Adaptive resource scaling methods for Multi-tenant cloud system

Dr. Amit Chaturvedi[#1], Zahoor Ahmad Bhat [*2]

[#]*1Assistant Prof.MCA Deptt., Govt. Engineering College, Ajmer*

**2 M.Phil. Scholar, Ajmer, India*

Abstract: *Cloud resource scaling is an important issue to address proper resource utilization in multi-tenant cloud computing environment. Analysis of Resource Scaling Infrastructure as Service is the main objective in this paper. Multi-tenant environment applications use virtualized technologies to encapsulate and segregate application performance by using separate virtual machines (VM). We have discussed three issue of resource scaling i) high resource scaling ii) low resource scaling iii) internet speed. So, we proposed that researcher should focus on developing such solutions that will allocate the resources dynamically and on demand. The solutions should be so dynamic, if some resources are occupied for a long time and another resource requirement occurs. The provider should adjust the balance from the already booked resource this will be a most economical solution.*

Keywords: *Allocation, Multi-tenant, Virtualization, Scalability, Cloud Environment, Dynamic, Iaas.*

I. Introduction

Cloud resource scaling is an important issue to address proper resource utilization in multi-tenant cloud computing environment. Analysis of Resource Scaling Infrastructure as **S**ervice is the main objective in this paper. 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 snow shifted to cloud systems and it 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.

As we are aware that even not more than 2% applications rarely static, thus static partition of applications will lead to fixed allocation of memory. 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 clouds scaling. Cloud scaling Resource conflict can arise in three ways a) low allocation resource scaling b) high allocation resource scaling c) low speed 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. Hence to overcome with the problem of resource scaling automatic resource scaling is best option to do the job.

II. Review of the related work

Cloud scaling of resources is essential to hold utilization of resources by proper ways during transformation of any stack of application, database, operating system, etc. Cloud scaling enables scale-up and down automatically. 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 [1]. 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 [2].

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 [3].

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 [4].

Spade efficiently fetches performance optimization and scalability to system applications. Spade works on code generation framework to develop highly optimised that execute on stream processing core (spa). Online information resources are increasingly hold the shape of data streams. [5]
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 [6].

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 in Virtualized machines for 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 [7]

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 [8].

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.[9].

Internet applications are emerging in every sector of life. The online portals like news, retails and 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. [10]

Cloud computing permits tenants to rent resources as required and go way. The potential to offer cost reliable resolution rather than maintain their own or arrange and maintain their complex, much costlier infrastructure themselves. 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 objective. And it financial benefits. Allocating resources with over provisioning wastes resources. [11]
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 [12].

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 of our 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. [14]

III. Need of the proposed work

Today we are living in new technological paradigm of cloud computing. It brings revolution in computing world. This paradigm is under research and there are ways to improve in every sector to leave fewer margins to vulnerabilities. Our analysis is focused on resource scaling in cloud where infrastructure as a service (IaaS) is platform for customer. This sector has important need of evaluation so that we can get more improved way of resource utilization in multi-tenant cloud environment. Resource scaling is all about better way of utilization of computing resources. So we are doing focus on need of improve some important features in our analysis.

IV. Proposed Model and Methodology :

As there are three important issues to be consider when discussing the allocation of the resources (a) Low resource allocation, (b) High resource allocation, and (c) Low speed of internet. These issues in detail are discussed below:

a. Low resource allocation:-

As we have stated above in the figure 1, that low resource scaling is causing problems because sometimes users are not sure how much space they want to take on rent. in the above proposed diagram we have showed how multiple users are using resources of cloud and we know that any point of time other n-number of multiple user can try to access same cloud resources. This may lead to low resource scaling and it is possible as we have seen that if the load doubles instead of what we have available in cloud. There are no methods in place in cloud resource scaling to stop any user from access same cloud.

b. High resource allocation:

High resource scaling is another aspect of multi-tenant cloud environment. Where users allocate resources without knowing proper need and rent more resources than their requirement. In high resource scaling we are not only wasting resources but also preventing many other users from accessing the same resources. Cloud computing environment is based on cost as per use of the requirements. There should be mechanism in place to adjust resources dynamically. So that customer will not only use exactly what they need but also save their money. Thus we need proper scaling of resources in the cloud environment to make it dynamically adjustable as per need. Because we know that online users can grow or shrink manifolds at any point time.

c. Low speed of internet:

