ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 7, July 2016

Survey On Scalability In Cloud Environment

Ab Rashid Dar, Dr.D.Ravindran

Abstract: Cloud computing in recent times is most talked technology which offers resources from the large data centers. The main purpose of cloud computing is to allow clients to take the advantage from all of these technologies. Cloud Computing is made available as a pay on demand service to the clients. Cloud services are "pay-per-use" over the internet. . As the services provided by cloud service provider are chargeable so the clients are only charged for the desired services. It is on demand access to virtualized IT services and products. Some of the renowned Cloud service providers in IT sector are as RackSpace, Sales force, Amazon, Google, IBM, Dell and HP.It has many features that include measured services, availability, security and scalability. Among all the most and exciting feature of cloud computing is Scalability which offers the clients the ease and comfort to use the resources as per their expectations and demand. The unpredictable nature of cloud applications and services and to handle it. Scalability mechanism is more important than anything else. Resources can be scaled in and scaled out when the demand and situation arises like that. It might be static or dynamic in nature. But dynamic or Autoscaling is most widely implemented and preferred over static scaling in Cloud environment. The cloud service providers mostly use the Autoscaling mechanism where the resources offered to the clients are only as per their interest and requirements and can be added and removed any time without the intervention of third party and even clients. The main purpose of this paper is to give an overview of cloud computing and more importantly it emphasizes the auto scaling mechanism. It also put some light on its types and usage in cloud computing applications.

Index terms: Cloud Computing, Scalability, Autoscaling, Vertical& Horizontal Scalability

I.INTRODUCTION

Cloud Paradigm is an acronym used for Internet, its services, is the latest and highly evolved computational model which is basically based on different cloud computing paradigms that already exist and is only functional with the help of DSL as it provide the maximum internet speed and bandwidth. Cloud computing came into its existence because of the evolution and adoption of already existed technologies and paradigms. The main purpose of using the cloud is to get the benefits from different Computing paradigms. The cloud aims to reduce the costs, and helps the clients to focus on their business rather to care about the obstacles and barriers within the information technologies. NIST defines cloud as "Cloud computing paradigm is a model for enabling global, suitable, on demand network accesses to the related configurable computing resources like networks, servers, storage, applications and services that can be rapidly equipped and released with less management effort or service provider intervention"[1].

Ab Rashid Dar,
Research Scholar, Department of Computer Science, St. Joseph's College, Tiruchirappalli-02, India.

Dr.D.Ravindran,
Dean, School of Computing Science, St. Joseph's College, Tiruchirappalli-02, India.

The clients are also unconcerned and unaware about the number of servers or details of the other resources that are necessary to support their currently desired needs and requirements, the clients simply want to pay for the computing capacity or the services which they use and expect the capacity to scale up or down to meet their current requirements in an on-demand fashion.

II.CLOUD COMPUTING SERVICE OR BUSINESS MODELS

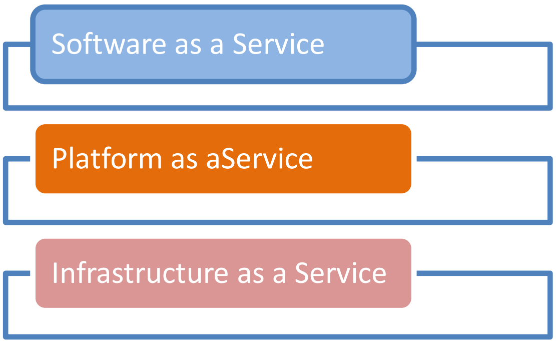

Fig.1.SPI Model of Cloud Computing

A. *Software as a Service*

It is the model in which an application is hosted as a service to clients who access the cloud through internet. Clients can access and use their application anywhere on the globe when having an internet connection.It is a Deployment model client uses provider's applications running on cloud infrastructure applications are designed for end-users, delivered over the webhosted and managed by vendor, delivered across the internet [2]. Cloud providers install and operate application software in the cloud. As the services are chargeable, so the service providers take some monthly fee or clients pay as per the services and facilities they use, so price is scalable and adjustable if users are added or removed at any point.Some of the examples of Software as a Service include, Google Apps, innkeypos, Quickbooks Online, Limelight Video Platform, Salesforce.com, and Microsoft Office 365, Gmail.

