International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

# Review on: Resource Provisioning in Cloud Computing Environment

**Sagar Girase[1], Rahul Samant[2], Mayank Sohani[3], Suraj Patil[4]**

[1]M.Tech. Department of Computer Engineering
MPSTME SVKM's NMIMS Shirpur, Maharashtra, India

[2]Associate Professor, Head Department of Information Technology
MPSTME SVKM's NMIMS Shirpur, Maharashtra, India

[3, 4] Assistant Professor, Department of Computer Engineering
MPSTME SVKM's NMIMS Shirpur, Maharashtra, India

**Abstract:** ***Today, Cloud computing become an emerging technology which will has a significant impact on IT Infrastructure. Still, Cloud computing is infancy. In the current cloud computing environment there is numerous of application, consist of millions of module, these application serve from large quantity of users and the user request becomes dynamic. So there must be provision that all resources are dynamically made available to satisfy the needs of requesting users. The resource provisioning was done by considering Service Level Agreements (SLA) and with the help of parallel processing using different types of scheduling heuristic. In this paper we realize such various policies for resource provisioning and issues related to them in current cloud computing environment.***

**Keywords:** Cloud computing; Scheduling, Service Level Agreements (SLA), Virtualization, Virtual Machines (VM).

## 1. Introduction

Cloud computing is one of the most promising technologies in the modern world having a broad array of web-based services aimed at allowing users to obtain a wide range of functional capabilities on a 'pay-as-you-use' basis. Cloud computing is still in its early stages, but the public sector it offers various benefits such as, Cost savings, Highly automated, Flexibility, More Mobility, Increase storage, Business agility, availability of resources etc.

### 1.1. Cloud Deployment Models

The entire concept of cloud computing is divided into three forms of cloud. All three have significant characteristics; however their choice depends on the personal requirements of business environment. These include Private Cloud, Public Cloud and Hybrid Cloud which is exceptionally flexible.

- Private Cloud: A Private cloud not promotes shared environment. This means private cloud is beneficial for those organizations that do not want to share their confidential data with any third party.
- Public Cloud: In this type of cloud form, data stored is in cloud server, which is located at a distant place elsewhere. It enables users to share and access data from anywhere and at any point of time. This means public cloud promotes shared environment. Although, a bit risky in terms of data security as business operations are done through Internet, but offers highly scalable environment.
- Hybrid Cloud: A Hybrid cloud is a combine of both and gives users or business entities advantage of both the cloud environments. Suppose, a business enterprise wants to share its services and products with its clients across the globe, but at the same time wants to hide the confidential information from them, Hybrid cloud architecture would suit best for such types of businesses.

### 1.2. Cloud Service Models

Cloud services are classified into three models, it includes:

- Software-as-a-Service (SaaS) - In this model, providers offers a complete application to the client's for use on demand. EMC Mozy is an example of SaaS.
- Platform-as-a-Service (PaaS) – This model has capability provided to the client's is to deploy applications, supported by the providers. It is also used as an application development environment. Google App Engine and Microsoft windows Azure Platform are examples of PaaS.
- Infrastructure-as-a-Service (IaaS) - This model has capability to provides scalable computing, storage, network and other fundamental computing resources where the client's can able to deploy and run their own software. Amazon Elastic Compute Cloud (Amazon EC2) is an example of IaaS.

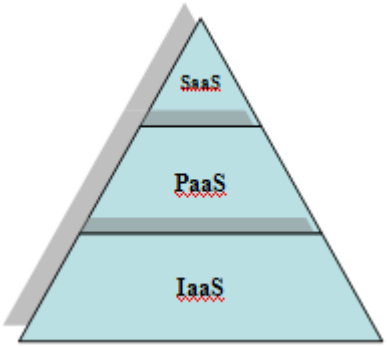


**Figure 1:** Cloud Service Model

Cloud computing has enabled IT organizations and individual to gain benefits, such as automated and rapid resource provisioning, flexibility, high availability and faster

time to market at a reduced total cost of ownership. Although there are concerns and challenges, the benefits of cloud computing are compelling enough to adopt it.

