to genetic algorithm. Where the chromosomes (individuals among the population) are encoded using direct representation. The fitness function will ensure that the solution has the minimum cost and is completed within the deadline. For this, select the chromosomes for applying genetic operations by using the binary tournament selection technique. On the selected chromosomes apply the crossover and mutation genetic operators to produce the new children (generation). Validate the new children by the fitness function and add the good quality off-springs (valid) into the new population. Best optimized solution obtained from the Genetic algorithm is the best solution and evaluate the performance parameters including execution time, execution cost and termination delay.

V. Result and Discussion

In this section proposed algorithm is compared with existing algorithm. The comparison is based upon two parameters Finish time and Execution cost. The results are evaluated on scientific workflows such as Montage, LIGO and Epigenomics. Workflows of different size are taken.

5.1 Finish Time

Finish time is the total execution time of task ti on the virtual machine of type VM that has minimum execution time among all types of VMs available in cloud and its finish time is defined as follows:-

Finish Time : $FT(t_i)=ST(t_i)+\{ET_{t_i}^{VM_k}/(1\text{-PerVar})\}$

PerVar is the amount of variation in the performance of VM of type VM_k

The below table shows the Finish Time of Hybrid GAFFA and GA. Different datasets of different size are taken to conduct an experiment. The scientific workflow MONTAGE, INSPIRAL and EPIGENOMICS are taken of different size.

Datasets	GAFFA (Proposed)/ms	GA (Existing)/ms
Dataset1[MONTAGE -25]	18.52	22.95
Dataset2 [MONTAGE-50]	39.05	44.85
Dataset3[INSPIRAL,25]	525.89	623.92
Dataset4[INSPIRAL,50]	455.39	805.03
Dataset5[EPIGENOMICS,24]	1659.4.	2179.98
Dataset6 [EPIGENOMICS,46]	3542.62	4133.38

Table 1.Finish Time matrix of GAFFA and GA

We have evaluated the performance of the algorithm with the help of a bar graph.It shows that proposed Hybrid GAFFA includes less Execution cost in comparision with GA..The different datasets shows an improvement in proposed algorithm.The GA has longest excution cost for each type of scientific workflow.The Proposed algorithm GAFFA gives the cost effective schedule for each workfllow and it has much lower cost then GA.

