

Resource Provisioning for Scientific Workflow Applications using Aws Cloud

S.M.Jaybhaye, V.Z.Attar



Abstract: Cloud computing play a very important role in day to day life of everyone. In recent years cloud services are much popular for hosting the applications. Virtual Machine Instances are the Images of physical machines which are described with its specification and configurations such as number of microprocessor (CPU) cycles, Memory access and network bandwidth. Cloud provider must contribute special interest while designing and implementing the IaaS. The role of quality and service performance is crucial aspects in application execution Scientific workflow based applications are both compute and data intensive. These application can take advantage of cloud features. Resource provisioning approaches varies based on the user's requirements and the metric which is used to allocated resources.
Index Terms: Cloud computing, Resource provisioning, Scheduling

I. INTRODUCTION

Cloud computing is on demand provisioning of different resources for computing, storage and communication. Different services of cloud provides speedy access to flexible and low-cost IT resources. Cloud user can access different types of resources as per their requirement, almost instantly, and only pay for what you use. Cloud provider preserves its own network connected hardware resource pool and applications, which are provided to the end user through web services. Cost calculations and projections of a workflow based applications majorly depends on type of Virtual Machine Instance (VMI), which will have different combinations of CPU, storage and network, number of virtual machines, cost (charges) of VMI, pricing model, monitoring services. To enhance the workload demands of small startups, businesses and organizations cloud services are used. They provide low cost, high availability, reliable and resizable and scalable resources. Scientific workflow is becoming more popular model for scientists. It helps to formalize and structure complex scientific processes. It speeds up many significant scientific discoveries. Specification, modification, execution, failure handling and monitoring of scientific workflows can be done on cloud.

Scientific Workflow Scheduling and mapping in cloud computing is the problem which has been in research for long time. It is an optimization problem where the tasks need to be scheduled in such a way that it minimizes the cost of execution and time with required quality of service parameter.

The very high importance is given by researchers & cloud service providers for optimization due to the fact that the optimal scheduling of virtual resources could save the millions of dollars every year. It is very crucial to select the right task for right resource and schedule it. Due to changing demands of the quality of services required by the user, many a times a lot of expensive resources are wasted and the execution time is also increased. Scientist often need to

- 1) Integrate diverse services and data
- 2) Automate data processing steps
- 3) Repeat processing steps on new data
- 4) Share their analysis steps with other researchers
- 5) Track the provenance of data analysis
- 6) Execute analyses in parallel on distributed resources

II. COST MANAGEMENT FOR WORKFLOW BASED APPLICATION

Cloud computing is remotely accessible service which provides cloud based resources to its consumer. Cloud based resources can be physical or virtual (software based). Physical server, storage device, network device are physical resources and virtual server, virtual storage and virtual networks are

software based virtual resources. Infrastructure as a Services (IaaS) uses virtualization technology for resource provisioning. It is a web service which provides many computing resources called virtual instances. To control the cost of cloud services best suitable model should be selected and the best practices should be selected Better resource selection and mapping of right sized resource to the task application, Resource utilization, workload management, budget review, monitoring plays important role in resource provisioning. Cost optimization approaches can be improved by these metrics. They also facilitates variety of pricing models such as On-Demand, Spot Instances, Dedicated Instances and Reservation.

1. On-Demand Instances— computing power can be minimized or can increased or reduced as per user's requirement change of the applications. Purchasing and maintenance and security issues will be handled by provider only.
2. Spot Instances— bidding is done for available and extra resources.
3. Dedicated Instances : More secure and more controlled, physically isolated resources are provided

Revised Manuscript Received on February 10, 2020.

* Correspondence Author

S.M.Jaybhaye*, PhD Scholar, Department of Computer Engineering and Information Technology, College of Engineering Pune, Pune, India.

Dr. V.Z.Attar, Associate Professor & Head, Department of Computer Engineering and Information Technology, College of Engineering Pune, Pune, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

4. Reserved Instances—substantial concessions are provided but planning is required. Scientific Workflow Scheduling and mapping in cloud computing is the problem which has been in research for long time.

It is an optimization problem where the tasks need to be scheduled in such a way that it minimizes the cost of execution and time with required quality of service parameter.

