# Analysis of Resource Pooling and Resource Allocation Schemes in Cloud Computing

Dr. Amit Chaturvedi[#1], Aaqib Rashid[*2]

*#Assistant Prof., MCA Deptt., Govt. Engineering College, Ajmer*

**ABSTRACT:** *Cloud servers always do resource pooling for providing to their customers. In multi-tenant cloud environment, multiple tenants may demand for the same resource or multiple resources may be occupied by the same tenanct for long time and there may be lack of resource due to this reason. So, efficient Resource Allocation Schemes are required to manage the resources.*
*Cloud computing basically is resource pooling and allocation or sharing technology of pooled resources. In this paper, we are analysing various resource pooling and resource allocation schemes proposed by researchers for cloud computing.*

**Keywords :** *Resource scaling, cloud computing, virtual machine, multi-tenant.*

## I. INTRODUCTION

Resources on rental bases in cloud computing is one of the influential and cost effective manner. Multi-tenant cloud environment needs to take advanced and reliable mode on Resource scaling so that the important objective could take on that advanced level. The Cloud Computing creates new opportunities to align IT resources and business goals. In order to consume and use Cloud Computing resources to their best advantage, it is essential for business owners, IT managers, startup founders, and developers to understand the options that Cloud Computing provides. The complete range of Cloud and traditional resources will assure that our online IT resources can expand to meet the business needs even in high growth situations. Cloud computing resources refer to —applications, platforms, raw computing power and storage, and managed service detection of viruses delivered Over the Internet.

Analysis of Resource Scaling Infrastructure as Service is the main objective in resource scaling in cloud computing systems. Multi-tenant environment applications use virtualized technologies to encapsulate and segregate application performance by using separate virtual machines (VM).Virtualization technologies evolved to help IT organizations and improve the efficiency of their hardware resources by partitioning hardware to provide simultaneous support to multiple applications and their corresponding software stacks (operating system, database, application server, etc.).

However, if resource utilization is not properly allocated to application it will lead to the faulty services to the customers. In multi-tenant cloud, scaling resource paradigm application shifted to cloud systems will reduce cost of services to clients or customers. Resource scaling can resolve issues of application migration conflicts, which is due to compatibility issues of new paradigm of multi-tenant cloud environment. Certain applications may have issues with cloud infrastructures as those applications are developed in different environment. So, when shifted they may have conflict with resources on cloud and will lead to compatibility issues which may be costly to make applications compatible to multi-tenant cloud environment. Resource scaling should do check of compatibility of applications during application transfer to cloud. Resource scaling is also very important in case of processor availability so that there should be no issue of service due to technical snag during execution of applications.

We know that cloud service demand is increasing day by day and more applications will be on peak demand in future, so dynamic partitioning is very essential in cloud scaling. Cloud scaling Resource conflict can arise in three ways, low allocation of resources, high allocation of resources, and low bandwidth of internet. In low allocation resource, scaling the service will not up to the mark as the resources are not allocated as per need of application. In high resource scaling allocation, too much resources will be allocated that will lead to over-charges to the user. Internet is life line of cloud computing and low speed can cause inconvenience to users due to non-allocation of resources on time. Resource scaling should be done in that manner so that client taking resources on rental basis neither over charged nor did less charge for the usage of cloud resources. We need mechanism in cloud computing to control resources scheduling both ways.

## II. Related Work

A.Singh, D. Juneja, M. Malhotra, proposed a novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing. The main role of

service provider is to effectively distribute and share the resources which otherwise would result into resource wastage. In addition to the user getting the appropriate service according to request, the cost of respective resource is also optimized. They proposed a new Agent based Automated Service Composition (A2SC) algorithm comprising of request processing and automated service composition phases and is not only responsible for searching comprehensive services but also considers reducing the cost of virtual machines which are consumed by on-demand services only.[1]