### 1.3. Challenges of Cloud

Following are the major challenges needs to overcome, it includes:
Dynamic Scalability: The most of the application is ready to scale up and scale down the compute nodes dynamically as per the response time of the user's queries. The resource allocation and task scheduling delay is the one of the major factor which leads to the need of dynamic and effective load management system [15].

Interoperability and Portability- When user wants to move their services from one cloud to another is a lack of compatible, time-consuming and labour-intensive because currently there is no standard way exists for interfacing with a different cloud and each provider exposes its own APIs [16]. Cloud Security and Privacy - In cloud computing, Security is one of the most critical issues because people are not worried about the location of data they store in the cloud. Some CSPs may have less transparency than others about their policy of information security.

Reliable Service Allocation - Several reliability problems that arise when allocating applications to processing resources in a Cloud computing platform [17]. Compliance Audit Requirements – Now a day's cloud computing services can challenge various compliance audit requirements currently in place. Data locations, cloud computing security policy transparency are all challenging issues in compliance auditing efforts [15]. Automated service provisioning - Resource provisioning decisions must made online for satisfying service level objectives while minimizing operational costs.

Scalable Querying and Secure Access: In both grid and cloud environment are two open problems such as scalable provenance querying and secure access of provenance information. Multi-tenancy: Many CSPs have multi-tenant applications in which multiple independent application are serviced using a single set of resources. When a number of applications executing by the single compute node, then the amount of bandwidth allocated to each application reduces which may lead to performance degradation.

Reliability and fault-tolerance: fault tolerance is required which help to improve the reliability of develop system. Cloud computing leads an opportunity to offering testing tools for testing the application against compute failures in clouds [18]. Power: An energy aware resource management offers many type of services to meet the needs of users because it require enormous amount of power for that in cloud computing environment [15].

Data location - cloud computing technology allows cloud servers to reside anywhere, thus the enterprise may not know the physical location of the server used to store and process their data and applications. Although from the technology point of view, location is least relevant, this has become a critical issue for data governance requirements. It is essential to understand that many Cloud Service Providers (CSPs) can also specifically define where data is to be located.

Disaster recovery - It may be possible that the data may be commingled and scattered around multiple servers and geographical areas for a specific point of time cannot be identified. In the cloud computing model, the primary CSP may outsource capabilities to third parties, who may also outsource the recovery process. This will become more complex when the primary CSP does not ultimately hold the data.

One of the important requirements for a Cloud computing environment is providing reliable QoS. It can be defined in terms of Service Level Agreements (SLA) that describes such characteristics as minimal throughput, maximal response time delivered by the deployed system. Although modern virtualization technologies can ensure performance isolation between VMs sharing the same physical computing node, due to aggressive consolidation and variability of the workload some VMs may not get the required amount of resource when requested. This leads to performance loss in terms of increased waiting time, time outs or failures in the worst case. Therefore, Cloud providers have to deal with resources provisioning for user request, while meeting QoS requirements. The rest of the paper is organized as follows: Section 2 provides a literature survey, Section 3 Improvement methodology, Section 4 concludes.

## 2. Literature Survey

Efficient scheduling problems and resource management are related to the efficiency of the whole cloud computing facilities. The scheduling algorithms in distributed systems usually have the goals of spreading the load on processors and maximizing their utilization while minimizing the total task execution time. Several heuristic algorithms have been introduced in scheduling.

Quiroz et al. [1] introduced a Decentralized; robust Online Clustering mechanism to drive workload (i.e. VM) provisioning on enterprise grids and clouds. In order to deal with inaccurate resource request that leads to over-provisioning provided by application job requests, their mechanism has demonstrated a model-based approach for estimating the application service time given its provisioning. They also presented a quadratic response surface model (QRSM), which was found to best capture the behaviour of the workload for application-specific. It is used to model the application in the cloud computing environment dynamically.

