ON-DEMAND OPTIMUM RESOURCE PROVISIONING ON CLOUD

Sneha Jani[1], Gayatri Pandi(Jain)[2]

[1]PG Researcher, Computer Department, L.J. Institute of Engineering and Technology, Gujarat, India
[2]Head of department, Post Graduation Department, L.J. Institute of Engineering and Technology, Gujarat, India

ABSTRACT

Cloud computing is an advanced model for delivering information technology services in which resources are retrieved from the internet through web based tools and applications. The important challenges in cloud computing are security, resources allocation and resources provisioning. The important aim in resource provisioning is maximum performance in minimum time and reduce the amount of data transfer with minimum cost. When the workload of services increases rapidly, the Quality of Service(QoS) of the hosted application may degrade and the Service Level Objective(SLO) will be violated.

Keyword : - *Cloud Computing, Resources Provisioning, Workload, Quality of Service, Service Level Objective*

I. INTRODUCTION

Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in third-party data centres. It is the practice of using a network of remote servers hosted on the Internet to store, manage and process data , rather than a local server or a personal computer[6].

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories:

1. Software-as-a-Service (SaaS)
2. Platform-as-a-Service (PaaS)
3. Infrastructure-as-a-Service (IaaS)

A cloud deployment model represents a specific type of cloud environment, primarily distinguished by ownership, size, and access. There are four common cloud deployment models:[7]

1. Public Cloud
2. Private Cloud
3. Hybrid Cloud
4. Community Cloud

Cloud computing is a large-scale distributed computing paradigm in which a pool of computing resources is available to users (called cloud consumers) via the Internet. Computing resources, e.g., processing power, storage, software, and network bandwidth, are represented to cloud consumers as the accessible public utility services. Infrastructure- as-a-Service (IaaS) is a computational service model widely applied in the cloud computing paradigm [2].

Elasticity has now become the elemental feature of cloud computing as it enables the ability to dynamically add or remove virtual machine instances when workload changes. However, effective virtualized resource management is still one of the most challenging tasks. When the workload of a service increases rapidly, existing approaches cannot respond to the growing performance requirement efficiently because of either inaccuracy of adaptation decisions or the slow process of adjustments, both of which may result in insufficient resource provisioning [4].Cloud computing has experienced a rapid development and gained a huge success in the past ten years. A large number of commercial cloud service providers(CSPs) begin to deliver various public cloud computing services. More and more enterprises and organization build their own cloud computing infrastructure or resort to hybrid cloud. Efficient resource management not only enhances the quality of service, but also reduces resources consumption. The growing challenge is how to efficiently provision resources to meet the requirement of quality of service (QoS).Resource provision is the most fundamental issue of cloud computing service deployment [9].

II. RESOURCE PROVISIONING

In cloud computing, a resource provisioning mechanism is required to supply cloud consumers a set of computing resources for processing the jobs and storing the data. Cloud providers can offer cloud consumers two resource provisioning plans, namely short-term on-demand and long-term reservation plans. Efficient resource provision which can guarantee the satisfactory cloud computing services to the end user, lays the foundation for the success of commercial competition [9].Resource provisioning is the allocation of a cloud provider's resources to a customer. When a cloud provider accepts a request from a customer, it must create the appropriate number of virtual machines (VMs) and allocate resources to support them.

The process is conducted in several different ways:

- ➔ **Advance provisioning :** With **advance provisioning**, the customer contracts with the provider for services and the provider prepares the appropriate resources in advance of start of service. The customer is charged a flat fee or is billed on a monthly basis.
- ➔ **Dynamic provisioning :** With **dynamic provisioning**, the provider allocates more resources as they are needed and removes them when they are not. The customer is billed on a pay-per-use basis. When dynamic provisioning is used to create a hybrid cloud, it is sometimes referred to as cloud bursting**.**
- ➔ **User self-provisioning :**With **user self-provisioning** (also known as cloud self-service), the customer purchases resources from the cloud provider through a web form, creating a customer account and paying for resources with a credit card. The provider's resources are available for customer use within hours, if not minutes [8].