S.A. Hussain, M. Fatima, A.Saeed, I. Raza, R.K. Shahzad, discussed the Multilevel classification of security concerns in cloud computing. Threats jeopardize some basic security requirements in a cloud. These threats generally constitute privacy breach, data leakage and unauthorized data access at different cloud layers. This paper presents a novel multilevel classification model of different security attacks across different cloud services at each layer. It also identifies attack types and risk levels associated with different cloud services at these layers. The risks are ranked as low, medium and high. The intensity of these risk levels depends upon the position of cloud layers. The attacks get more severe for lower layers where infrastructure and platform are involved. The intensity of these risk levels is also associated with security requirements of data encryption, multi-tenancy, data privacy, authentication and authorization for different cloud services. The multilevel classification model leads to the provision of dynamic security contract for each cloud layer that dynamically decides about security requirements for cloud consumer and provider.[2]

Saraswathi AT, Kalaashri.Y.RA, Dr.S.Padmavathi, presented Dynamic Resource Allocation Scheme in Cloud Computing. Cloud Computing environment provisions the supply of computing resources on the basis of demand, as and when needed. It builds upon advances of virtualisation and distributed computing to support cost efficient usage of computing resources, emphasizing on resource scalability and on-demand services. It allows business outcomes to scale up and down their resources based on needs. Managing the customer demand creates the challenges of on demand resource allocation. Virtual Machine (VM) technology has been employed for resource provisioning. It is expected that using virtualized environment will reduce the average job response time as well as executes the task according to the availability of resources. Hence VMs are allocated to the user based on characteristics of the job. Effective and dynamic utilization of the resources in cloud can help to balance the load and avoid situations like slow run of systems. [3]

M.Verma, GR Gangadharan, NC Narendra, R Vadlamani, V.Inamdar, L. Ramachandran, discussed Dynamic resource demand prediction and allocation in multi-tenant service clouds. Cloud computing is emerging as an increasingly popular computing paradigm, allowing dynamic scaling of resources available to users as needed. This requires a highly accurate demand prediction and resource allocation methodology that can provision resources in advance, thereby minimizing the virtual machine downtime required for resource provisioning. They present a dynamic resource demand prediction and allocation framework in multi-tenant service clouds. The novel contribution of our proposed framework is that it classifies the service tenants as per whether their resource requirements would increase or not; based on this classification, our framework prioritizes prediction for those service tenants in which resource demand would increase, thereby minimizing the time needed for prediction. Furthermore, their approach adds the service tenants to matched virtual machines and allocates the virtual machines to physical host machines using a best-fit heuristic approach. Performance results demonstrate how our best-fit heuristic approach could efficiently allocate virtual machines to hosts so that the hosts are utilized to their fullest capacity. [4]

Z. Shen, S. Subbiah, X Gu, J. Wilkes, presented CloudScale: Elastic Resource Scaling for Multi-Tenant Cloud Systems. Elastic resource scaling lets cloud systems meet application service level objectives (SLOs) with minimum resource provisioning costs. CloudScale employs online resource demand prediction and prediction error handling to achieve adaptive resource allocation without assuming any prior knowledge about the applications running inside the cloud. CloudScale can resolve scaling conflicts between applications using migration, and integrates dynamic CPU voltage/frequency scaling to achieve energy savings with minimal effect on application SLOs. We have implemented CloudScale on top of Xen and conducted extensive experiments using a set of CPU and memory intensive applications (RUBiS, Hadoop, IBM System S). The results show that CloudScale can achieve significantly higher SLO conformance than other alternatives with low resource and energy cost. [5]

W. Lin, J.Z. Wang, C. Liang, D. Qi, proposed a Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing. Compared to traditional distributed computing paradigms, a major advantage of cloud computing is the ability to provide more reliable, affordable, flexible resources for the applications (or users). The need to manage the applications in cloud computing creates the challenge of on-demand resource provisioning and allocation in response to dynamically changing workloads. Currently most

of these existing methods focused on the optimization of allocating physical resources to their associated virtual resources and migrating virtual machines to achieve load balance and increase resource utilization. Unfortunately, these methods require the suspension of the cloud computing applications due to the mandatory shutdown of the associated virtual machines. They propose a threshold-based dynamic resource allocation scheme for cloud computing that dynamically allocate the virtual resources (virtual machines) among the cloud computing applications based on their load changes (instead of allocating resources needed to meet peak demands) and can use the threshold method to optimize the decision of resource reallocation.[6]