In [3] Van et al. proposed an autonomic resource manager to address the problem of autonomic virtual resource management for hosting service platforms while optimizing a global utility function that integrates the operating cost of the platform provider and the application level SLAs. They also used utility function with constraint programming approach to achieved self optimization by defining business level SLAs of the application and the resource exploitation cost of the hosting provider as constraints. The proposed system attempts to maximize the performance of the hosted applications with an optimal operating cost for the hosting provider.

Parekh et al. [4] addressed the problem of building an effective external controller for automated adaptive scaling of applications deployed in the cloud. They recommended the Proportional Thresholding approach which dynamically adjusts the target range i.e. high and low thresholds based on the number of accumulated virtual machine instances. Thus the relative effect of allocating resources becomes finer as the number of accrued resources increases; eventually resulting in being adaptive and more resource efficient.

Silva et al. [5] proposed a heuristic for optimizing the number of machines for processing an analytical job with predefined number of independent tasks so that maximum speedup can be achieved within a limited budget. However, the traffic of a web application is dynamic and random in nature; hence predicting the optimal number of machines for the fulfillment of the application level SLAs in real time.

Parekh et al. and Silva et al. [4], [5] presented different new algorithms for adaptive scaling, there is a need for an effective prediction scheme for prediction of the future application-level SLAs in order to minimize the system overheads. In this regard, Caron et al. [2] initiated the groundwork for a new approach to workload prediction algorithm based on past usage pattern. Since dynamic allocation and de-allocation of virtual machine instances include some overheads such as, setup time, performance improvement and responsiveness could be achieved if the system can predict and scale in advance to adapt with the changing workload. Based on similar characteristics of web-traffic, they proposed a pattern matching algorithm that is used to identify closest resembling patterns similar to the last usage measure of the present usage pattern in a set of past usage traces of the cloud client. The resulting closest patterns are then interpolated by using a weighted interpolation to forecast approximate future values that are going to follow the present pattern. This prediction finally aids in making dynamic scaling decisions in real time. However, this approach is unable to adapt with any new pattern that might appear as a result of the dynamic nature of the web traffic.

In [6], Author proposed the design and implementation of an automated system that uses virtualization technology. Here they discussed the process of allocating data center resources dynamically based on application demands and support green computing by optimizing the number of servers in use. In this, author introduced the concept of “skewness”. Author introduced concept is used to measure the unevenness in the multidimensional resource utilization of a server and also for prevent overload in the system effectively while saving energy used they proposed a set of heuristics.

In [7], Author has proposed an adaptive resource allocation algorithm for the cloud system with preempt able tasks. There are two major contribution of this work done by author. First, author has present a resource allocation mechanism in cloud systems which enables preempt able task scheduling suitable for the autonomic feature within clouds and the diversity feature of VMs. Second, they have proposed two adaptive algorithms for resource allocation and task scheduling in IaaS cloud computing. These algorithms adjust the resource allocation adaptively based on the updated of the actual task executions.

SLA-based Resource allocation algorithms in [8] for SaaS providers, who want to minimize infrastructure cost and SLA violations. In this, author proposed algorithms are designed in a way to ensure that SaaS providers are able to manage the dynamic change of customers, mapping customer requests to infrastructure level parameters and handling heterogeneity of Virtual Machines. They take into account the customers’ Quality of Service parameters such as response time, and infrastructure level parameters such as service initiation time. Proposed algorithms minimize the SaaS provider’s cost and the number of SLA violations in a dynamic resource sharing in cloud environment.

In [9], Author proposed a VM-based architecture for adaptive management of virtualized resources in cloud computing environments. VM-based architecture provides strong isolation between applications which simultaneously run in the virtual resource pool and allows the dynamic allocation of resources to applications and applications to grow and shrink based on resource demand to achieve SLOs. But, how to allocate resources to every application on-demand and in response to time-varying workloads is the major problem in VM-based architecture.

