IJCSC 0973-7391 Volume 7 • Number 2 March 2016 - Sept 2016 pp. 12-15 available online at www.csjournalss.com

Resource Provisioning Strategies in Cloud: A Survey

Shivani Vasoya[1], Lata Gadhavi[2,] Jitendra Bhatia[3], Madhuri Bhavsar[4]
[1]M.Tech Student, [2,3]Assistant Professor, [4]Professor
Department of Computer Science and Engineering, Nirma University, Ahmedabad,

Abstract: In the recent field of Information and Technology, the cloud computing has been introduced as an most emerging technology which provides on-demand self-service and broader access to the network by managing the provision and the release of shared pool of computational resource in a manner that reduce the resource utilization cost. Resources needed to maintain and implement the management system and to address customer requirement including personal, work environment, information, and financial resources. Verification of resource needs for particular activities is integrated with the process of defining and initiating the activity. Resource provisioning means to provide resources as per the requirement of the application with keeping in mind the Quality-of-service defined in the SLA. It is a very challenging task to provide resources in a manner to achieve the QoS as defined in the SLA. CSP should ensure that all the applications running on the cloud get the resources as per their requirements. In this paper, we have reviewed the various resource provisioning techniques proposed by researchers considering different parameters.
Key Terms: Cloud Computing, Resource Provisioning.

I. INTRODUCTION

Cloud computing is the fastest growing technology in IT industry. A very large amount of IT enterprises have started migrating their application development work to the cloud environment as it provides broad network access, resource pooling, elasticity, on-demand services. As users get the required resources from cloud, they don't have to buy hardware, software licenses / install software, drivers etc and the time and the money can be saved. Cloud service providers make the job of an application developer easy by providing necessary services to its users. Therefore, cloud users can make use of the existing services provided by the cloud service providers rather creating new services which is difficult and time taking process. The cloud provides various services like Infrastructure-as-a-Service and Platform-as-a-Service which reduces the cost for the establishment of necessary infrastructure to develop an application. In cloud, users can acquire the resources provided by the cloud service provider and get their applications developed and deployed on cloud. The cloud service providers can provide the resources to the cloud user as per the static requirements or the workload of the user or by considering various parameters. To decide the proper number of resources needed for the implementation of user task so that one can reduce the commercial cost from user's point of view and increase resource usage from service provider's point of view is challengable task.

Resource provision refers to determination, establishment and run time management of software and hardware resources. It considers SLA(Service Level Agreement) to provide resources to the cloud users. Specific characteristics of the service (scope, quality, responsibilities) are defined in SLA. This is a primary agreement between cloud user and cloud service provider. Static/ dynamic assignment of resources must be done as per the requirements of an application. So the final objective of the user is to decrease the resource usage cost and from service provider's perspective to increase the profit. To accomplish the goal user must request to service providers to provide resources either statically or dynamically. According to the user request, service provider come to know that how many resources are required to fulfill the request. Service provider allocates resources such as QoS parameters like waiting time, availability, security must be reached without violating a SLA.

This paper aims to contemplate the resource provisioning variants. Section II describes resource provisioning types. Section III lists various resource provisioning techniques and their performance benefits. Section VI describes conclusion and future directions.

II. RESOURCE PROVISIONING TYPES

As per the requirement of the application, resource provisioning techniques can be categorized into two types:

1. Static Resource Provisioning

If any application is having expected and fixed demands, one can use static provisioning effectively. User has to mention fixed requirements so that service provider will be able to provide the same while starting the application.

2. Dynamic Resource Provisioning

If any application is having unexpected and varying demands, dynamic provisioning is used whereby VM may be migrated on-the-fly to new VM. In this case, service provider assigns extra VMs to the users if needed and removes them.

DOI: 10.090592/IJCSC.2016.103

In resource provisioning, various parameters are considered such as response time, workload, reduced SLA violation etc. The resource provisioning algorithm designed must take minimum time to respond while executing the task. Also, it must consider the workload of the each VM and able to reduce SLA violation.

For effective use of the cloud resources, resource provisioning techniques are to be used. Many researchers have proposed methods for resource provisioning with respect to various parameters like resource provision based on deadline provided, resource provision based on cost analysis or budget constraints, resource provision based on Service Level Agreements[11], etc.

