Resource Provisioning in Single Tier and Multi-Tier Cloud Computing: "State-of-the-Art"

Marwah Hashim Eawna
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

Salma Hamdy Mohammed
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

El-Sayed M. El-Horbaty
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

***Abstract*—Cloud computing is a new computation trend for delivering information as long as an electronic device needs to access of a web server. One of the major pitfalls in cloud computing is related to optimizing the resource provisioning and allocation. Because of the uniqueness of the model, resource provisioning is performed with the objective of minimizing time and the costs associated with it. This paper reviews the state-of-the-art of managing resources of the cloud environments in theoretical research. This study discusses the performance and analysis for well-known cloud provisioning resources techniques, single tier and multi-tier.**

***Keywords*—Cloud Computing; Resource Provisioning**

I. INTRODUCTION

The cloud environment describes a company, organization or individual that uses a web-based application for every task rather than installing software or storing data on a computer. All cloud environments are not common but a move toward this is a long-term goal for cloud computing enthusiasts and cloud capitalists.

Many challenges are influenced to adopt the cloud computing technology such as security, resources allocation, resources provisioning and others. In provisioning resource for cloud computing environment the major challenge is to determine the right amount of resources required for the execution of work in order to minimize the financial cost from the perspective of users and to maximize the resource utilization from the perspective of service providers [1].

The resource provisioning must meet Quality of Service (QoS) parameters like availability, throughput, response time, security, reliability etc., and thereby avoiding Service Level Agreement (SLA) violation [2]. There are entirely two generic way of resource provisioning, Static and dynamic.

1) Static Resource Provisioning: usually provides the peak time needed resource all the time for the application. In this kind of provisioning most of the time the waste of resource due to workload is not in a peak, but resource providers provide the maximum required resource to prevent SLA violation.

2) Dynamic Resource Provisioning: the basic fundamental idea in the latter way is providing the resources based on the application needs, this helps the provider to assign the Non-loaded resources (which become free to used now) to the new users. This method reduces a fraction of providers' development costs by utilizing current available resources and beside that the user can happily just pay for the amount of the resources which were really used [3]. Moreover, the parameters of resource provisioning is presented as:

1) Response time: The resource provisioning algorithm designed must take minimal time to respond when executing the task.

2) Minimize Cost: From the Cloud user point of view cost should be minimized.

3) Revenue Maximization: This is to be achieved from the Cloud Service Provider's view.

4) Fault tolerant: The algorithm should continue to provide service in spite of failure of nodes.

5) Reduced SLA Violation: The algorithm designed must be able to reduce SLA violation.

6) Reduced Power Consumption: virtual machine placement & migration techniques must lower power consumption [4].

The rest of this paper is organized as follows: Sections 2 contains a review of resource provisioning in single tier and multi-tier cloud environments. Section 3 illustrates the most popular efficiency of Single tier and Multi-tier architectures. Finally, Conclusions and future works are given in Section 4.

II. RELATED WORK

This section explains two basic architectures are dependent in the resources provision in cloud computing environment like single tier and multi-tier architecture.

A. Single-tier technique

A single-tier architecture of cloud computing has a set of servers that are used to provide resources by receiving

requests from user in presentation server and looking information in application server and store the information in database server.

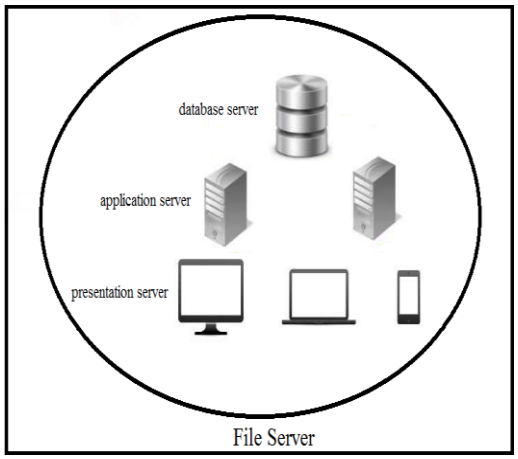

Fig. 1. Single tier architecture in cloud computing

As show in figure1, single tier architecture there are no interactive between the servers and is not allowed to share resources between server that cause consumption of resources and loss of time when compared with multi-tier architecture.