B. *Platform as a Service*

This is another delivery application model which provides resources required to build an application and services completely from internet without purchasing them [2].Usually the IT sector people like software developers and different information service providers use this platform on commercial scale basis, try to gain the market and mainly concentrate on get more and more money by publishing the advertisements and attract the cloud clients to buy their developed applications and services. The software developers without any cost and complication can develop and deploy their software on the cloud platform. Some of the examples of Platform as a Service are Amazon Elastic Beanstalk, Cloud Foundry, Heroku, Force.com, EngineYard, Mendix, Google App Engine, Microsoft Azure and OrangeScape

C. Infrastructure as a Service

The hardware tools and software applications that are need for the clients are offered by the vendors as Infrastructure as a Service and clients can put or run anything they wanted on cloud environment [2]. It offers the necessary resources like server space, CPU cycles, network equipments, memory and storage space to the cloud clients on rent basis, clients use and pay as per their demand and requirements. Clients do not have to bother about purchasing the servers, software, datacenter space or network equipment, Clients acquires only desired resources. Generally IaaS can be obtained as public or private infrastructure or a combination of the two. Examples of IaaS are as, Amazon Elastic Compute Cloud (EC2), Eucalyptus, GoGrid, FlexiScale, Linode, RackSpace Cloud, Terremark.

III.CLOUD DEPLOYMENT MODELS

A. Public Cloud

Public clouds are made available to the general public by a service provider who hosts the cloud infrastructure. Generally, public cloud providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access over the Internet. With this model, customers have no visibility or control over where the infrastructure is located [3]. It is important to note that all customers on public clouds share the same infrastructure pool with limited configuration, security protections and availability variances.

B. Private cloud

Private cloud is cloud infrastructure dedicated to a particular organization. Private clouds allow businesses to host applications in the cloud, while addressing concerns regarding data security and control, which is often lacking in a public cloud environment [3,4]. It is not shared with other organizations, whether managed internally or by a third-party, and it can be hosted internally or externally.

C. Hybrid cloud

Hybrid Clouds are a composition of two or more clouds (private, community or public) that remain unique entities but are bound together offering the advantages of multiple deployment models. In a hybrid cloud, you can leverage third party cloud providers in either a full or partial manner; increasing the flexibility of computing [3, 1]. Augmenting a traditional private cloud with the resources of a public cloud can be used to manage any unexpected surges in workload.

D. Community

A community cloud is a is a multi-tenant cloud service model that is shared among several or organizations and that is governed, managed and secured commonly by all the participating organizations or a third party managed service provider[3]. Community clouds are a hybrid form of private clouds built and operated specifically for a targeted group. These communities have similar cloud requirements and their main purpose is to work in co-ordination or collaboration to achieve their business objectives.

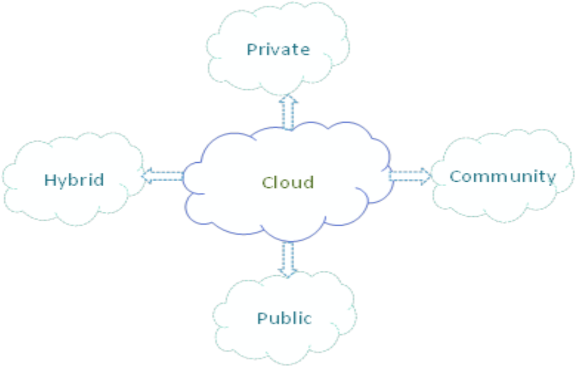

Fig.2. Deployment Cloud Computing Models

IV.CHARACTERISTICS OF CLOUD COMPUTING PARADIGM

A. Storage and Scalability

The storage is no more a limitation when clients are using cloud platform and they don't have to buy now the blocky and costly hardwarical components like servers and storage devices etc. Client basically has access to unlimited storage capability and scalability. On demand, clients can add or remove the resources at any point of time. Scalability is the unique feature of cloud computing where dynamic provisioning of the resources is being done by the clients themselves with in the real time slice [5].

B. Backup and Disaster Recovery

Those days of tape back-up where clients used to store their vital data are long gone. The cloud vendors provide their clients platforms and comfort to back up their vital data, at any point of disastrous situations, the vendors offer them the ease to recover their lost data any time [5].

C. Mobility

Mobility provides the cloud the “on the go” feature. Whether it's your development platform, suite office tools or custom content management system – cloud mobility enables access anywhere with a web connection [5]. It makes cloud easy to operate from anywhere on the globe and clients can access their applications and other resources from various devices like smart phones, tabs, desktops etc.

D. Cost Efficiency

Cost is one of the constraints that abide the clients to use and access the IT resources like storage, servers, and network. But since the advancement of cloud computing paradigm the Clients can use software or applications, with minimal service charges [5,6]. It's because cloud computing offers the most exciting feature as multi-tenancy, service level agreement and also cloud also offers some of open source products.