Internet is emerged and continuously emerging as global market place for companies, educational organisations, businessmen, and social networking. It is wish of every organisation if they are not doing business online they do wish in future. This has made internet a mesh of networks and work load is increasing day by day. Many organisations are now shifted their business from normal server and client architecture to new booming cloud computing environment. Any business on internet is based on speed of internet so as multi-tenant cloud environment. In our paper, low speed of internet is described as new problematic point for any multi-tenant cloud environment and presented in figure 2. Multi-tenant cloud environment is all about multiple services in cloud environment but these services can be ruined if the speed of internet in low. It will directly affect resource scaling in cloud based system. When resource is not allocated on time due to low speed of internet it will directly affect business organisations and those applications which hosted in that cloud. Low speed requires more response time to respond its users to allocate resources and this will lead to wastage of time and low response can cause inconvenience to customer.

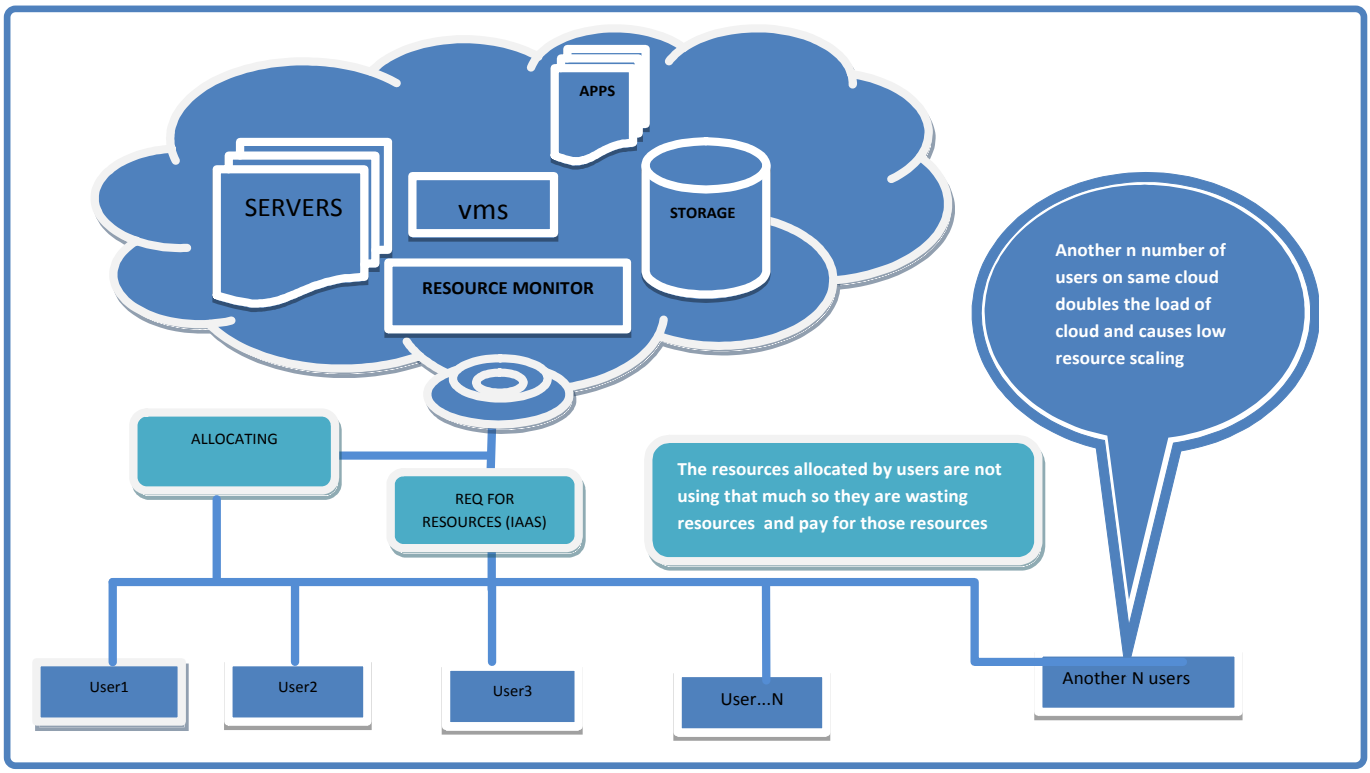

Figure 1 : Resource Allocation Process

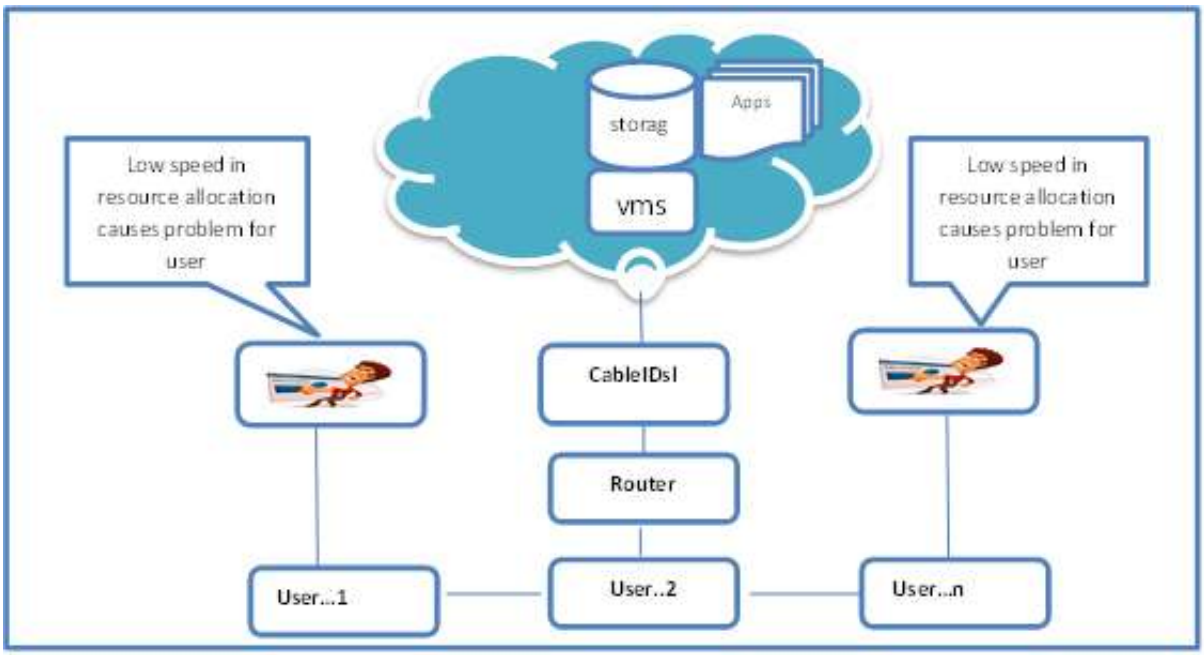

Figure 2: Low Speed of Internet in cloud

V. Conclusion

After analysing the various schemes and models proposed for resource scaling in cloud computing. We concluded that the resource scaling is important factor and dynamic scaling of resources is best option to do the job. In this analysis, we find that resources scaling has issues of allocation in multi-tenant environment. It has issue of low resource scaling; where resources are allocating but they lack the actual allocation requirement i.e. lack of resources many times is the problem in this case. Another aspect is high resource allocation in this analysis where resources are allocated to more than the requirement of users. Here