Some of the existing techniques involving nature-inspired meta-heuristics have become the new focus of resource allocation. For example, Tarun goyal et al. [5] Scheduling a model based on minimum network delay using Suffrage Heuristic coupled with Genetic algorithms for scheduling sets of independent jobs algorithm is proposed, the result show that the schedule of multiple jobs on multiple machines in an efficient manner such that the jobs take the minimum time for completion.

Othman et al. [6] proposed a novel Simulated Annealing (SA) algorithm for scheduling tasks in cloud environments. SA exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system. Results show that this approach for job scheduling not only guarantees the QoS requirement of customer job but also ensures to make best profit of cloud providers. It has also concern about real execution time of jobs in different systems as well as deadline and penalty cost in the algorithm.

Shaobin Zhan et al. [7] introduces an improved Particle Swarm Optimization (IPSO) algorithm in resources scheduling strategy by Add simulated annealing into every iteration of PSO, through experiments, the results show that this method can reduce the task average running time, and raises the rate availability of resources.

Talwinder Kaur et al. [8] Improved Particle Swarm Optimization (IPSO), Simulated Annealing (SA) Algorithm, and Hybrid Particle Swarm Optimization-Simulated Annealing algorithms based on utilizing and scheduling resources. The experiment show that by using this algorithm provisioning resource in very less time as compared to the existing algorithm.

Xiaotang Wen el al. [9] improved algorithm to provide resource by combine Ant Colony Optimization (ACO) with particle swarm optimization algorithm to improve the efficiency of resource scheduling in cloud computing environments. Table 1 summarizes the advantages and disadvantages of single tier mechanisms are surveyed in this section.

TABLE I. RESOURCE PROVISIONING ALGORITHMS IN SINGLE TIER TECHNIQUES

Parameters	Techniques	Attributes	Authors
Time	Scheduling model based on GA	minimize the make span	Tarun goyal et al. [5]
	SA-based approach for sche-duling in cloud envi-ronment	minimize execution time	Monir Abdullah et al. [6]
	PSO algorithm in resources scheduling strategy	reduce the task average running time, and raises the rate availability of resources	Shaobin Zhan et al. [7]
	Use PSO and SA algorithm	less execution time as compare with existing algorithm	Talwinder et al.[8]
	Use PSO and ACO algorithm	Efficiently and speed to provide resource	Xiaotang et al.[9]

B. Multi-tier technique

This architecture partitions the application process into multiple tiers. Each tier provides certain functionality. The benefit of such architecture is that it can provide a high level of scalability and reliability. However, the resource allocation among these tiers will be more difficult due to the interdependency between the tiers. A multi-tier cloud computing application may span multiple nodes. Specifically, most multi-tier cloud computing applications use 3-tier architecture.

As shown in figure 2, the first tier, named presentation tier, consists of Web servers. It displays what is presented to the user on the client side within their Web browsers. For the Web server tier, it mainly has three functions:

1) Admitting/denying requests from clients and services static Web requests.

2) Passing requests to the Application server.

3) Receiving response from Application server and sending them back to clients. Examples of Web servers include Apache Server and Microsoft Internet Information Server (IIS).

The second tier, named business tier, consists of Application servers. Business logic processing is performed at this tier. There are also three functions at the Application server tier:

1) Receiving requests from the Web server.

2) Looking up information in the database and processing the information.

3) Passing the processed information back to the Web server.

The last tier, named data tier, consists of database servers. It handles database processing and data accessing. Database server tier is used to store and retrieve a Web site's information (e.g., user accounts, catalogs to reports, and customer orders) [10].

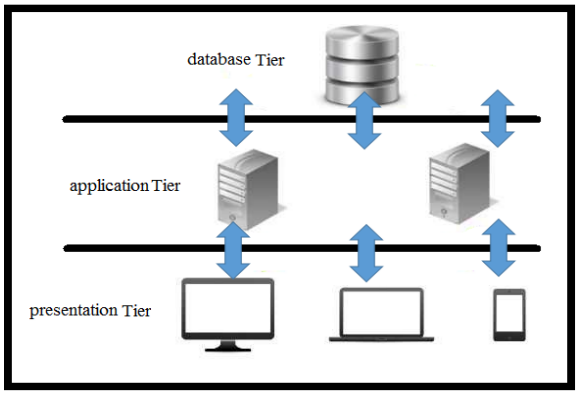

Fig. 2. Multi-tier architecture of cloud computing

Hadi Goudarzi el al. [11] considered the problem of the resource allocation to optimize the total profit gained from the SLA contracts and lost from operational cost. They assume that servers are characterized by their maximum capacity in three dimensions: processing power, memory usage, and communication bandwidth. While guaranteeing SLAs for clients with applications that require multiple tiers of service to complete.