The very high importance is given by researchers & cloud service providers for optimization due to the fact that the optimal scheduling of virtual resources could save the millions of dollars every year. It is very crucial to select the right task for right resource and schedule it. Due to changing demands of the quality of services required by the user, many a times a lot of expensive resources are wasted and the execution time is also increased. Thus there is a need to achieve close match between the plan and reality. Mixed Integer non Linear Programming includes a mathematical model in which some variables are integer and some real, this type of a problem turns into a NP-hard problem. An NP-hard problem is generally categorized with having a higher mathematical and computational complexity, this makes handling solution more difficult.

The NP-hard complexity arises due to the branch and bound technique in finding the solution of integer variables. One of the constraints in the mathematical model thus generally exhibits nonlinearity. This makes the problem difficult to solve as the classic simplex algorithm cannot work here. One of the way of solving these problems is by decomposing the problem in linear and non-linear constraints and another by solving the problem using Evolutionary algorithm.

III. REAL-WORLD WORKFLOW APPLICATIONS

Scientific applications are dependent on other task as well as complex for execution. These applications may require a number of analysis tools for data processing. These applications are time constrained applications and may require a large supercomputing support as well as different types of VMI. Scientific workflows are used as research purpose i.e. many changes are done in dataset and tasks and it may require some human interventions. After research human interventions can be reduced and automation can be increased. Following are some of the well-known applications which are used for experimentation and analysis.

Montage workflow: It is an astronomy application. Montage workflow enables an astronomical study for generating sky mosaics in the scale of hundreds of GBs.

Cybershake: It is an application to characterize earthquake hazards. CyberShake workflows are composed of more than 800,000 tasks and have input data larger than 200TBs.

LIGO: Laser Interferometer Gravitational-wave Observatory (Memory intensive): The Application is used for detecting gravitational waves.

SIPHT: It is used to automate the process of searching for sRNA encoding-genes for all bacterial replicons in the National Centre for Biotechnology Information database.

Epigenomics: It is a biological application that studies the set of epigenetic modifications on human cells. The workflow is a CPU intensive application that automates the execution of various genome sequencing operations.

Workflow representation: Fig. 1. shows various workflows

used for scientific Application. These are a) Montage b) Epigenetics c) LIGO d) CyberShake e) SIPHT



Fig 1. Scientific Workflow representation

IV. EVOLUTIONARY ALGORITHMS

Genetic algorithms are good for large, discontinuous problems and more chances to lead global solution. As compared to other classical techniques these algorithms don't use fix transition rules for the exploitation of the solution. Chromosome structure considers conditions like dependency between the different tasks. Tasks must be scheduled on the resources which are able to handle them. The solution after the application of Genetic algorithm, gives the pair of task and allocated resource. Chromosome structure considers the dependency condition between the different tasks i.e. the successor cannot be executed before the execution of its predecessor. The task with smallest ready time is executed first, if many tasks with same ready time then the task with earliest finish time will be executed initially. Task execution order is decided by some rule set based on users requirement. Tasks must be mapped on the resources which fulfill its requirement and handles them. The solution gives the pair of task and mapped resource. Genetic operators are the core of a genetic algorithm. They define how to generate new solutions from the existing ones. The two genetic operators are the crossover and the mutation operators. Some of the terms used for Adaptive Genetic algorithm are as

1. Gene: A Discrete variable
2. Chromosome: A set of genes
3. Population: Chromosomes present with same length forms the population.
4. Fitness: The value assigned to an individual
5. Fitness function: A function that assigns value to the individual gene.
6. Mutation: Small change in the gene
7. Selection: Select individuals for creating the next generation.

V. LITERATURE SURVEY

J. Durillo, Radu Prodan [1] developed Multi-Objective Heterogeneous Earliest Finish Time (MOHEFT) to find tradeoffs between time and execution cost using Amazon cloud for workflow executions.

It has been compared with single objective HEFT algorithm and multi objective evolutionary SPEA2 which shows better results for MOHEFT. G. Juve, Mats Rynge, Ewa Deelman [2] described their work on FutureGrids and Amazon EC2 cloud providers which reveals advantages and disadvantages of infrastructures usage for application deployment.

They hosted computationally intensive “periodogram” astronomy workflow application and considered different metric such as performance, usability and availability.

Availability of resources plays very vital role in scheduling if the workflows are deadline sensitive [22][27]. In this case effective resource procurement and task mapping helps.