III. Resource Provisioning Variants

1. Resource Provisioning based on Execution Cost:
In [12], author had proposed an algorithm for resource provisioning and scheduling on Infrastructure as a Service (IaaS) clouds. The strategy contained two methods meta-heuristic optimization technique and particle swarm optimization (PSO) which tries to minimize the execution cost by keeping in mind the deadline constraints. Also, heuristics are considered to generate schedules with lower implementation cost. The whole technique is divided in two phases. In the first phase, all the resources that will be used to run the task are selected for resource provisioning, while in second phase, each task is associated with the computing resources such that the requirements of QoS are met.

This strategy works well for smaller size of work-flows but for larger size of work-flows, the scheduler generates lower make spans and higher costs to to meet the deadline.

2. Resource Provisioning based on Budget Constraint:
In the Resource provisioning with budget constraints for adaptive applications, the provisioning of resources takes place by considering the budget which is predefined for each application in cloud[13]. To allocate the resources,following parameters are used.

- Resource budget,
- Time constraint
- QoS parameters i.e. response time, CPU utilization.

CPU and memory usage take in consideration for calculating cost. Each VM has a predefined share in the host system's CPU cycle. There are two ways to to run the jobs: **Capped mode-** The VM running on this mode can execute their jobs within their share only. In a case, if an application requires more amount of share than predefined from the CPU cycle, they can't execute their jobs even if there is any idle CPU cycle. **Non-capped mode-** The VM running on this mode can execute their jobs if they need extra amount of share in CPU cycle. While allocating the CPU cycles to the VM, the controller first checks for the cost and compare it with the budget of VM user. If it crosses the maximum budget limit then it will allow only those resources which are best fit to the budget. Otherwise it allocates all the requested resources.

As shown in table 1, resource cost can be decreased significantly by using this approach. Also, there is a provision to perform parameter adaption to encounter different time and budget requirements. This approach seems unrealistic because exact amount of CPU and memory usage cannot be known in advance.

Execution Time (Hour)	Resource Cost			
	Linear Pricing Model		Expo. Pricing Model	
	General model	Proposed model	General model	Proposed model
1	450	390	550	400
3	600	410	750	395
5	800	400	1150	406

Table 1: Resource cost of conventional and proposed model for Linear and Non-linear pricing model [2]

3. Automated Resource Provisioning
In [8], author proposed a method for automated provision of cloud services where the services are a set of components that could be mapped and remapped to the resources which are created dynamically. This approach has described an architecture for deploying and managing the virtual infrastructure and deployed services in cloud. Also, it consists of four layers named infrastructure layer, service orchestrator layer, abstraction layer and design layer respectively. In this architecture, the service provision has been done based on a template which incorporates the requirements of services and options and also behavior of a group of resources. To integrate the deployment and to reconfigure the behavior for services where the description of logical components is given for scaling the components and for changing the configuration as per the need.

As shown in table 2, time needed to provide software component is significantly small as compare to time needed to create infrastructure. This model uses parallel deployment method so that large cloud deployment time can be decreased. This approach is lacking for the automatic computation features like fault-tolerance, management of SLA and assurance for QoS.

No. of VMs	Execution time(s)	
	Infrastructure Time	Deployment Time
1	60	35
2	68	32
3	75	40
4	74	41
5	75	33
6	78	34

Table 2: Execution time for VM deployment and infrastructure[3]

4. Resource Provisioning based on Heuristics

In [16], Static algorithm had been proposed to schedule a work-flow of an instance based on deadline constraints in IaaS cloud. This algorithm considers partial critical paths for each work-flows and also some parameters like heterogeneity of VMs, pay-per-use and time interval pricing model. They calculate latest finish time of each task based on maximum time limit and availability of an instance. Based on heuristic, they try to minimize the execution cost of the task. They are not able to use complete work-flow structure because they have used optimization at task level. They can generate more optimal solution by developing optimization technique at global level.