E. Availability

Cloud possesses the property of being available 24X7 hours. The availability feature makes cloud every organizations their first choice to run their business [5]. The e-commerce organizations like Amazon, Flipkart, Snapdeal etc. are dependent on the availability of cloud.

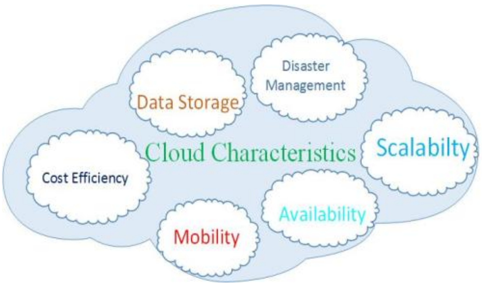

Fig.3.Characteristics of Cloud Computing

V.SCALABILITY IN CLOUD COMPUTING: AN EXCITING & ADVANTAGEOUS FEATURE MAKING CLOUD WORTHWHILE TO ADOPT

A. Cloud Scalability

One of the key benefits of using cloud computing paradigm is its scalability. It supports the long term strategies and business needs and is entirely different than elasticity. It is the mechanism by which clients dynamically provision their resources like hardware devices and software applications when demand and situation arise like that. Cloud computing allows clients or cloud vendors business to easily scale up or scale down their IT requirements as and when required[5]. For example, most cloud service providers will allow clients to increase their existing resources to accommodate increased business needs or changes. This will allow clients to support their business growth without expensive changes to the existing systems. Because of the highly scalable nature of cloud computing paradigm, many organizations are now relying on managed data centers where there are cloud experts trained in maintaining and scaling shared, private and hybrid clouds. Cloud computing allows for quick and easy allocation and reallocation of resources in a monitored environment where overloading or load balancing is no more a concern as long as the system is managed and maintained properly. The most important technology which enables the cloud paradigm to scale up and scale down the resources is virtualization [6], without it cloud computing is not sufficient, it provides the agility and speed up the execution of processes. Autoscaling reduces the client's manual involvement and an intervention thus minimizes the possibility of client's errors, provides automation to the resources, increases the speed and reduces the laborer costs. Provisioning the resources automatically by implementing Autoscaling mechanism making the cloud first choice for the different e-commerce organizations residing on it.

B.Cloud Scalability Types

1. Horizontal Scalability (Scaling out)

Horizontal cloud scalability is the ability of the system or resources to connect multiple hardware or software entities, such as servers or networks so that they work as a one logical unit [5]. It means adding more individual units of resource doing the same job. For example, in the case of servers it could increase the speed or availability of the logical unit by adding more servers as per the needed. Instead of one server here one can have two, ten, or more of the same server doing the same work. It is also referred to as scaling out

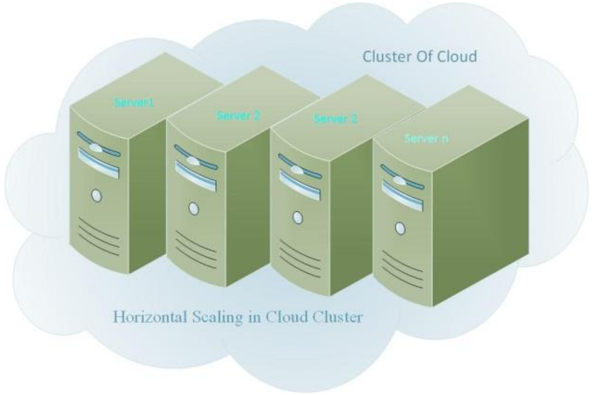

Fig.4.Horizontal Scaling of Servers in a Cluster

2. Vertical cloud scalability:

Vertical scalability is the ability to increase the capacity of existing single hardware or software by adding more resources to the same server or hardware. For example, adding processing power to a server to make it faster [7]. It can be achieved through the addition of extra hardware to the same entity such as hard drives, servers, CPU's, etc. It provides more shared resources for the operating system and applications. This type of scalability may also be referred to as scaling up or scaling in.

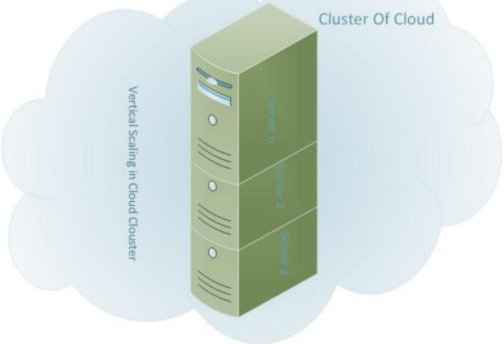

Fig.5.Vertical Scaling of Servers in a Cluster

VI.RELATED WORK

M.Kriushanth et al.,[5] discussed about the basics of cloud computing concepts like service models, deployment models and the various dimensions of cloud scalability. The dimensions they given are vertical and horizontal scalability. They presented infrastructure of auto scaling and some challenges and issues that are occurring in auto scaling. The issues they given are is taken into account for future work.