III. RELATED WORKS

VARIOUS RESOURCE PROVISIONING TECHNIQUES

A. Particle Swarm Optimization(PSO) algorithm and Simulated Annealing(SA) algorithm

Marwah Hashim Eawna et al., 2015[1] presented dynamic resources provisioning in multi-tier application by using meta-heuristic technique such as Particle Swarm Optimization (PSO) algorithm, Simulated Annealing (SA) algorithm and hybrid algorithm that combine Particle Swarm Optimization (PSO) and Simulated Annealing (SA). In PSO algorithm, there is calculated average computation cost of all tasks on all the compute resources. There is used PSO as a local searching select local best position (*Lbest*) and global searching to select global best position (*Gbest*). To improve optimal performance of PSO, *Gbest* can be searched by SA after every iteration of particle swarm, whose result can be taken as new *Gbest* of PSO system.

PSO algorithm in multi tier application

Step 1: Generate the initial population.

Step 2: Calculate the objective function value for each individual depend on following equations:

ECT=ST+DU+EET (1)

EET= (ST-FT) + DU (2)

The ECT represent expected completion time, ST represent first time, DU represent duration time between arrival requests till to start, EET represent the estimated execution time and FT represent Finish time.

Step 3: Sort the initial population based on the objective function values. .

Step 4: Select the local best position (*Lbest*).

Step 5: Select the global best position (*Gbest*) based on *Lbest.*

Step 6: i=i+1, check if the end of iteration, else go back to Step 4.

Step 7: Check if *Gbest* need to migrate and get the output.

SA algorithm in multi tier application

Step 1: Generate the initial population.

Step 2: Calculate the objective function value for each individual based on Equ.1 and Equ2.

The ECT represent expected completion time, ST represent first time, DU represent duration time between arrival requests till to start, EET represent the estimated execution time and FT represent Finish time.

Step 3: Sort the initial population depend on the objective function values. .

Step 4: Select the global best position (*Gbest*).

Step 5: Apply SA to search around the global solution *Gbest*. If the solution obtained by SA is better than previous *Gbest* then swap with new *Gbest*.

Step 6: i=i+1, check if the end of iteration else go back to Step 4.

Step 7: Check if *Gbest* need to migrate and get the output.

PSO-SA algorithm in multi tier application

Step 1: Generate the initial population and initial velocity.

Step 2: Calculate the objective function value for each individual based on Equ1 and Equ2.

The ECT represent expected completion time, ST represent first time, DU represent duration time between arrival requests till to start, EET represent the estimated execution time and FT represent Finish time.

Step 3: Sort the initial population depend on the objective function values. .

Step 4: Select the local best position (*Lbest*).

Step 5: Select the global best position (*Gbest*) based on *Lbest.*

Step 6: Apply SA to search around the global solution *Gbest*. If the solution obtained by SA is better than previous *Gbest* then swap with new *Gbest*.

Step 7: i=i+1, check if the end of iteration else go back to Step 4.

Step 8: Check if *Gbest* need to migrate and get the output.

The resource provisioning based on PSO-SA algorithm in multi-tier application is much faster than resource provisioning in multi-tier application based on PSO algorithm and SA algorithm, that is beneficial in the development of cloud computing.

B. Optimal Cloud Resource Provisioning(OCRP) algorithm

Sivadon Chaisiri et al.,2012[2] has proposed the optimal cloud resource provisioning algorithm for the virtual machine management. The optimization formulation of stochastic integer programming is proposed to obtain the decision of the OCRP algorithm as such the total cost of resource provisioning in cloud computing environments is minimized. The optimal solution obtained from OCRP is obtained by formulating and solving stochastic integer programming with multistage resource.