5. Resource Provisioning Based on Predictions:

The researchers have developed algorithms that handles the resource provisioning based on some predictions. In [14], the management of resource provisioning is done based on predictions where they made use of machine learning algorithms and sliding window techniques of time-series analysis. Machine learning algorithms like Linear Regression and Neural Network are used to handle the future demands. Prediction is better in Neural Network approach than Linear Regression.

This paper also considers sliding window technique for prediction. Prediction is more accurate by using sliding window as it gives more particular knowledge about pattern in neighborhood of the estimated data point. But, this algorithm works very slow as it takes time to train Neural Network. Another drawback of this algorithm is that the process of training the Neural Network must be repeated at specific time interval and that should be derived from the usage behavior of resources in cloud.

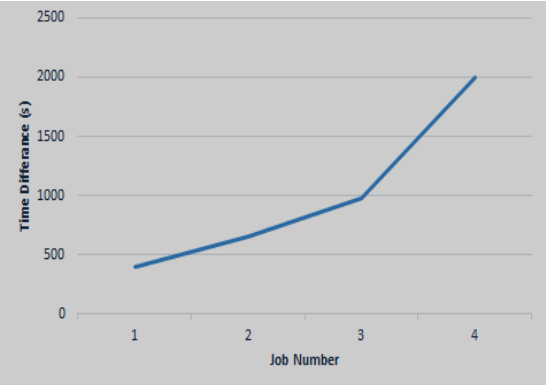
6. Resource Provisioning based on Priority

The proposed algorithm in the paper [18] aims to allocate the resources dynamically to the incoming requests. The VMs are allocated to the jobs as per the priority given to each jobs. The algorithm also considers the various parameters like predicted execution time of each job, priority given to each job, lease type associated to each job. In this proposed algorithm, first the cloud-broker has to pre-configure and store all the VM images in cloud which provides a platform to execute the user's jobs. All the submitted user's jobs are queued. A system-level scheduler, pre-configured by the service provider, runs at a particular system in cloud, which gets executed periodically. The scheduler carries out the following tasks each time when it gets executed: prediction of incoming work-loads in future, provision of VMs in advance, allocation of jobs to VM. releasing idle Vms and if the predicted execution time of newly arrived job is high, start the necessary number of VMs. This proposed approach attempts to develop a priority-based preemption policy which tries to improve resource utilization. VMs get allocated to each job by considering their priority and the lease type associated with each job. The following are the lease types associated with the jobs. **Cancellable :** These kind of requests can be scheduled to execute at any time once after they arrive and need not to be resumed later. These leases do not guarantee the deadline. **Suspendable :** These kind of leases guarantee the execution but not a deadline. Jobs associated this kind of lease can be suspended any time but should be resumed later. It is possible to schedule these jobs any time after they are ready to execute. **Non-Preemptable :** The jobs associated this type of lease should not be preempted.

When a high priority job request arrives and no VM is free to allocate to newly arrived job, then a low priority job, getting executed into a VM has to preempt it's resources and allow newly arrived high priority job to use the resources, preempted by the low priority jobs. When a new jobs arrives for execution when a VM is not free to execute this job, the algorithm finds two low priority jobs with their lease types. The job having the lease type as cancelled gets preempted. When two low priority jobs arrive, the job having a lease type as canceled and whose execution time is also low gets the first priority for execution. When scheduler finds two low priority jobs with suspendable lease type, the job having executed a minimum amount of work gets the high priority for preemption.

The disadvantage of this method is that the difference between predicted time and actual time is very large(as shown in fig. 1) so it is not feasible to apply in real time environment.

Fig. 1 Difference between actual and predicted time per job [19]



IV. Conclusion and Future Directions

This survey presents various resource provisioning variants with their merits and demerits. It shows the performance variance while considering different parameters and different methods. The efficient dynamic resource provisioning or automated resource provisioning is one of the elementary challenge in cloud, because a tradeoff between professional SLA and different constraint like max resource utilization, cost etc. In future, we will proposed a method for automated resource provisioning. It has adaptive nature so that resources can be fully utilized.