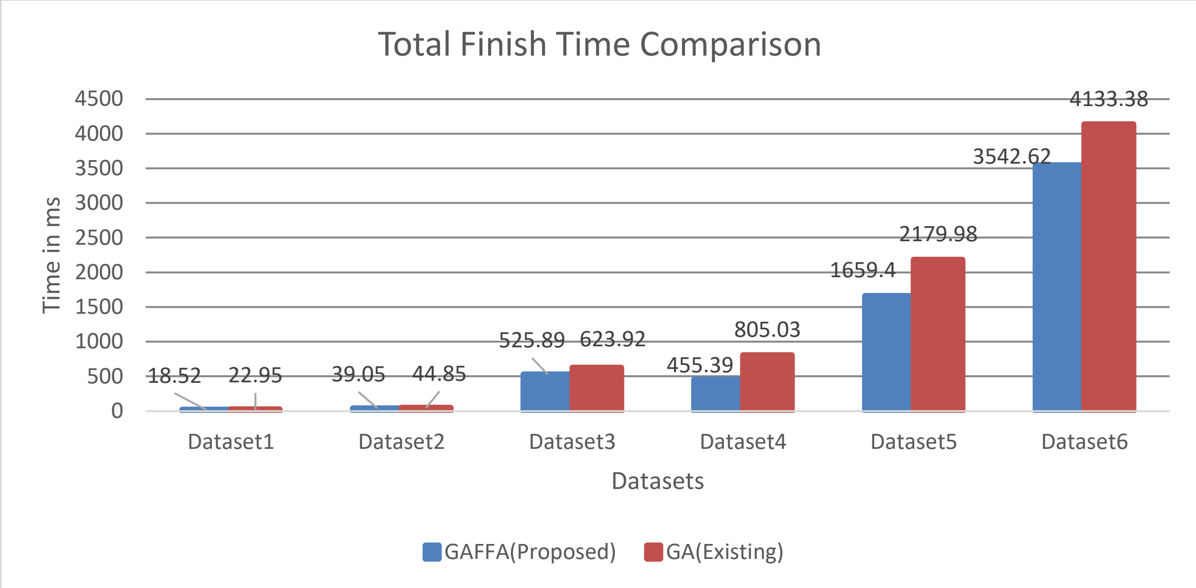

Figure 2. Total Finish Time Comparisons of GAFFA and GA

5.2 Execution cost

The Execution Cost is to find a feasible schedule(S) for a given workflow such that Total Execution Cost does not exceed to the deadline(D) of the workflow. The Execution Cost can be computed as follow:-

$$\textbf{Execution Cost}= \sum_{v=1}^{VM\_Pool} Ctype(vm_{k})*[\frac{LET_k - LST_k}{timeinterval}]$$

$Ctype(vm_k)$ is the cost of type vm_k per time interval

The below table shows the execution cost of Hybrid GAFFA and GA. Different datasets of different size are taken to conduct an experiment. The scientific workflow MONTAGE, INSPIRAL and EPEGENOMICS are taken of different size.

Datasets	GAFFA (Proposed)	GA (Existing)
Dataset1[MONTAGE -25]	5670.24	12043.33
Dataset2 [MONTAGE-50]	5356.66	44110.23
Dataset3[INSPIRAL,25]	6810.21	245310.0
Dataset4[INSPIRAL,50]	7923.43	614240.0
Dataset5[EPIGENOMICS,24]	6340.02	565220.0
Dataset6 [EPIGENOMICS,46]	15800.0	198518.0

Table 2. Execution Cost matrix of GAFFA and GA

We have evaluated the performance of the algorithm with the help of a bar graph.It shows that proposed Hybrid GAFFA includes less Execution cost in comparision with GA..The different datasets shows an improvement in proposed algorithm.The GA has longest excution cost for each type of scientific workflow.TheProposed algorithm GAFFA gives the cost effective schedule for each workfllow and it has much lower cost then the GA.