in this case, the wastage of resources is the problem found. Third issue is the internet speed issue that is life line of any internet services so as in multi-tenant cloud. Internet speed is widely affecting internet as whole multi-tenant cloud environment. Any kind of service is possible but when we place whole organisations, educational institutions, and business online services round the clock 24x7 any kind of obstacles to speed of internet will lead to big risk to whole organisation or business.

So, we proposed that researcher should focus on developing such solutions that will allocate the resources dynamically and on demand. The solutions should be so dynamic, if some resources are occupied for a long time and another resource requirement occurs. The provider should adjust the balance from the already booked resource this will be a most economical solution.

So, future researcher may work in this direction to develop some solutions.

References:

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

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

[3]. J. Chase, D. Anderson, P. N. Thakar, and A. M. Vahdat.

[4]. Managing energy and server resources in hosting centers. In*Proc. SOSP*, 2001.

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

[6]. 2008the System S declarative stream processing engine

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

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

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

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

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

[12]. Z. Gong, X. Gu, and J. Wilkes. PRESS: PRedictive ElasticReSource Scaling for Cloud Systems.In*Proc. CNSM*, 2010.

[13]. 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.

[14]. ZhimingShen, SethuramanSubbiah, XiaohuiGu, John Wilkes, CloudScale: Elastic Resource Scaling for Multi-Tenant Cloud Systems 2011

[15]. ArchanaGanapathi†, Harumi Kuno§, UmeshwarDayal§, Janet L. Wiener,

[16]. Armando Fox†, Michael Jordan†, David Patterson