C. Variable size bin packing greedy algorithm, GA algorithm

Alok Gautam Kumbhare et al., 2015[3] has developed the concept of " dynamic dataflows" which utilize alternate tasks as additional control over the dataflow's cost and QoS. They formalize an optimization problem to represent deployment and runtime resource provisioning that allows to balance the application's QoS, value, and the resource cost. They proposed two greedy heuristics, centralized and sharded, based on the variable-sized bin packing algorithm and compare against a Genetic Algorithm (GA) based heuristic that gives a near-optimal solution. A large-scale simulation study, using the linear road benchmark and VM performance traces from the AWS public cloud, shows that while GA-based heuristic provides a better quality schedule, the greedy heuristics are more practical, and can intelligently utilize cloud elasticity to mitigate the effect of variability, both in input data rates and cloud resource performance, to meet the QoS of fast data applications.

D. SPRNT strategy

Jinzhao Liu et al., 2015[4] introduce SPRNT, a novel resource management framework, to ensure high-level QoS in the cloud computing system. SPRNT utilizes an aggressive resource provisioning strategy which encourages SPRNT to substantially increase the resource allocation in each adaptation cycle when workload increases. This strategy first provisions resources which are possibly more than actual demands, and then reduces the over-provisioned resources if needed. By applying the aggressive strategy, SPRNT can satisfy the increasing performance requirement in the first place so that the QoS can be kept at a high level. The experimental results show that SPRNT achieves up to 7.7* speedup in adaptation time, compared with existing efforts. By enabling quick adaptation, SPRNT limits the SLO violation rate up to 1.3 % even when dealing with rapidly increasing workload.

E. Dynamical Request Redirection and Resource Provisioning(DYRECEIVE) method

As user demands are difficult to predict and the prices of the VMs vary in different time and region, optimizing the number of VMs of each type rented from datacenters located in different regions in a given time frame becomes essential to achieve cost effectiveness for VSPs. It is equally important to guarantee users' Quality of Experience (QoE) with rented VMs. Wenhua Xiao et al., 2016[5] give a systematic method called Dynamical Request Redirection and Resource Provisioning (DYRECEIVE) to address this problem. They formulate the problem as a stochastic optimization problem and design a Lyapunov optimization framework based online algorithm to solve it. This method is able to minimize the long-term time average cost of renting cloud resources while maintaining the user QoE.

IV. PROBLEM FORMULATION

The problem of efficient resource provisioning remains a challenging task in IaaS clouds, particularly when workload increases at a high speed. Existing efforts can deal with the general resource provisioning task well, but may fail when faced with rapidly increasing workload, thus resulting in degradation of QoS.

There is presented for resource provisioning in multi-tier cloud computing based on PSO and SA and hybrid algorithm that combine PSO and SA algorithm, Simulation of presented algorithms shows that provisioning resource based on hybrid PSO-SA algorithm are good that take less average execution time as compared with resources provisioning based PSO and SA algorithms as alone in multi-tier cloud computing.

In this paper, there is used meta-heuristic technique based on provisioning resources and considered only execution time but it is necessary to evaluate response time, storage memory space and optimal cost in dynamic cloud request scenario.

V. PROPOSED WORK

Different approaches have been used in resource provisioning, the only existing resources provisioning approach in cloud computing using meta-heuristic technique are based on multi-tier application which considered only execution time. In our proposed algorithm, there will be evaluated response time, storage memory space and optimal cost in dynamic cloud request scenario.

In our proposed algorithm, in each tier in multi-tier application we use PSO as a local searching select local best position (*Lbest*) and global searching to select global best position (*Gbest*), and use SA to search around *Gbest*; in other words, *Lbest* and *Gbest* changes in each iteration.

To implement the Enhanced PSO-SA algorithm the following steps should be taken and repeated to each tier of multi-tier application.

Enhanced PSO-SA Algorithm :

Step 1: Generate the initial population.
Step 2: Calculate the objective function value for each individual depend on following equations.