P. Pradhan, R.K.Behera, BNB Ray, presented Modified Round Robin Algorithm for Resource Allocation in Cloud Computing. Cloud computing is an attractive computing model since it allows for the provision of resources on-demand. Cloud computing has emerged as a new technology that has got huge potentials in enterprises and markets. Clouds can make it possible to access applications and associated data from anywhere. Companies are able to rent resources from cloud for storage and other computational purposes so that their infrastructure cost can be reduced significantly. Hence there is no need for getting licenses for individual products. Cloud Computing offers an interesting solution for software development and access of content with transparency of the underlying infrastructure locality. The Cloud infrastructure is usually composed of several data centres and consumers have access to only a slice of the computational power over a scalable network. The provision of these computational resources is controlled by a provider, and resources are allocated in an cloud computing is related to optimizing the resources being allocated. The other challenges of resource allocation are meeting customer demands and application requirements. In this paper, modified round robin resource allocation algorithm is proposed to satisfy customer demands by reducing the waiting time.[7]

We can create thousands of server instances and allocate them simultaneously. Every instance can be controlled separately by the medium of middleware known as virtual machine. Flexible cloud resource hosting services can be provided with multiple choices of instances and could configure the memory, operating system, instances in boot partition [9]. Every day millions of new internet users get themselves registered for new connections of internet thus this automatically increase more traffic over the internet manifolds so as workload. To tackle the workload on servers we need dynamic provisioning of different data centres with guarantee. Such approach is based on a dedicated or shared model [10].

Servers provide number of multiple services from common hardware base for resources. The resources are managed by centrally hosted operating system. The resources provisioning of servers on co-hosted services automatically to offered load, improve the energy services of server clusters by dynamically hosting centres. A greedy resource algorithm adjusts resources costs to balance demand and supply [11]. Light weight dynamic voltage and frequency technique implemented on modern multi-tasking system. The techniques implement on processors execution statistics and an online learning algorithms to power-up the accurate suited voltage and frequency settings [12]. Spade efficiently fetches performance optimization and scalability to system applications. Spade works on code generation framework to develop highly optimized that execute on stream processing core (spa). Online information resources are increasingly hold the shape of data streams. [13]Advancement in servers, computers networks and data storage virtualization are permitting the development of resource pools of servers that permits multiple application workload to share server in the pools. The trace based method to tackle management a) describing required availability b) the characteristics of load patterns c) the prediction of synthetic load played by required services [14].

Data virtualization permits price influenced server consolidation and consume less power with increased results. It is hard to do proper resource management of virtualized servers. The control based theory on resource management has shown the important advantages of scaling allocations to identify changing loads of work. The kalman filter is proposed to feedback controllers to dynamically allocate processor resources. Virtualized machines several applications. The optimal technique of filtering states for the calculate in the summation of square sense to trace the processor utilization and upgrade the allocation accordingly [15].Elasticity of computing resources systems gain and relieve resources to dynamic workloads, and paying for those only the needed, this character of cloud computing. The core of any elastic system is with automated controls. The multi-tier applications services that allocates and relieves resources in segments, such as virtual server instances of predefined sizes. It highlights on elastic control of the storage tier, in which storage and removing from machines or brick needs re-balancing stored data on all the machines .the new challenges for storage tier presents for elastic controls. Elastic resources scaling needs mechanism to present a wide range of applications [16].

The increasing interests towards the information technology on virtualization technologies and utility computing have developed for more balancing workload management tool. One that

achieves quality services (Qos) and also dynamically control resource allocated on the application services. These ways can in turned and dragged by the utility of services provided, they are all depended on those application service level agreements (S.L.A) and cost of resources allocated to the applications.[17]. Internet applications are emerging in every sector of life. The online portals like news, retails, ecommerce and financial online portals of banks have become market place in recent years. The applications on internet are becoming hard and complex systems of software that employs multi-tier architecture and are replicated or distributed on a cluster of servers. Tiers have separate functionality and its preceding tier and created the functionality provided by its successors to take out its part of the total requested of processing. For instance an ecommerce application consists of three tiers- front of web is one tier that is responsible to requesting to server by using portal of http (hypertext transfer protocol). Java is a middle tier of server which imposes core application functionality and database as backend tier stores data catalogues various products. [18] Cloud computing permits tenants to rent resources as required and go way. The application running on the cloud required right number of computing resources to achieve the advantage. It is very important to cloud resources scaling infrastructure to maintain service level objectives and it financial benefits. Allocating resources with over provisioning wastes resources. [19]