A.Alsarhan, A.Itradat and A.Y.Zomaya[3] formulated resource provisioning with guarantee of Quality of Service (QoS) as specified in Service Level Agreement. Changes in service, cost and service demand are adapted and Virtual machine hiring is done using reinforcement learning.

Q.Zhu, Gagan Agrawal [4] evaluated their framework which supports dynamic adaptation using feedback control algorithm in consideration with time and cost to improve Quality of Service with less overhead.

G.Juve, Ewa Deelman, Karan Vahi, Gaurang Mehta .et.al [5] examined performance of Montage, Broadband, Epigenome scientific workflow applications. Resource Cost, Storage Cost, Transfer Cost analysis is done for workflow execution on EC2.

A.Tchernykh,Uwe Schwiegelsohn, et al. [6] reviewed role of cloud computing uncertainties while scheduling and resource provisioning .There are many sources of uncertainty such as data size, dynamic cost of virtual machines ,resource availability , scalability, job arrival rate etc. Most of the uncertainties depends on the following parameters of cloud providers such as resource capacity, bandwidth, processing time, memory, processing capacity, storage, and network capacity.

G.Juve, Ewa Deelman, Karan Vahi, *et al.* [7] described the problems of workflow deployment on EC2.Data management can significantly improve the performance of workflow execution.

J.Rao, Xiangping Bu, *et al.* [8] proposed generic framework named iBalloon which is an adaptive framework which provides resources. It uses reinforcement learning and uses individual Virtual Machines feedback.

C.Chen,Wei Wang, and Bo Li [9] used and find elasticity demand of data parallel jobs .Aware Fair (PAF) scheduler is used to get near optimal performance. Initially it fairly allocates resources and then transfer them from one to another. This approach improves performance by 13% without penalizing resource giver.

P.Armstrong and Ashok Agarwal [10] developed a new cloud scheduler based on Condor by considering user priorities and fair allocation, utilization of resources. Initially as per the jobs requirement Virtual machines are allocated to all the job which are in queue. Over allocated or under allocated virtual machines are balanced again for new job submission.

R.Garg, Awadhesh Kumar [11] proposed Adaptive Workflow Scheduler (AWS) based on the availability of resources in the grid .Resource load is of computing and communication is considered to check availability of resources.

Ali Nikravesh, Yadavar, Samuel A. Ajila et al. [12] proposed prediction approach to improve accuracy using TPC-W web application as workbench on Amazon cloud. Most of the researchers [13][19][20] worked on heterogeneous resource and cloud infrastructure and

workflow [14, 17] deployment by deploying them on Amazon EC2 [15].

Evolutionary approaches such as genetic algorithms [16][24][25] are able to solve multi objective scheduling problem. As the optimizing objectives increases search space and complexity is also increased [26]. Local exploitation and global exploration abilities are improved by adjusting parameters of evolutionary approaches By changing bidding strategies according to policies of market [23] in spot instance pricing model helps in cost reduction [18].

A detailed survey of scheduling algorithm [21] shows that For further study security issues must be considered for data management of workflow.

Now a days, researchers are trying to optimize the different parameters of cloud for workflow based applications by considering different metric such as resource cost, time , security, availability, deadline , priority, ranking etc

VI. EXPERIMENTAL SETUP

Following steps are implemented to deploy scientific applications on T2.micro Virtual machines of AWS cloud.

1. Select required Virtual Machines and launch the Instances.
2. select existing key pair before launching instances
3. Select the IP address of created Instance.
4. Use gitbash which is a linux like command prompt in windows
5. Change the IP address to connect new instances.
6. Install required software such as Java and Tomcat Server in the Virtual Machine using Sudo yum install java-1.8.0 command.
7. Open the browser and VM, provide Java and tomcat on the. VM go to browser and open VM with IP
8. Deploy application as projectname.war file on Virtual Machine.
9. List all the running files on that VM, using ls
10. Cp projectname.war