Bhuvan Urgaonkar et al. [12] propose a novel data center architecture based on virtual machine monitors to reduce provisioning overheads, this technique reduced the overhead of switching servers across applications from several minutes to less than a second, while meeting the performance targets of residual sessions.

Chandra at el. [13] considered the problem of resource allocation on shared data centers, where they modeled a server resource as a generalized process sharing server and used a time-domain description of the server to model transient system states. These techniques can judiciously allocate system resources, especially under transient overload conditions

Heng et al. [14] use a benefit-aware approach with feedback control theory to solve the problem of continuously guarantees the SLA in the new configuration in multi-tier application. This approach can reduce resource provisioning cost by as much as 30% compared with a cost oblivious approach, and can effectively reduce SLA violations compared with a cost-aware approach.

TABLE II. RESOURCE PROVISIONING ALGORITHMS IN MULTI-TIER TECHNIQUES

Parameters of	Techniques	Attributes	References
Time	novel dynamic provisioning technique	double the application capacity reduced the overhead of switching servers from several minutes to less than a second	Bhuvan Urgaonkar et al. [11]
SLA	model for SLA-based multi-dimensional resource allocation scheme	meet SLA and effectiveness	Hadi Goudarzi el al. [12]
	novel benefit-aware provisioning approach	effective in reducing both cost and SLA violations	Heng et al. [13]
	model the server resource by use a time-domain description of (GPS) server	judiciously allocate system resources	Abhishek .[14]

Table 2 summarizes resource provisioning in multi-tier technique, these techniques considered SLA and real execution time of job in different system as well as soft deadline and penalty cost in the algorithm. Table2 ensures that prevent SLA violated and give a good profit for the different cloud provider.

III. EFFICIENCY OF SINGLE TIER AND MULTI-TIER ALGORITHMS

There are several algorithm has been developed to provide a better scheduling in a single tier and multi-tier cloud environment. Experiment by using CloudSim shows that the improved algorithm not only accelerated the convergence speed, but also avoided falling into local optimum solution, and achieved the purpose that the user tasks were efficiently provided appropriate resources in cloud computing, which improved the resource utilization ratio.

As shown in figure 3, the algorithm that combine between PSO and ACO takes time more than other algorithm that combine between PSO and SA.

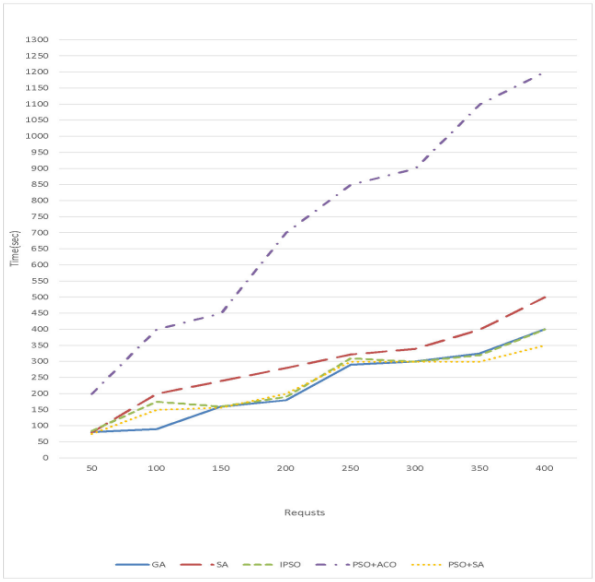

Fig. 3. Comparison among algorithms of single tier architecture

For instance, if the number of requests is 400 in the algorithm that combine between PSO and ACO the complete execution time will be 1200 sec. and in the algorithm that combine between PSO and SA the complete execution time will be 350 sec.