References

[1] A.S.Weber, "Cloud computing in education," *in Ubiquitous and mobile learning in the digital age,* pp.19-36, Springer, 2013

[2] N. Sultan, "Cloud computing for education: A new dawn?" *International Journal of Information Management,* vol. 30, no. 2, pp. 109-116, 2010

[3] M. M. Alabbadi, "Cloud Computing for education and learning: Education and learning as a Service (class)" *in Interactive Collaborative Learning (ICL), 2011 14th International Conference on,* pp. 589-594, IEEE, 2011

[4] Q. Chen and Q. Deng, "Cloud computing and its key techniques" *Journal of Computer Applications,* Vol. 29, No. 9, p. 2565, 2009.

[5] D.G. Chandra and M.. D. Borah, "Cost benefit analysis of cloud computing in education" *Computing, Communication and Applications (ICCCA), International Conference on,* pp. 1-6, IEEE, 2012.

[6] M. Mircea and A.I. Andreescu, "Using Cloud Computing in higher education: A strategy to improve agility in the current financial crisis," *Communications of the IBIMA, 2011.*

[7] Y. Hu, J. Wong, G. Iszlai, M. Litoiu, "Resource provisioning for cloud computing," in *proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research,* pp. 101-111, IBM Corp., 2009.

[8] L. Zhang, Z. Li, C. Wu, "Dynamic resource provisioning in cloud computing: A randomized auction approach," in *INFOCOM, 2014 Proceedings IEEE,* pp. 433–441, IEEE, 2014.

[9] P. Jamshidi, A. Ahmad, C. Pahl, "Autonomic resource provisioning for cloud based software," *in Proceeding of 9th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, pp. 95-104, ACM, 2014.

[10] J. Hu, J. Gu, G. Sun, T. Zhao, "A Scheduling strategy on load balancing of virtual machine resources in cloud computing environment," *in Parallel Architectures, Algorithms and Programming (PAAP), 2010 Third International Symposium on*, pp. 89–96, IEEE, 2010.

[11] A. Kertesz, G. Kecskemeti, I. Brandic, "An sla-based resource virtualization approach for on-demand service provision," *in Proceedings of the 3rd international workshop on Virtualization technologies in distributed computing*, pp. 27-34, ACM, 2009.

[12] M. A. Rodriguez, R. Buyya, "Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds," *Cloud computing, IEEE Transactions on ,* Vol. 2, no. 2, pp. 222-235, 2014.

[13] Q. Zhu, G. Agrawal, "Resource provisioning with budget constraints for adaptive applications in cloud envirnments" *in Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pp. 304-307, ACM, 2010.

[14] J. Kirschnick, J. M. A. Calero, L. Wilcock, N. Edwards, "Toward an architecture for the automated provisionig of cloud services," *Communications Magazine*, IEEE, Vol. 48, no. 12, pp. 124-131, 2010.

[15] Q. Zhang, Q. Zhu, R. Boutaba, "Dynamic resource allocation for spot markets in cloud computing environments," *in Utility and Cloud Computing (UCC). 2011 Fourth IEEE International Conference on*, pp. 178–185, IEEE, 2011.

[16] S. Islam, J. Keung, K. Lee, A. Liu, "Empirical prediction models for adaptive resource provisioning in the cloud," *Future Generation Computer Systems,* Vol. 28, no. 1, pp. 155–162, 2012.

[17] L. Gadhavi, D. Korat, M. Bhavsar, "Proposed cloud architecture for automated and reliable service provisioning of engineering educational domain" in *Engineering (NuiCONE),2013 Nirma University International Conference on*, pp. 1-7, IEEE, 2013.

[18] Saraswathi AT a, Kalaashri.Y.RA b, Dr.S.Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing" *Elsevier B.V. 2015, Procedia Computer Science 47 (2015) 30 – 36*

[19] Jitendra Bhatia , Harshal Trivedi, Vishrut Majmudar, Tirth Patel "HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud " *2012 International Symposium on Cloud and Services Computing*

[20] Jitendra Bhatia, " A Dynamic Model for Load Balancing in Cloud Infrastructure" *Nirma University Journal of Engineering and Technology Vol. 4,No. 1, Jan-Jun 2015*

[21] Jitendra Bhatia, Malaram Kumhar, "Perspective Study on Load Balancing Paradigms in Cloud Computing" IJCSC Vol 6, Issue-1, Sep-Mar 2015 pp.112-120