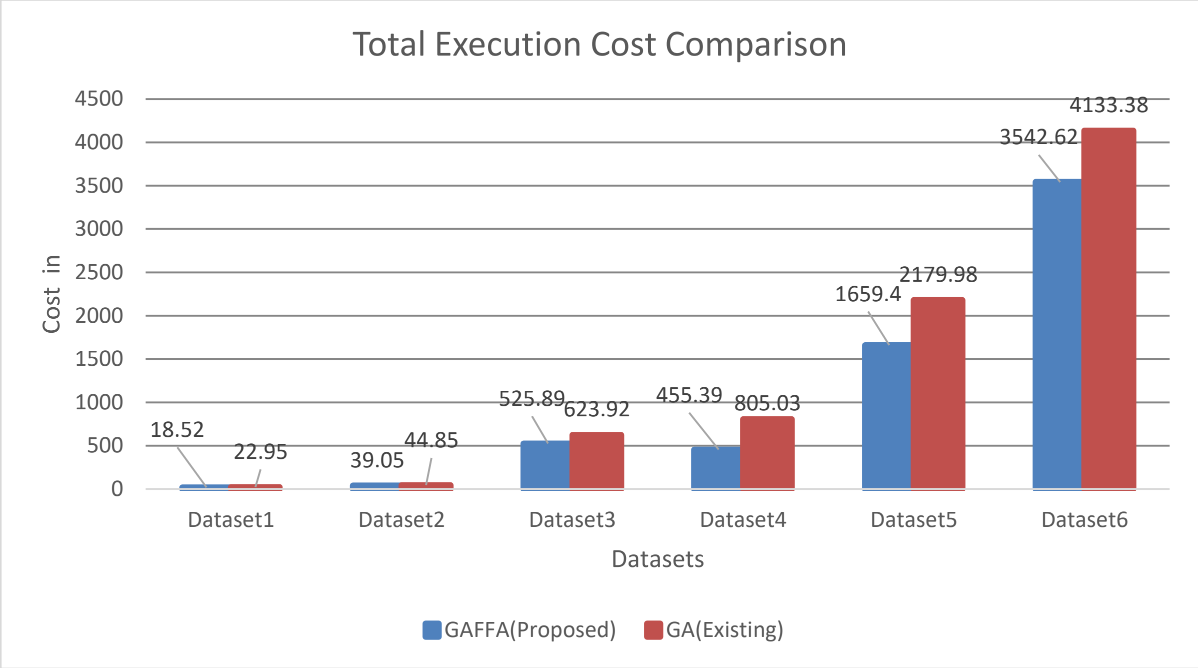

Figure 3. Total Execution Cost comparison of GAFFA and GA

VI. Conclusion and Future Scope

As cloud computing is an emerging field and workflow scheduling is one of the focused area in it. Workflow scheduling is one of the critical issue in cloud computing which aims to find a feasible schedule that minimizes the finish time and execution cost while meeting the quality of service requirements. The cloud system works on different algorithms but there is need to develop a novel scheme that will improve the quality of the algorithm. The proposed algorithm CEGF exhibits that it performs better than the other algorithms in terms of execution cost and finish time. The performance evaluation on well-known scientific workflows such as Montage, LIGO and Epigenomics of different size shows that our proposed hybrid algorithm performs better than the existing algorithm. In the future work we would like to integrate Firefly Algorithm with some other meta heuristic algorithm to improve the quality of service.

References

[1] Yu, J., Buyya, R.: "*Scheduling Scientific Workflow Applications with Deadline and Budget Constraints using Genetic Algorithms. Scientific Programming*" 14,217–230 2006

[2] E. Ilavarsan and P. Thambidurai. "*Low complexity performance effective task scheduling algorithm for heterogeneous computing environments*" Journal of Computer Science, 3(2):94–103, 2007.

[3] Amir MasoudRahman and Mohammad Ali Vahedi, "*A Noval Task Scheduling in Multiprocessors System with Genetic Algorithm by Using Elistism Stepping Method*", Science and Research Branch,Tehran, Iran,May 26,2008.

[4] J. Yu, R. Buyya and K. Ramamohanarao, "Workflow scheduling algorithms for grid computing", Metaheuristics for scheduling in distributed computing environments, SpringerBerlinHeidelberg,pp.173-214,2008

[5] Talbi and E. Ghazali, "Metaheuristics: from design to implementation", John Wiley & Sons. , vol. 74, 2009.

[6] AnjuBala, Dr.InderveerChana, "*A Survey of Various Workflow Scheduling Algorithms in Cloud Environment*", 2nd National Conference on Information and Communication Technology NCICT ,Proceedings published in International Journal of Computer Applications (IJCA) 2011

[7] M. I. Daoud and N. N. Kharma, "A hybrid heuristic-genetic algorithm for task scheduling in heterogeneous processor networks", Journal of Parallel and Distributed Computing, Elsevier, vol. 71, no. 11, pp. 1518–1531, 2011 .

[8] A. Yousif, A. H. Abdullah, S. M.Nor, A. A. abdelaziz, Scheduling jobs on grid computing using firefly algorithm. Theoretical and Applied Information Technology,33,155-164 2011.

[9]ArashGhorbanniaDelavar, Yalda Aryan, "*HSGA: a hybrid heuristic algorithm for workflow scheduling in cloud systems*", Journal of Cluster Computing, Springer, 2012.

[10] S. Kaur and A. Verma, "An efficient approach to genetic algorithm for task scheduling in cloud computing environment", International Journal of Information Technology and Computer Sciencevol.4,no.10,pp.74,September,2012

[11] C. Wang, J. Gu, Y. Wang, and T. Zhao, "A hybrid heuristic-genetic algorithm for task scheduling in heterogeneous multi-core system", in Algorithms and Architectures for Parallel Processing (Y. Xiang, I. Stojmenovic, B. Apduhan, G. Wang, K. Nakano, and A. Zomaya, eds.), Springer Berlin Heidelberg ,vol. 7439 of Lecture Notes in Computer Science, pp. 153–170, 2012.