Social or business websites are built on rated of a traditional databases has then own problems when scaling resources at the storage at backend. The high request rate of social networking sites with increasingly powerful hardware but has low latency. Building top relational database clusters and due to low latency of these systems. Face book is one of the examples of the popular networking websites. It has dynamically two billion Web pages per day. Traffic management results are over 23000 page views in a second. Each of which could results in many quires of the database. The architecture of face-book has forced to respond to this load by federating their 1800 plus instances in database and many of them are independent, geographically distributed clusters [20].

Elastic resource scaling lets cloud systems meet application service level objectives (SLOs) with minimum resource provisioning costs. A prediction-driven elastic resource scaling system for multi-tenant cloud computing. The goal is to develop an automatic system that can meet the SLO requirements of the applications running inside the cloud with minimum resource and energy cost. [22] Infrastructure as a Service is a provision model in which an organization outsources the equipment used to support operations, including storage, hardware, servers and networking components. The service provider owns the equipment and is responsible for housing, running and maintaining it. The client typically pays on a per-use basis.[23] Public infrastructure-as-a-service clouds, such as AmazonEC2, Google Compute Engine (GCE) and Microsoft Azure allow clients to run virtual machines (VMs) on shared physical infrastructure. This practice of multi-tenancy brings economies of scale, but also introduces the risk of sharing physical server with an arbitrary and potentially malicious. Past works have demonstrated how to place a alongside a target victim (co-location) in early-generation clouds and how to extract secret information via side channels.[24]. Cloud computing has transformed the way applications are created and run tremendously in recent years. It employs the Infrastructure as a Service (IaaS) model in which customers outsource their computing and software capabilities to third party infrastructures and pay for the service usage on demand. Compared to the traditional computing model that uses dedicated, in-house infrastructure, cloud computing provides a number of advantages, including economies of scale, dynamic provisioning, and low capital expenditures. [25].

## III. OUTCOME OF THE ANALYSIS

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud application over the internet. Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment to meet the needs of the cloud application. The order and time of allocation of resources are also an input for an optimal RAS. An optimal RAS should avoid the following criteria:

i. **Resource contention**
ii. **Scarcity of resources**
iii. **Resource fragmentation**
iv. **Over-provisioning**
v. **Under-provisioning**

From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical. For the cloud users, the job should be completed on time with minimal cost. Hence due to limited resources, resource heterogeneity, locality restrictions, environmental necessities and dynamic nature of resource demand, we need an efficient resource allocation system that suits cloud environments.

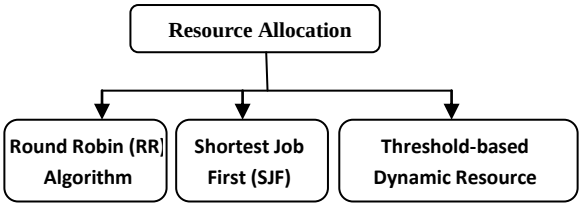


**Fig 1 : Resource Allocation Schemes**

In general, the workload of network applications (such as Web applications) fluctuates over the application lifetime. If we use a static resource allocation scheme to assign fixed resources to an application, the application may be slowed down sometimes due to insufficient resources, or excessive resources are wasted when application is not at its peak load. Therefore, it is ideal to design a dynamic resource allocation scheme that can adjust the resources allocated to an application according to its workload, thus, improving the resource utilization. The Round Robin (RR) algorithm allocates the resources to the process for the fixed time slots in round robbing fashion. The Shortest Job First (SJF) algorithm, first evaluates the required time to complete the process and then allocates the resource(s) first to the process that will take less time. The main idea of the threshold-based dynamic resource allocation scheme is to monitor and predict the resource needs of the cloud applications and adjust the virtual resources based on application's actual needs. A dynamic resource allocation scheme needs to address two issues: when to reallocate resources and how much resource to be adjusted.