R. Anandhi et al.,[8] presented the basics of scalability and its scalability factors. Here they distinguish the scalability into four by its scalability factors. Then they described why and how the scalability has been chosen based on the user requirement. Further they describe the two types of approach in messaging system of scalability. They had given the way to improve the scalability through auto scaling, scaling the database horizontally and EBS.

Jorge M. Londoño-Peláez et al.,[9] explained about the way to solve the two problems like over provisioning and under provisioning. To address these problems they present an autonomic auto-scaling controller that based on the stream of measurements from the system maintains the optimal number of resources and responds efficiently to workload variations, short duration peaks in the workload. Their technique consists of three components. It has been explained with clearly with its parameter tuning. These techniques have also been analyzed.

HaniehAlipour et al.,[10] presented a survey that explores definitions of related concepts of auto-scaling and taxonomy of auto-scaling techniques. Based on the survey results, they outline open issues and future research directions for this important subject of auto scaling in cloud computing. They explained each and every concept of the auto scaling taxonomical areas. This gives the new various areas if research sectors especially in the area of auto scaling.

Che-Lun Hung et al., [11] proposed the novel virtual cluster architecture for dynamic scaling of cloud applications in a virtualized Cloud Computing environment. An auto-scaling algorithm for automated provisioning and balancing of virtual machine resources based on active application sessions will be introduced. Also, the energy cost is considered in this proposed algorithm. This work has demonstrated the proposed algorithm is capable of handling sudden load requirements, maintaining higher resource utilization and reducing energy cost. They proposed the two algorithms for auto scaling of web applications and for the distributed systems.

Zhen Xiao et al.,[12] presented a system that provides automatic scaling for Internet applications in the cloud environment. They encapsulated each application instance inside the virtual machine (VM) and use virtualization technology to provide fault isolation. They model the architecture as the Class Constrained Bin Packing (CCBP) problem where each server is a bin and each class represents an application. The class constraint reflects the practical limit on the number of applications a server can run at the same time. They explain this with an efficient semi-online color set algorithm that achieves good demand satisfaction ratio and saves energy by reducing the number of servers used when the load is low. Their experimental results demonstrate that our system can improve the throughput by 180% over an open source implementation of Amazon EC2 and restore the normal QoS five times as fast during flash crowds. The main disadvantage is that the authors concentrate on the two-tier architecture and now the internet applications have been developing to three tier architecture.

Sushil Deshmukh et al.,[13] they provides automatic scaling for web application in cloud environment. So every application instance encapsulated inside virtual machine and model it as the Class Constraint Bin Packing (CCBP) problem. It is concentrating on the two areas of the application placement and load distribution. Where each class represents an application and each server is a bin and uses virtualization technology for fault isolation. Now many business customers need good satisfy response services from cloud. So they designed and developed a semi online color set algorithm that achieve good demand satisfaction ratio and as well as when load becomes low it reducing number of server and save energy. It explained about all the any fit algorithm and take supports of green computing to adjusting the placement of application instance adaptively and putting ideal machine into the standby mode.

Maram Mohammed Falatah et al.,[14] describing about the definitions of cloud scalability. Scalability is the ability of the systems to do the works which are giving by the user in a fast manner. It has to take care of the many parameters of load balancing, resource allocation, and optimization. They had given the scalability levels and its performance considerations. This paper then presents the scaling approaches and further it gives the details about the study about the web application in cloud.

PranaliGajjar et al.,[15] had given objective of this paper was to present a comprehensive study about the auto-scaling mechanisms available today. They started their work by explaining the initial taxonomy of scalability. Auto-scaling techniques are diverse, and involve various components at the infrastructure, platform and software levels. Many techniques have been proposed for auto-scaling. They proposed classification of these techniques into five main categories: static threshold-based rules, control theory, reinforcement learning, queuing theory and time series analysis.

Trieu C. Chieu et al.,[16] described a novel architecture for the dynamic scaling of web applications based on thresholds in a virtualized Cloud Computing environment. They illustrated scaling approach with a front-end load-balancer for routing and balancing user requests to web applications deployed on web servers installed in virtual machine instances. A dynamic scaling algorithm for automated provisioning of virtual machine resources based on threshold number of active sessions will be introduced. The on-demand capability of the Cloud to rapidly provision and dynamically allocate resources to users will be discussed. This work has demonstrated the compelling benefits of the Cloud which is capable of handling sudden load surges, delivering IT resources on-demands to users, and maintaining higher resource utilization, thus reducing infrastructure and management costs.