[12] X. S. Yang, Multiobjective firefly algorithm for continuous optimization,Engineeringwith Computers, Online Fist, DOI: 10.1007/s00366-012-0254-1 2012.

[13] M.-H. Horng, Y.-X.Lee, M.-C.Lee and R.-J.Liou, Fireflymetaheuristic algorithm fortraining the radial basis function network for data classification and disease diagnosis,in:Theory and New Applications of Swarm Intelligence pp. 115-132 2012

[14] Gideon Juve, Ann Chervenak, EwaDeelman, ShishirBharathi, Gaurang Mehta, Karan Vahi "Characterizing and Profiling Scientific Workflows". 2013

[15] R.M. Rizk-Allah, Elsayed M. Zaki, Ahmed Ahmed El-Sawy" Hybridizing ant colony optimization with firefly algorithm for unconstrained optimization problems" Applied Mathematics and ComputationVolume 224, 1, Pages 473-483,2013

[16] S. Abrishami, M. Naghibzadeh, and D. H. J. Epema, ''Deadlineconstrained workflow scheduling algorithms for infrastructure as a service clouds,'' Future Generat.Comput. Syst., vol.29,no.1,pp.158–169,2013.

[17] Xin-She Yang and Xing-shi He,(2013). 'Firefly Algorithm: Recent Advances and Applications',Int. J. Swarm Intelligence,Vol. 1,No. 1, pp. 36–50,2013 .

[18] Y.Xu, K. Li, J. Hu, and K. Li, "*A genetic algorithm for task scheduling on heterogeneous computing systems using multiple priority queues*," *Information Sciences*,vol 270, no.0,pp. 255-287, 2014.

[19] SaurabhBilgaiyan, SantwanaSagnika and Madhabananda Das, "*Workflow Scheduling in Cloud Computing Environment Using Cat Swarm Optimization*", International Journal of Computer Applications, Vol .89, No.2, pp. 680 – 685,2014

[20] Hamid, Arabnejad and JG. Barbosa, "*List scheduling algorithm for heterogeneous systems by an optimistic cost table"*, IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 3, pp. 682-94, 2014.

[21] A. Verma and S. Kaushal, "Cost-Time Efficient Scheduling Plan for Executing Workflows in the Cloud", Journal of Grid Computing, Springer, vol. 13, no. 4, pp. 495-506, 2015.

[22] Delavar, A. Ghorbannia and Y. Aryan, "HSGA: a hybrid heuristic algorithm for workflow scheduling in cloud systems", Cluster computing, Springer, vol. 17, no. 1, pp. 129-137, 2014.

[23] M. Kalra and S. Singh, "A review of metaheuristic scheduling techniques in cloud computing", Egyptian Informatics Journal, Elsevier, vol. 16, no. 3, pp. 275-295, 2015.

[24] S. Smanchat and K. Viriyapant, ‘‘Taxonomies of workflow scheduling problem and techniques in the cloud,’’ Future Generat.Comput. Syst., vol. 52, pp. 1–12, Nov. 2015.

[25] S.G. Ahmad, C.S. Liew, E.U. Munir, T.F. Ang and S.U. Khan, "*A hybrid genetic algorithm for optimization of scheduling workflow applications in heterogeneous computing systems*", Journal of Parallel and Distributed Computing, Elsevier, vol. 87, pp. 80-90,2016.

[26] JasrajMeena, Malay Kumar, And Manu Vardhan, "*Cost Effective Genetic Algorithm for Workflow Scheduling in Cloud Under Deadline Constraint*", IEEE Access, Vol.4, pp. 5065-5082, 2016.

[27] A. Verma, S. Kaushal, "*Deadline constraint heuristic based genetic algorithm for workflow scheduling in cloud,*" Forthcoming article in international journal of grid and utility computing,2016

[28] G. Wang, Y. Wang, H. Liu and H. Guo, "HSIP: A Novel Task Scheduling Algorithm for Heterogeneous Computing", Scientific Programming, Hindawi Publishing Corporation, 2016.