ECT = ST + DU + EET
EET = (ST - FT) + DU
Response Time = Service Time/(1-CPU Utilizaton) ^ no of effective servers
Memory space = Total memory – Used memory
Cost = Hours of usage + VM type + Storage + Bandwidth used

Step 3: Sort the initial population depend on the objective function values. .
Step 4: Select the local best position (*Lbest*).
Step 5: Select the global best position (*Gbest*) based on *Lbest.*
Step 6: Apply SA to search around the global solution *Gbest*. If the solution obtained by SA is better than previous *Gbest* then swap with new *Gbest*.

Step 7: i=i+1, check if the end of iteration else go back to Step 4.
Step 8: Check if *Gbest* need to migrate and get the output.

Where,

ECT represents expected completion time,
ST represent first time,
DU represents duration time between arrival requests till to start,
EET represent the estimated execution time
FT represents Finish time
Service Time = 1 – (Service Time Workload/No of server)^No of server

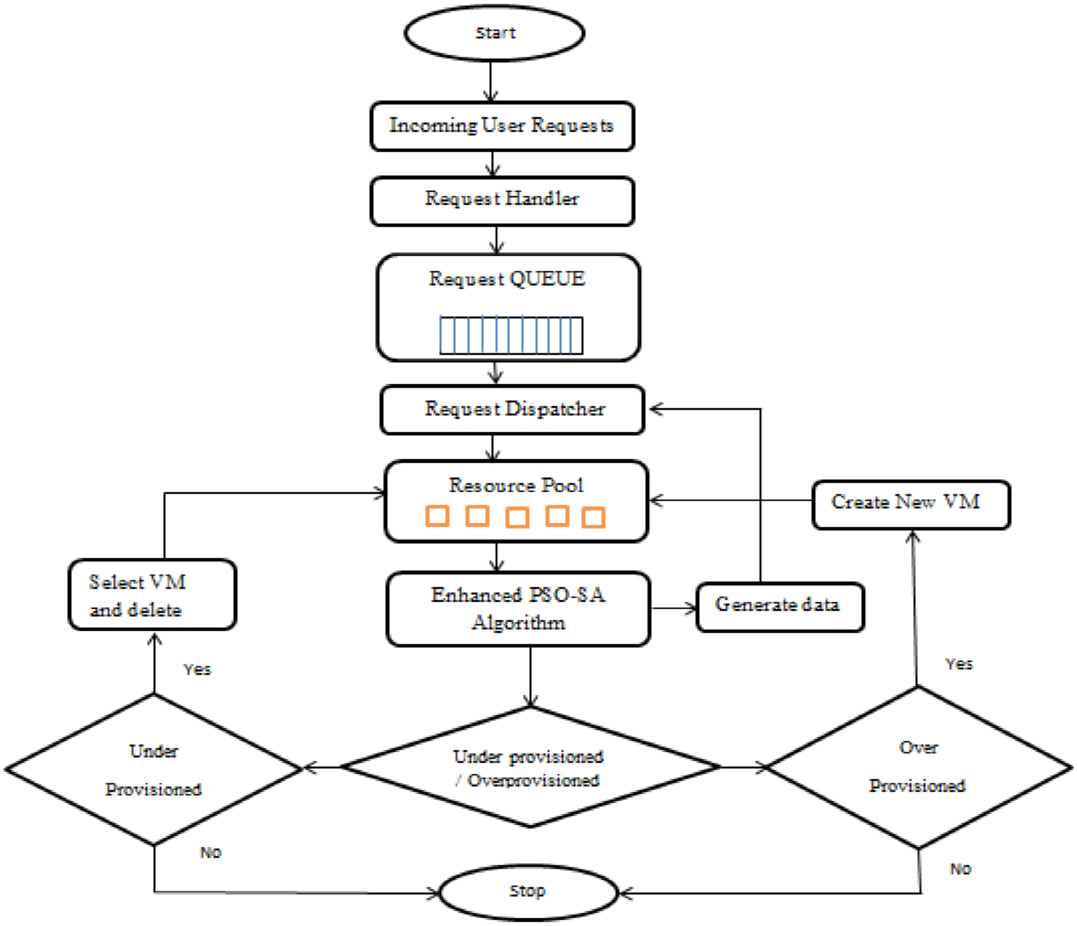


Fig-1 Proposed Flowchart

Steps of workflow :