An approach for dynamic autonomous resource management in computing clouds [10]. The main contribution of this work is two-fold. First, they adopt a distributed architecture where resource management is decomposed into independent tasks, each of which is performed by Autonomous Node Agents that are tightly coupled with the physical machines in a data center. Second, the Autonomous Node Agents carry out configurations in parallel through Multiple Criteria Decision Analysis using the PROMETHEE method. Simulation results show that the proposed approach is promising in terms of scalability, feasibility and flexibility.

A decentralized architecture presented in [11] of the energy aware resource management system for Cloud data center. Author has defined the problem of minimizing the energy consumption while meeting QoS requirements and stated the requirements for VM allocation policies. Moreover, they proposed three stages of continuous optimization of VM placement and presented heuristics for a simplified version of the first stage.

Architecture for the dynamic scaling of web applications [12], based on thresholds in a virtualized Cloud Computing environment. In this work, front-end load-balancer is used in scaling approach for balancing user requests to web applications. They were introduced; scaling algorithm based on threshold number of active sessions for dynamic provisioning of virtual machine resources. Dynamically allocate and rapidly provision of resource to users is to be discussed based on-demand capability of the cloud. The existing works have been shows that maximum work is done in the investigation of resource provisioning from the cloud service provider perspective.

## 3. Improvement Methodology

In Cloud Computing, earlier most of the work was done by different authors for static resource allocation as well as dynamic resource allocation. But here, Requesting VM is

required to select one appropriate task for execution which is best suitable for available resources. Hence, Improvement methodology will deals with the dynamic resource provisioning with time constraint in Cloud Computing environment which will considering On-line mode scheduling heuristic, preempt able tasks execution and multiple SLA parameters such as memory, required CPU time, network bandwidth and waiting queue to provide reliable QoS.

## 4. Conclusion

The review shows that the challenges in here is for applications hosted in the cloud need to be elastic in order to achieve economy of scale while preserving the application-specific Service Level Agreements (SLA) with time constraints such as, response time, and throughput etc. The usage prediction and dynamic provisioning of resources is one of the fundamental research challenges, because a balanced trade-off between the business-specific SLAs and other constraints such as maximum utilization of resources, cost effectiveness, etc. will need to be achieve. Hence, Improvement heuristic will provide efficient resource provisioning to the multiple cloud users.

## 5. Future Scope

In future, we will present a model which deals with the efficient allocation of the resources and their implementation as well as analyze the performance level of cloud system with time constraints. The future work can be a sequence of research on other parameters which implies performance of a cloud system.

## 6. Acknowledgment

We would like to thanks to all the experts from MPSTME, SVKM's NMIMS Shirpur-Campus for his helps and advices.