Furthermore, the Pseudo code of the algorithm combine between PSO and SA is iullstarted in figure 4.

```
1: procedure PSO-SA algorithm
2: CalculateExecTimes();
3: initSwarm ();
4: initGlobalBest();
5: for i=0 to numberIterations do
6: for j=0 to numberParticles do
7: calculateInertiaValue();
8: calculateNewVelocities();
9: calculateNewPositions();
10: calculateFitnessValue();
11: evaluateSolution();
12: updateParticleMemory();
13: updateGlobalBest();
14.UpdateGlobalBestDependonSAalgorithm
15: end for
16: end for
17: end procedure
```

Fig. 4. Pseudo code of algorithm that combine between PSO and SA based resource provisioning

On the other side, the multi-tire algorithms are surveyed in [11], [12], [13], and [14] have been implemented based on CloudSim environment. Figure 5 shows the efficiency of these algorithms.

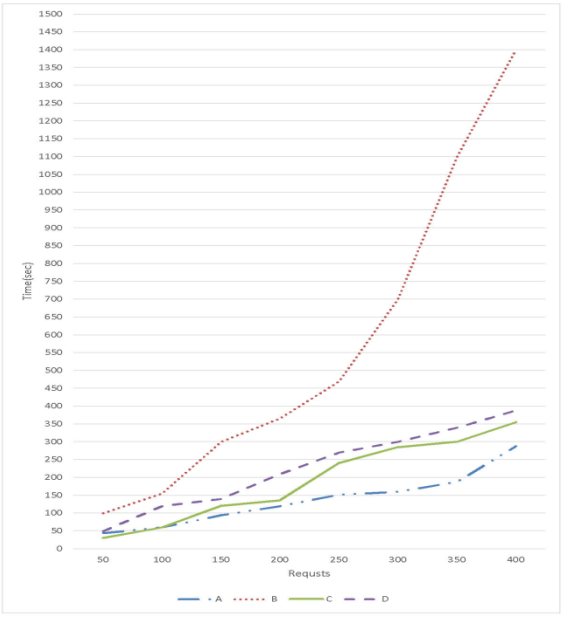

Fig. 5. Comparison among algorithms of multi-tier architecture

According to figure 5, the algorithm of novel dynamic provisioning technique [B] takes time more than others. However, less time is needed for the algorithms that use multi-dimensional SLA-based resource allocation [A]. Finally, figure 6 shows the Pseudo code of the multi-dimensional SLA algorithm in multi-tier technique**.**

Algorithm *Resource_Consolidate* ()
// Search the solution space to find better profit
TP = total profit;
Initialize the forces between clients and servers;
// calculate force differentials
$D_{i,j\to k}^{\alpha,t} = F_{ik}^{\alpha,t} - F_{ij}^{\alpha,t}$; $\forall j,i,t,\alpha$
ΔF = 1;
While (ΔF > 0) {
ΔF = max ($D_{ij\to k}^{\alpha,t}$); // client i and α
j = selected source server;
k = selected destination server type;
g = selected destination server;
If (ΔF is toward an ON server in server type k){
g = find the least busy server in k, assigned to tier t;
If (lower bound constraints satisfied) goto Re-Assign;
Else goto skip Re-Assign;}
Else If (ΔF is toward an OFF server in server type k){
g = find an OFF server in k;
If (found an OFF server) goto Re-Assign;
Else goto skip Re-Assign;}
Else If (ΔF is toward a server serving client i) goto Re-Assign;
Re-Assign: Re-assign α portion of the requests to g from j;
Update force related to j, g and client i;
P = total profit;
If (P>TP) TP = P; save the state;}
Skip Re-Assign: Update the move limit; }

Fig. 6. Pseudo code of the algorithm using multi-dimensional SLA-based resource provisioning

IV. Conclusions and Future Work

The investigations which were studied above are trying to optimize and utilize the resources. Several methods were mentioned here which used different parameters as a goal for resource provisioning such as response time, rejection rate, service level agreement (SAL) violation rate, cost etc. For provisioning planning should take appropriate provisioning times, Provisioning resources too soon will wastes our resources and therefore our money, on the other side provisioning resources too late will cause potentially SLA violations and makes the users angry.