Step 1 : Request handler captures all user requests.
Step 2 : Request handler is continuously monitoring Queueing Agent to allocate request to optimum resource which is provisioned.
Step 3 : Request Dispatcher will allocate requests to the specific machines based on Enhanced PSO-SA algorithm.
Step 4 : Enhanced PSO-SA algorithm will also take care of overprovisioned and underprovisioned resources by scaling mechanism.

VI. IMPLEMENTATION

To verify the efficiency and effectiveness of proposed algorithms, we use the Amazon Web Services(AWS) to provide resource based on the proposed algorithms. The implementation results of resource provisioning based on enhanced PSO-SA algorithm in multi-tier application are shown below:

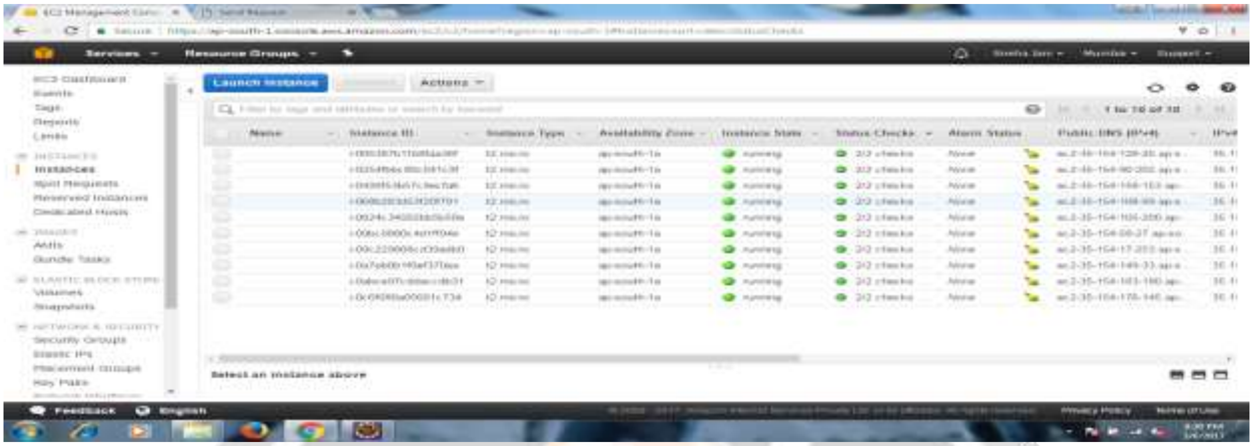
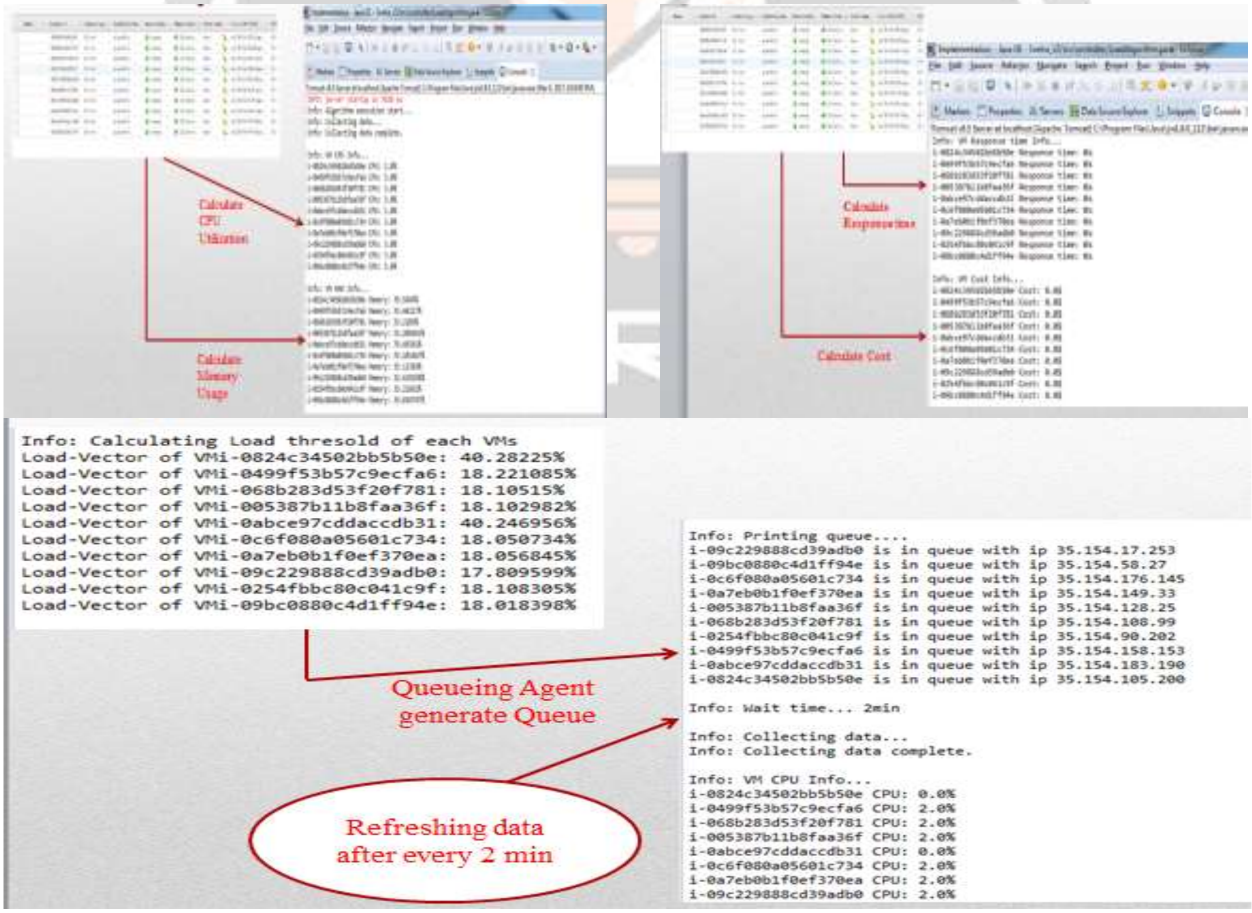