## References

[1] Andres Quiroz, Hyunjoo Kim, Manish Parashar, Nathan Gnanasambandam, Naveen Sharma, "Towards Autonomic Workload Provisioning for Enterprise Grids and Clouds", presented at the 10th IEEE/ACM International Conference on Grid Computing, 2009.
[2] Eddy Caron, Frederic Desprez and Adrian Muresan, "Forecasting for Cloud computing on-demand resources based on pattern matching," 2010.
[3] Hien Nguyen Van, Frederic Dang Tran, Jean-Marc Menaud, "Autonomic virtual resource management for service hosting platforms," presented at the ICSE'09 Workshop, Vancouver, Canada, 2009.
[4] Harold C. Lim, Shivnath Babu, Jeffrey S. Chase, Sujay S. Parekh, "Automated Control in Cloud Computing: Challenges and Opportunities".
[5] L Joao Nuno Silva, Luis Veiga, Paulo Ferreira, "Heuristic for resources allocation on utility computing infrastructures".
[6] Zhen Xiao, Weijia Song, and Qi Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment", IEEE Transaction on Parallel and Distributed system", VOL. 24, NO. 6, JUNE 2013.
[7] Jiayin Li, Meikang Qiu, Yu Chen., "Adaptive Resource Allocation for Preemptable Jobs in Cloud Systems", IEEE 10thInternational Conference on Intelligent Systems Design and Applications, 2010.
[8] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, "SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments", 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2011.
[9] Qiang Li, Qinfen Hao, Limin Xiao, Zhoujun Li. "Adaptive Management of Virtualized Resources in Cloud Computing Using Feedback Control" The 1st International Conference on Information Science and Engineering (ICISE2009) IEEE.
[10] Ya gız, Onat Yazır, Chris Matthews, Roozbeh Farahbod, Stephen Neville, Adel Guitouni, Sudhakar Ganti and Yvonne Coady. "Dynamic Resource Allocation in Computing Clouds using distributed Multiple Criteria Decision Analysis" IEEE 3rd International Conference on Cloud Computing, 2010.
[11] Anton Beloglazov and Rajkumar Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers", 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. 2010.
[12] Trieu C. Chieu, Ajay Mohindra, Alexei A. Karve and Alla Segal. "Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment", IEEE International Conference on e-Business Engineering, 2009.
[13] T.R. Gopalakrishnan Nair, Vaidehi M. "Efficient Resource Arbitration and Allocation Strategies in cloud computing through virtualization" Proceedings of IEEE CCIS-2011.
[14] Daniel Warneke and Odej Kao. "Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the cloud", IEEE Transaction on Parallel and Distributed Systems, January 2011
[15] Naidila Sadashiv, S. M Dilip Kumar, "Cluster, Grid and Cloud Computing: A Detailed Comparison", The 6th International Conference on Computer Science & Education (ICCSE 2011) August 3-5, 2011. SuperStar Virgo, Singapore, pp. 477- 482.
[16] Zehua Zhang and Xuejie Zhang, "Realization of Open Cloud Computing Federation Based on Mobile Agent", Intelligent Computing and Intelligent Systems, ICIS 2009. IEEE International Conference. Vol 3.
[17] Olivier Beaumont, Lionel Eyraud-Dubois, Hubert Larcheveque, "Reliable Service Allocation in Clouds", IEEE 27th International Symposium on Parallel & Distributed Processing, 2013.
[18] Jerry Gao, Xiaoying Bai and Wei-Tek Tsai, "Cloud Testing- Issues, challenges, Needs and Practive", Software engineering: an international journal (SEIJ), Vol. 1, No. 1, September 2011.

## Author Profile

**Mr. Sagar Girase** received the bachelor's degree from North Maharashtra University, Jalgaon in 2012. Currently, He is a Student of Master of Technology from Department of Computer Engineering at of

Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, (Maharashtra) of SVKM's NMIMS (Deemed to be University). His current research focuses on resource scheduling problems in cloud systems.

**Prof. Rahul Samant** is an Associate Professor, Head Dept. of Information Technology at Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, Dist. Dhule (Maharashtra) of SVKM's NMIMS (Deemed to be University). He received the bachelor's degree and Master's degree from Mumbai University in 1997 and 2005. His research interests include cloud computing and data warehouse & mining.

**Prof. Mayank Sohani** is an Assistant Professor in the Department of Computer Engineering at of Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, Dist. Dhule (Maharashtra) of SVKM's NMIMS (Deemed to be University). He received Degree of MCA from Rajiv Gandhi Technical University, Bhopal, M.P. in 2005 and Master of Technology from DAVV University, Indore, M. P. in 2011. His research interests include cloud computing and software testing.

**Prof. Suraj Patil** is an Assistant Professor in the Department of Computer Engineering at of Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, Dist. Dhule (Maharashtra) of SVKM's NMIMS (Deemed to be University). He received the bachelor's degree from Bharatiy vidyapeath, Shivaji University, Kolhapur 2006 and Master's degree from North Maharashtra University, Jalgaon in 2013. His research interests include cloud computing and data mining.