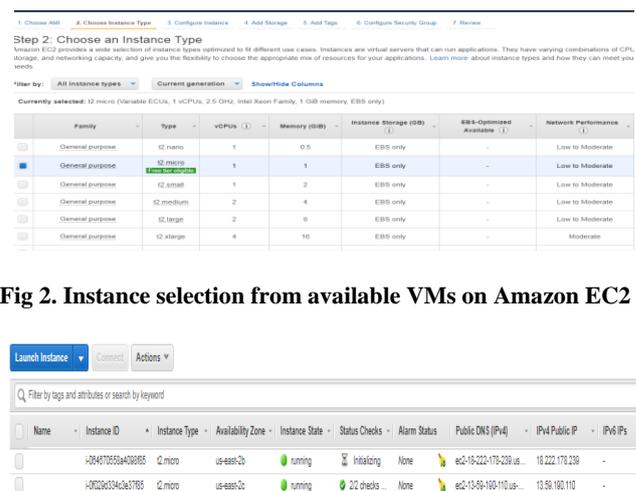


Fig 2. Instance selection from available VMs on Amazon EC2

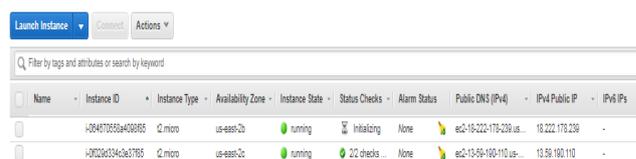


Fig 3. Launched and running instances

VII. RESULTS

Table 1,2,3 and 4 represents Execution time in seconds required for Inspiral ,Montage, SIPHT workflow respectively

Table 1: Execution time required for Inspiral workflow

No. of VM	Inspiral (Execution time)		
	Inspiral_30	Inspiral_50	Inspiral_100
2	6,268	16,821	60,984
3	4,529	11,694	42,449
4	3,639	9,094	31,964

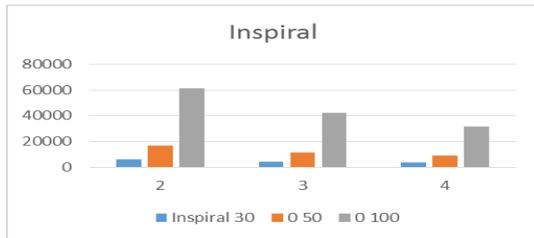


Fig. 4 Comparison of Inspiral with no. of VM

Table 2: Execution time required for Montage workflow

No. of VM	Montage(Execution time)		
	Montage_25	Montage_50	Montage_100
2	2,005	5762	32,734
3	1,530	4606	22,417
4	1,353	2799	17,192

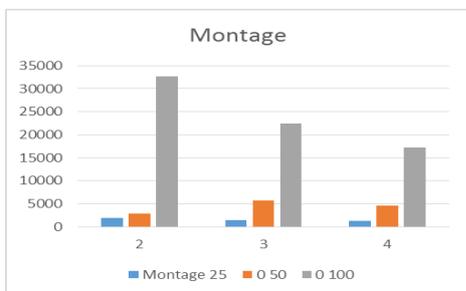


Fig. 5 Comparison of Montage with no. of VM

Table 3: Execution time required for SIPHT workflow

No. of VM	SIPHT(Execution time)		
	SIPHT_30	SIPHT_60	SIPHT_100
2	7,236	19,666	48,593
3	5,273	12,299	32,530
4	6,444	14,535	25,671

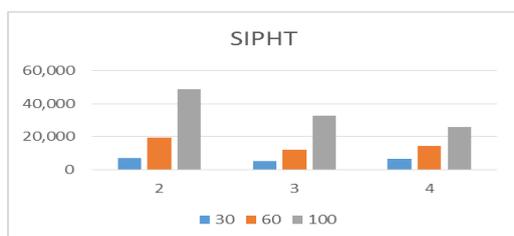


Fig. 6 Comparison of SIPHT with no. of VM

Table 4: Execution time required for Inspiral workflow

No. of VM	Cybershake ((Execution time)		
	Cybershake_30	Cybershake_50	Cybershake_100
2	85,372	188,450	706,173
3	59,458	130,818	492,010
4	45,784	97,110	364,004

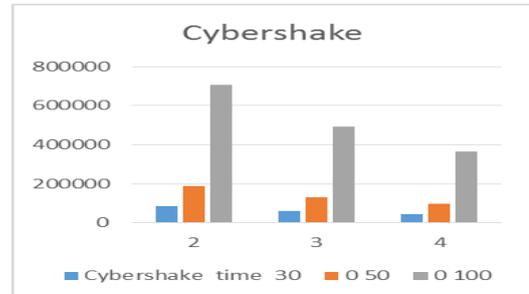


Fig. 7 Comparison of Cybershake with no. of VM

Figure 4,5,6, and 7 represents comparison of Inspiral ,Montage, SIPHT and Cybershake with no, of VM respectively.