Several ideas were reviewed and it can be concluded from them that managing such big resources for human administrators is not possible anymore and administrators are going to be replaced with managing systems. These systems must use techniques that are able to estimate and allocated resource in the most efficient way while avoiding SLA violations. New trends are needed with less human intervention so some new search techniques involving nature-inspired meta-heuristics have become the new focus of resource allocation research by using it provide very good solutions in a reasonable time. Provisioning resources by using multi-tier architecture that given continuously guarantees the SLA and provide a high level of scalability and reliability but the resource provisioning by multi-tier architecture is more difficult due to the fact that the resource demand at each tier is different. However, Single-tier architecture has relatively simple structure and is easy to setup.

So far, there are no attempts to use algorithms of meta-heuristic technique to provide resources in multi-tier architecture, our Improvement methodology will deals with the resource provisioning in multi-tier architecture in cloud computing by using of meta-heuristic technique such as SA, PSO, PSO-SA algorithm. Finally, compare our allocation techniques with other allocation techniques to evaluate their relative effectiveness.

Acknowledgements

This work was supported by Faculty of Computer and Information Sciences, Ain Shams University

References

[1] Eun-Kyu Byuna, Yang-Suk Keeb, Jin-Soo Kimc,Seungryoul Maenga, "Cost optimized provisioning of elastic resources for application workflows," *Future Generation Computer Systems,vol. 27,* pp. 1011–1026, 2011.

[2] Guruprasad, Bhavani B H and H S, "Resource Provisioning Techniques in Cloud Computing Environment: A Survey," International Journal of Research in Computer and Communication Technology, Vol 3, no. Issue3, pp. 395-400, March-2014.

[3] S. J. Hamid Reza Qavami, "A Survey On Resource Provisioning In Cloud Computing," International Journal of Research in Computer and Communication Technology, Vol.2, no. Issue.2, pp. 160-167, February 2014.

[4] Guruprasad, Bhavani B H and H S, "Resource Provisioning Techniques in Cloud Computing Environment: A Survey," . Vol 3, no. Issue3, pp. 395-401, March-2014.

[5] Agrawal, Tarun goyal & Aakanksha, "Host Scheduling Algorithm Using Genetic Algorithm In Cloud Computing Environment," international journal of research in engineering & technology (ijret), Vol. 1, no. Issue 1, pp. 7-12, june 2013.

[6] Othman, Monir Abdullah and Mohamed, "Simulated Annealing Approach To Cost-Based Multi- Quality Of Service Job Scheduling In Cloud Computing Enviroment," American Journal of Applied Sciences,vol.11, pp. 872-877, 2014.

[7] Huo, Shaobin Zhan and Hongying, "Improved PSO-based Task Scheduling Algorithm in Cloud Computing," Journal of Information & Computational Science, vol.9, pp. 3821–3829, 2012.

[8] Talwinder Kaur, Seema Pahwa. s.l., "An Upgraded Algorithm of Resource Scheduling using PSO and SA in Cloud Computing.," International Journal of Computer Applications, vol. 74, pp. 28-32, July 2013.

[9] Xiaotang Wen, Minghe Huang, Jianhua Shi. s.l., "Study on Resources Scheduling Based on ACO Algorithm and PSO Algorithm in Cloud Computing.," IEEE, pp. 219-222, 2012.

[10] Dong Huang, Bingsheng He and Chunyan Miao, "A Survey of Resource Management in Multi-Tier Web Applications," IEEE, vol. 16, no. 3, pp. 1574 - 1590, 29 January 2014.

[11] Pedram., Hadi Goudarzi and Massoud, "Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems," Cloud Computing (CLOUD), IEEE International Conference on, pp. 324 - 331. 2011.

[12] Bhuvan Urgaonkar, Prashant Shenoy, Abhishek Chandray, and Pawan Goyal, "Agile, Dynamic Provisioning of Multitier".

[13] Chandra, W. Gong, and P. Shenoy, "Dynamic Resource Allocation for Shared Data Centers Using Online Measurements," Proceedings of the 11th International Conference on Quality of Service, vol. 2707, pp. 381-398, 2003.

[14] Heng WU ,Wenbo ZHANG, Jianhua ZHANG, JunWEI, Tao HUANG, "A benefit-aware on-demand provisioning approach for multi-tier application in cloud computing," Frontiers of Computer Science, pp. 459–474, 2013.