The design goals of the Data Centre Network (DCN) architecture are :

- Improving Availability and Fault tolerance
- Scalability: Incrementally increase DCN size as and when needed
- Low cost: Lower power and cooling costs
- Throughput : The number of requests completed by the data center per unit of time. (Compute + Transmission + Aggregation Time)
- Economies of scale : utilize the benefits of its huge size
- Scalable interconnect bandwidth: Host to host communication at full bisection bandwidth
- Load balancing: Avoid hot-spots, to fully utilize the multiple paths

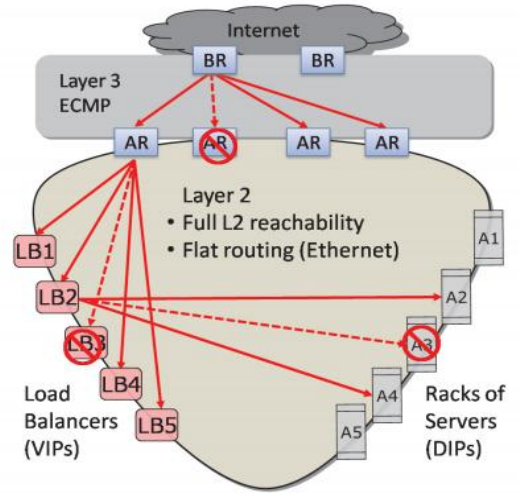


**Figure 3: Overview of the Monsoon architecture**

The following are the challenges for establishing a DCN:

- Reduce Utilization
- Scale-out vs Scale-up : per-port cost, cabling and packaging complexity, scalable cooling
- Placement, Air-flow and rack-density
- TCP Incast, Large Buffer switches
- Resource fragmentation: VLANs
- Manual configuration
- Oversubscription : 1:1 vs 1:240
- Flooding and Routing n/w overhead

Conventional Data Centre have many drawbacks like Ethernet is hard to scale out, fragmentation of resources, no performance isolation, poor server to server connectivity, need very high reliability. So, new architectures have been proposed like The Monsoon Architecture, The VL2 Architecture, The SEATTLE Architecture, The Portland Architecture, and The TRILL.

## IV. CONCLUSION

As internet users are continuously increasing, the requirement of the available IT resources and the need of computing for efficiently using these resources is also increasing. After analyzing the architectures proposed for cloud data centre networks for the resource pooling and allocating, it is observed that the new architectures like, The Monsoon Architecture, The VL2 Architecture, The SEATTLE Architecture, The Portland Architecture, and The TRILL are the good solutions. Even though the above mentioned architectures are the good solutions, but still the demand of research in the area of developing new architecture is still required.

The researchers may work on Automated Request Processing Layer to handle the requests for improving the resources availability and fault tolerance. The various agents like Interface Agents, Broker Agents, Directory Agents, Resource Manager Agents etc should be more error prone to improve the Data Centre Networks. Resource Allocation Schemes like, Round Robin (RR) algorithm, SJF algorithm, and Threshold-based Dynamic Resource Allocation Scheme are used to allocated the resources efficiently in cloud computing architecture.

## ACKNOWLEDGMENTS

We feel grateful to the anonymous referees for their comments and for their valuable suggestions that have helped immensely in preparing the revised manuscript.

## REFERENCES

[1] A.Singh, D. Juneja, M. Malhotra, "A novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing", Journal of King Saud University – Computer and Information Sciences (2015), pp. 1-10, 1319-1578.