For the purpose of research the most appropriate Quality of Service requirements are selected. The investigation includes a meticulous description of Workflow based applications and different existing approaches.

The contributions of this paper are:

- (1) Analysis of parameters used in scheduling of cloud computing.
- (2) Evolutionary approach is implemented and deployed on cloud for time and cost constrained scientific workflow.
- (3) Comparison of different workflows on variety of VM configurations.

VIII. CONCLUSION

Quality of Service (QoS) attributes impacts on the specified utility function by the user. Workflow scheduler is applied for different scientific workflows which are available in different sizes. The survey of workflow scheduling using evolutionary is conducted which shows that economy and execution time can be improved. Evolutionary approach and AWS Cloud is used to carry out the research work. It identifies a resource provisioning mechanism to map most suitable resource for the task. Selection strategies, crossover and mutation methods are applied to change the execution order. The proposed AGA makes balance of Quality of Service attributes without violating soft and hard constraints. As shown in fig. by varying the number of resources such as 2, 3 and 4 for Inspiral, Montage, SIPHT and Cybershake workflows, makespan is reduced. Reduced makespan ultimately results in less cost. When montage_25 workflow is executed on two, three, four VMs respectively makespan is reduced by 24 % and 33%.

The results shows that given genetic algorithm beats to other approaches in terms of CPU time and cost.