Fig-2 Running instances on Amazon EC2


Calculate CPU Utilization
Calculate Memory Usage
Calculate Responsetime
Calculate Cost
Queueing Agent generate Queue
Refreshing data after every 2 min

```
Info: Calculating Load thresold of each VMs
Load-Vector of VMi-0824c34502bb5b50e: 40.28225%
Load-Vector of VMi-0499f53b57c9ecfa6: 18.221085%
Load-Vector of VMi-068b283d53f20f781: 18.10515%
Load-Vector of VMi-005387b11b8faa36f: 18.102982%
Load-Vector of VMi-0abce97cddaccdb31: 40.246956%
Load-Vector of VMi-0c6f080a05601c734: 18.050734%
Load-Vector of VMi-0a7eb0b1f0ef370ea: 18.056845%
Load-Vector of VMi-09c229888cd39adb0: 17.809599%
Load-Vector of VMi-0254fbbc80c041c9f: 18.108305%
Load-Vector of VMi-09bc0880c4d1ff94e: 18.018398%
```

```
Info: Printing queue....
i-09c229888cd39adb0 is in queue with ip 35.154.17.253
i-09bc0880c4d1ff94e is in queue with ip 35.154.58.27
i-0c6f080a05601c734 is in queue with ip 35.154.176.145
i-0a7eb0b1f0ef370ea is in queue with ip 35.154.149.33
i-005387b11b8faa36f is in queue with ip 35.154.128.25
i-068b283d53f20f781 is in queue with ip 35.154.108.99
i-0254fbbc80c041c9f is in queue with ip 35.154.90.202
i-0499f53b57c9ecfa6 is in queue with ip 35.154.158.153
i-0abce97cddaccdb31 is in queue with ip 35.154.183.190
i-0824c34502bb5b50e is in queue with ip 35.154.105.200

Info: Wait time... 2min

Info: Collecting data...
Info: Collecting data complete.

Info: VM CPU Info...
i-0824c34502bb5b50e CPU: 0.0%
i-0499f53b57c9ecfa6 CPU: 2.0%
i-068b283d53f20f781 CPU: 2.0%
i-005387b11b8faa36f CPU: 2.0%
i-0abce97cddaccdb31 CPU: 0.0%
i-0c6f080a05601c734 CPU: 2.0%
i-0a7eb0b1f0ef370ea CPU: 2.0%
i-09c229888cd39adb0 CPU: 2.0%
```