VII.CONCLUSION

This paper starts with brief introduction to cloud computing paradigm, its service and deployment models and whole attention is being given to the most advantageous feature of cloud computing paradigm i.e. Scalability and its types. It can be both static and dynamic in nature but Autoscaling is being preferred over static mostly. Brief and basic definitions of the auto scaling techniques and the related works to it. Cloud Computing is a vast and open field where researchers can carry on their research on its different aspects, Autoscaling is one among them , where still research has to be done .As most probably in this area still not too much work has been explored yet. Scalability issues in cloud computing paradigm and security concerns at different levels of cloud computing are the topics of interest and focused for its future work.

REFERENCES

1. Fang Liu, Jin Tong, Jian Mao, Robert Bohn, John Messina, Lee Badger and Dawn Leaf,"NIST Cloud Computing Reference Architecture", NIST Special Publication 500-292, September 2011.
2. PankajSareen, "Cloud Computing: Types, Architecture, Applications, Concerns, Virtualization and Role of IT Governance in Cloud", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 3, March 2013 ISSN: 2277 128X
3. K C Gouda, AnuragPatro, Dines Dwivedi, NagarajBhat, "Virtualization Approaches in Cloud Computing", International Journal of Computer Trends and Technology (IJCTT) – volume 12 Issue 4–June 2014 ISSN: 2231-5381
4. Tania Lorido-Botran _, Jose Miguel-Alonso , Jose A. Lozano, "A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments", ARTICLE in JOURNAL OF GRID COMPUTING · DECEMBER 2014, Impact Factor: 1.51 · DOI: 10.1007/s10723-014-9314-7.
5. M.Kriushanth, L. Arockiam and G. JustyMirobi,"Auto Scaling in Cloud Computing: An Overview", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013, ISSN (Print) : 2319-5940,ISSN (Online) : 2278-1021.
6. ChenhaoQu, Rodrigo N. Calheiros, and RajkumarBuyya,"A Reliable and Cost-E_cient Auto-Scaling System for Web Applications Using Heterogeneous Spot Instances", Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computing and Information Systems, The University of Melbourne, Australia, September 17, 2015.
7. KhosroMogouie, MostafaGhobaeiArani, MahboubehShamsi, "A Novel Approach for Optimization Auto-Scaling in Cloud Computing Environment", *I. J. Computer Network and Information Security*, 2015, 11, 46-53 Published Online October 2015 in MECS.
8. R.Anandhi, K. Chitra,"A Challenge in Improving the Consistency of Transactions in Cloud Databases - Scalability", International Journal of Computer Applications (0975 – 8887) Volume 52– No.2, August 2012.
9. Jorge M. Londoño-Peláez,Carlos A. Florez-Samur,"An Autonomic Auto-scaling Controller for Cloud Based Applications", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 9, 2013.
10. HaniehAlipour,YanLiu,AbdelwahabHamou-Lhadj,"Analyzing Auto-scaling Issues in Cloud Environments",.
11. Che-Lun Hung, Yu-Chen Hu and Kuan-Ching Li,"Auto-Scaling Model for Cloud Computing System",International Journal of Hybrid Information Technology Vol. 5, No. 2, April, 2012.
12. Zhen Xiao, Senior Member, IEEE, Qi Chen, and HaipengLuo,"Automatic Scaling of Internet Applications for Cloud Computing Services", IEEE, Vol. 63, No. 5, May 2014.
13. SushilDeshmukh1, Sweta Kale, "Automatic Scaling Of Web Applications For Cloud Computing Services: A Review", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.
14. Tania Lorido-Botr_an, Jos_e Miguel-Alonso, Jos_e A. Lozan,"Auto-scaling Techniques for Elastic Applications in Cloud Environments", September 5, 2012.
15. Maram Mohammed Falatah, Omar Abdullah Batarfi, "Cloud Scalability Considerations",International Journal of Computer Science & Engineering Survey (IJCSES) Vol.5, No.4, August 2014.
16. 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.
17. DivyakantAgrawal, Amr El Abbadi, Sudipto Das, and Aaron J. Elmore, "Database Scalability, Elasticity, and Autonomy in the Cloud".
18. PranaliGajjar, Brona Shah,"Survey on Different Auto Scaling Techniques in Cloud Computing Environment", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015, ISSN (Online) 2278-1021 ISSN (Print) 2319 5940.