REFERENCES

1. J.Durillo, Radu Prodan, "Multi-objective workflow scheduling in Amazon EC2", Cluster computing 17, no. 2 (2014): 169-189.
2. G.Juve, Mats Rynge, Ewa Deelman, Jens-S. Vockler, and G. Bruce Berriman, "Comparing future grid, amazon ec2, and open science grid for scientific workflows", Computing in Science & Engineering 15, no. 4 (2013): 20-29.
3. A.Alsarhan, A.Itradat, A.Y. Zomaya,"Adaptive Resource Allocation and Provisioning in Multi-Service Cloud Environments." (2019).
4. Q.Zhu, Gagan Agrawal,"Resource provisioning with budget constraints for adaptive applications in cloud environments", IEEE Transactions on Services Computing 5, no. 4 (2012): 497-511.
5. G.Juve, Ewa Deelman, Karan Vahi, Gaurang Mehta, Bruce Berriman, Benjamin P. Berman, and Phil Maechling, "Scientific workflow applications on Amazon EC2", In 2009 5th IEEE international conference on e-science workshops, pp. 59-66. IEEE, 2009.
6. A.Tchernykh,Uwe Schwiegelsohn, Vassil Alexandrov, and El-ghazali Talbi,"Towards understanding uncertainty in cloud computing resource provisioning", Procedia Computer Science 51 (2015): 1772-1781
7. G.Juve, Ewa Deelman, Karan Vahi, Gaurang Mehta, Bruce Berriman, Benjamin P. Berman, and Phil Maechling, "Data sharing options for scientific workflows on amazon ec2", In Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1-9. IEEE Computer Society, 2010.
8. J.Rao, Xiangping Bu, Kun Wang, and Cheng-Zhong Xu, "iBalloon: Self-Adaptive Virtual Machines Resource Provisioning", 2010, pp 1-12
9. C.Chen,Wei Wang, and Bo Li, "Performance-aware fair scheduling: Exploiting demand elasticity of data analytics jobs", In IEEE INFOCOM 2018-IEEE Conference on Computer Communications, pp. 504-512. IEEE, 2018.
10. P.Armstrong,Ashok Agarwal, A. Bishop, Andre Charbonneau, R. Desmarais, K. Fransham, N. Hill et al. "Cloud scheduler: a resource manager for distributed compute clouds", arXiv preprint arXiv:1007.0050 (2010).
11. R.Garg,Awadhesh Kumar Singh,"Adaptive workflow scheduling in grid computing based on dynamic resource availability", Engineering Science and Technology, an International Journal 18, no. 2 (2015): 256-269.
12. Ali Nikravesh, Yadavar, Samuel A. Ajila, and Chung-Hong Lung, "An autonomic prediction suite for cloud resource provisioning", Journal of Cloud Computing 6, no. 1 (2017): 3.
13. G.Juve, Ewa Deelman,"Automating application deployment in infrastructure clouds", In 2011 IEEE Third International Conference on Cloud Computing Technology and Science, pp. 658-665. IEEE, 2011.
14. G. Juve, Ann Chervenak, Ewa Deelman, Shishir Bharathi, Gaurang Mehta, and Karan Vahi,"Characterizing and profiling scientific workflows", Future Generation Computer Systems 29, no. 3 (2013): 682-692.
15. G,Juve, Ewa Deelman, Karan Vahi, Gaurang Mehta, Bruce Berriman, Benjamin P. Berman, and Phil Maechling,"Scientific workflow applications on Amazon EC2", In 2009 5th IEEE international conference on e-science workshops, pp. 59-66. IEEE, 2009.
16. F. Deng, Fuhu, Miao Lai, and Ji Geng, "Multi-Workflow Scheduling Based on Genetic Algorithm", In 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 300-305. IEEE, 2019.
17. G.Juve, Ewa Deelman,"Scientific workflows in the cloud", In Grids, Clouds and Virtualization, pp. 71-91. Springer, London, 2011.
18. V.Khandelwal, Anand Kishore Chaturvedi, and Chandra Prakash Gupta,"Bidding Strategies for Amazon EC2 Spot Instances-A Comprehensive Review", In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1-5. IEEE, 2018.
19. G.Lee,Randy H. Katz, "Heterogeneity-Aware Resource Allocation and Scheduling in the Cloud," In HotCloud. 2011.
20. Young Lee, Hyuck Han, Albert Y. Zomaya, and Mazin Yousif,"Resource-efficient workflow scheduling in clouds", Knowledge-Based Systems 80 (2015): 153-162.
21. Mohammad Masdari, Sima ValiKardan, Zahra Shahi, and Sonay Imani Azar, "Towards workflow scheduling in cloud computing: a comprehensive analysis", Journal of Network and Computer Applications 66 (2016): 64-82.
22. Lavanya Ramakrishnan, Jeffrey S. Chase, Dennis Gannon, Daniel Nurmi, and Rich Wolski, "Deadline-sensitive workflow orchestration without explicit resource control", Journal of parallel and distributed computing 71, no. 3 (2011): 343-353.
23. Mohsen Amini Salehi, and Rajkumar Buyya,"Adapting market-oriented scheduling policies for cloud computing", In International Conference on Algorithms and Architectures for Parallel Processing, pp. 351-362. Springer, Berlin, Heidelberg, 2010.
24. Henrique Yoshikazu Shishido, Júlio Cezar Estrella, Claudio Fabiano Motta Toledo, and Marcio Silva Arantes, "Genetic-based algorithms applied to a workflow scheduling algorithm with security and deadline constraints in clouds", Computers & Electrical Engineering 69 (2018): 378-394.
25. Quanwang Wu, MengChu Zhou, Qingsheng Zhu, Yunni Xia, and Junhao Wen,"MOELS: Multiobjective Evolutionary List Scheduling for Cloud Workflows", IEEE Transactions on Automation Science and Engineering (2019).
26. Miao Zhang, Huiqi Li, Li Liu, and Rajkumar Buyya, "An adaptive multi-objective evolutionary algorithm for constrained workflow scheduling in Clouds", Distributed and Parallel Databases 36, no. 2 (2018): 339-368.
27. Zhaomeng Zhu, and Xueyan Tang, "Deadline-constrained workflow scheduling in IaaS clouds with multi-resource packing", Future Generation Computer , 101 (2019): 880-893.

AUTHORS PROFILE



Mrs.S.M.Jaybhaye is PhD Scholar under the guidance of Dr.V.Z.Attar at Computer and Information Technology, College of Engineering Pune. Her educational qualification is ME (Computer Engineering). Her area of research is cloud computing, optimization, resource provisioning.



Dr.V.Z.Attar is working as Associate Professor and Head of Computer Engg. and Information Technology Department, College of Engineering Pune. Her educational qualification is Ph.D(Computer Engg) and having 24 years' experience in teaching & research. Her area of research interest are Cloud computing Multimedia Databases, algorithms.