Fig-3 Console Outputs

There are created instances on AWS. There is taken information of each instances at every 2 minutes and generate the queue which gives the optimal decisions to find LBest and GBest resource.

VII. CONCLUSION

As user demands are difficult to predict and the prices of the resources vary in different time and region, optimizing the number of resources of each type rented from datacenters located in different regions in a given time frame becomes essential to achieve cost effectiveness for CSPs. It is necessary to find the algorithm to provision resources dynamically with fulfilling QoS efficiently.

The work presented in this paper provides an optimal result provisioning resources for a cloud computing. A new method is proposed for resource provisioning in multi-tier cloud computing based on PSO and SA and hybrid algorithm that combine PSO and SA algorithm. Implementation of our proposed algorithms shows that provisioning resource based on enhanced PSO-SA algorithm are good that take less average execution time, less response time less memory storage space, less CPU usage and optimal cost in multi-tier cloud computing.

VIII. REFERENCES

[1] Eawna, Marwah Hashim, Salma Hamdy Mohammed, and El-Sayed M. El-Horbaty. "Hybrid Algorithm For Resource Provisioning Of Multi-Tier Cloud Computing". *Procedia Computer Science* 65 (2015): 682-690. Web.

[2] S. Chaisiri, B. Lee and D. Niyato, "Optimization of Resource Provisioning Cost in Cloud Computing", *IEEE Transactions on Services Computing*, vol. 5, no. 2, pp. 164-177, 2012.

[3] Kumbhare, Alok Gautam et al. "Reactive Resource Provisioning Heuristics For Dynamic Dataflows On Cloud Infrastructure". *IEEE Transactions on Cloud Computing* 3.2 (2015): 105-118. Web.

[4] Liu, Y. Zhang, Y. Zhou, D. Zhang and H. Liu, "Aggressive Resource Provisioning for Ensuring QoS in Virtualized Environments", *IEEE Transactions on Cloud Computing*, vol. 3, no. 2, pp. 119-131, 2015.

[5] W. Xiao, W. Bao, X. Zhu, C. Wang, L. Chen and L. Yang, "Dynamic Request Redirection and Resource Provisioning for Cloud-Based Video Services under Heterogeneous Environment", *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 7, pp. 1954-1967, 2016.

[6] Lee Badger,TimGrance, Robert Patt-Corner,JeftVoas,"Cloud Computing Synopsis and Recommendations", National Institute of Standards and Technology, Special Publication 800- 146, , May 2012.

[7] Rajkumar Buyya et. el., Cloud Computing: Principles and Paradigms, Wiley India Edition

[8] "What is cloud provisioning? - Definition from WhatIs.com", *SearchCloudProvider*, 2016. [Online]. Available: http://searchcloudprovider.techtarget.com/definition/cloud- provisioning. [Accessed: 28- Aug- 2016].

[9] Jiangtao Zhang, HejiaoHuang, XuanWang. “Resource provision algorithms in cloud computing : A survey”.ELSEVIER(2016): 23-42.