[2] S.A. Hussain, M. Fatima, A.Saeed, I. Raza, R.K. Shahzad, "Multilevel classification of security concerns in

[3] cloud computing", Applied Computing and Informatics (2016), pp.2-9, http://dx.doi.org/10.1016/j.aci.2016.03.001

[4] Saraswathi AT, Kalaashri.Y.RA, Dr.S.Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing", Procedia Computer Science 47 ( 2015 ) 30 – 36, doi: 10.1016/j.procs.2015.03.180

[5] M.Verma, GR Gangadharan, NC Narendra, R Vadlamani, V.Inamdar, L. Ramachandran, "Dynamic resource demand prediction and allocation in multi-tenant service clouds", Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.3767

[6] Z. Shen, S. Subbiah, X Gu, J. Wilkes, "CloudScale: Elastic Resource Scaling for Multi-Tenant Cloud Systems", ACM 978-1-4503-0976-9/11/10, October 27–28, 2011,

[7] W. Lin, J.Z. Wang, C. Liang, D. Qi, "A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing", Procedia Engineering 23(2011), pp. 695-703

[8] P. Pradhan, R.K.Behera, BNB Ray, "Modified Round Robin Algorithm for Resource Allocation in Cloud Computing", International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science 85 ( 2016 ), pp. 878 – 890

[9] Amazon Elastic Compute Cloud.http://aws.amazon.com/ec2/.

[10] Abhishek Chandra, Weibo Gong, PrashantSheno.Dynamic Resource Allocation for Shared DataCentres Using Online Measurements 2003

[11] J. Chase, D. Anderson, P. N. Thakar, and A. M. Vahdat.Managing energy and server resources in hosting centers. InProc. SOSP, 2001.

[12] X. Fan, W.-D.Weber, and L. A. Barroso. Power provisioningfor a warehouse-sized computer. In Proc. ISCA, 2007.

[13] D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper. Capacitymanagement and demand prediction for next generation datacenters. In Proc. ICWS, 2007.

[14] E. Kalyvianaki, T. Charalambous, and S. Hand. Self-adaptiveand self-configured CPU resource provisioning forvirtualized servers using Kalman filters. In Proc. ICAC,2009.

[15] H. Lim, S. Babu, and J. Chase. Automated control for elasticstorage. In Proc. ICAC, 2010.

[16] Xiaoyun Zhu, Zhikui Wang, SharadSinghal Utility-driven workloadmanagement using nested control design. In Proc. AmericanControl Conference, 2006.

[17] B. Urgaonkar, M. S. G. Pacifici, P. J. Shenoy, and A. N.Tantawi. An analytical model for multi-tier internet services and its applications. In Proc. SIGMETRICS, 2005.

[18] Z. Gong, X. Gu, and J. Wilkes. PRESS: Predictive Elastic Resource Scaling for Cloud Systems. In Proc. CNSM, 2010.

[19] M. Armbrust, A. Fox, D. A. Patterson, N. Lanham,B. Trushkowsky, J. Trutna, and H. Oh. Scads: Scale-independent storage for social computing applications. In Proc. CIDR, 2009.

[20] ZhimingShen, Sethuraman Subbiah, Xiaohui Gu, John Wilkes, "CloudScale: Elastic Resource Scaling for Multi-Tenant Cloud Systems" 2011

[21] VenkatanathanVaradarajan, Yinqian Zhang, Thomas Ristenpart_, and Michael Swift†, "Placement Vulnerability Study in Multi-Tenant Public Clouds"

[22] Jing Zhu_, Dan Li_z, Jianping Wu_, Hongnan Liu_, Ying Zhangy, JingchengZhang_ "Towards Bandwidth Guarantee in Multi-tenancy Cloud Computing Networks"

[23] T.Garg1, R.Kumar2, J. Singh, "A way to cloud computing basic to multitenant environment"

[24] Keng-Mao Cho, Pang-Wei Tsai, Chun-Wei Tsai, Chu-Sing Yang, "A hybrid meta-heuristic algorithm for VM scheduling with load balancing in cloud computing."

[25] Vasilios Andrikopoulos, Tobias Binz, Frank Leymann, Steve Strauch, "How to adapt applications for the Cloud environment Challenges and solutions in migrating applications to the Cloud"

[26] Zhang S, Qian ZZ, Wu J et al. Service-oriented resource allocation in clouds: Pursuing flexibility and efficiency. JOURNALOF COMPUTER SCIENCE AND TECHNOLOGY 30(2): 421–436 Mar. 2015. DOI Service-Oriented Resource Allocation in Clouds: Pursuing Flexibility and Efficiency.