ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

Auto Scaling in Cloud Computing: *An Overview*

M.Kriushanth[1], L. Arockiam[2] and G. Justy Mirobi[3]

Research Scholar, Department of Computer Science, St. Joseph's College (Autonomous) Tiruchirappalli, Tamilnadu[1]

Associate Professor in Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamilnadu, India[2]

Lecturer in Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[3]

Abstract: Cloud computing is a recent technology to provide resources from the large data centers. Cloud Computing is made available as a service to the users. It has become prominent IT to start a business or to utilize the resources without any capital investment. Cloud services are 'pay-per-use' over the internet. It is on demand access to virtualized IT services and products. Rackspace, Salesforce, Amazon, Google, IBM, Dell and HP are the well known service providers. Cloud services are chargeable, service providers charging the users using on demand service policy. In order to provide the excellent service, service providers have to improve the scalability factor. In recent trends, the providers use the auto scaling mechanism to scale the resources according to the users need. The aim of this paper is to give an overview of cloud computing and it emphasize the auto scaling.

Keywords: Cloud Computing, Virtualization, Scalability, Auto Scaling

I. INTRODUCTION

Cloud computing is a paradigm that focuses on sharing data and computations over a scalable network of nodes, spanning across end user computers, data centers, and web services. A scalable network of such nodes form a cloud. An application based on these clouds is taken as a cloud application. In recent years, most of the software, hardware and networking have grown, specially service-based cloud computing has changed the traditional computer and its centralized storage. It has tremendous potential to empowerment, agility, multi-tenancy, reliability, scalability, availability, performance, security and maintenance. The US National Institute of Standards and Technology (NIST) defines cloud computing as follows: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., Networks, servers, storage, applications, and services) that can be rapidly provisioned and released with a minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three delivery models, and four deployment models" [1].

This paper is organized as follows: Section 2 gives an overview of cloud computing. Section 3 describes the concept of scalability and types of scalability in cloud computing. Section 4 presents Auto Scaling and its features. Section 6 provides related works in cloud computing. Section 6 presents the challenges and issues. Finally, section 7 is provided with the conclusion.

II. CLOUD COMPUTING

A. Essential Characteristics

The five essential characteristics in cloud computing are On-demand self-service, Broad network access, Resource pooling, Rapid elasticity and Measured Service.

B. Cloud Service Models

There are three basic service models existing in cloud to provide the resources to the user. Recently the other service models also in action. Fig 1 shows an example of the basic cloud service models.

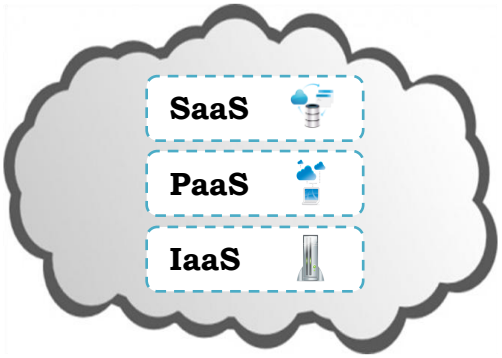

Fig. 1. Cloud Service Models

1. Software as a Service (SaaS): The user is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

through a thin client interface such as a web browser (e.g.,Web-based email, Google Docs).

2. Platform as a Service (PaaS): The user is to deploy onto the cloud infrastructure constomer-created or acquired applications created using programming languages and tools supported by the provider (e.g., Google App Engine, Microsoft Azure).

3. Infrastructure as a Service (IaaS): The user is to provision processing, storage, networks, and other fundamental computing resources from the service providers (e.g.., Amazon Web Services).

4. Network as a Service (NaaS): This is a category of cloud services where the capability provided to the cloud service user is to use network connectivity services and inter-cloud network connectivity services. NaaS involves the optimization of resource allocations by considering network and computing resources as a unified whole.

5. Anything as a Service (XaaS): Xaas is a common term and includes number of things. It can be used as "X as a Service ", "Anything as a Service" and "Everything as a Service". The most common examples of XaaS are Storage as a Service, Communication as a Service, Monitoring as a Service and Failure handling as a Service.

C. Deployment Models

Cloud computing services can be deployed in different methods depending on the organization and location. Four deployment models are usually distinguished, namely Private, Public, Community and Hybrid cloud.

1. Private Cloud: The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise. Fig. 2 is an example of private cloud.

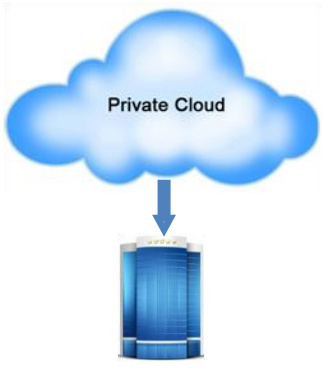

Fig. 2. Private Cloud

2. Public Cloud: The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services. Fig. 3 shows an example of public cloud.

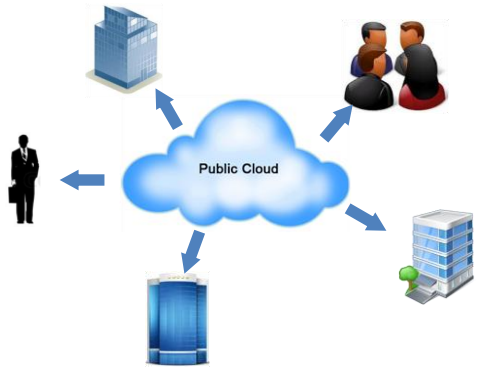

Fig. 3. Public Cloud

3. Community Cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., Mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise. Fig. 4. Represents an example of community cloud.

Fig. 4. Community Cloud

4. Hybrid Cloud: The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

application portability (e.g., cloud bursting for load balancing between clouds). Fig. 5 is an example of hybrid cloud.

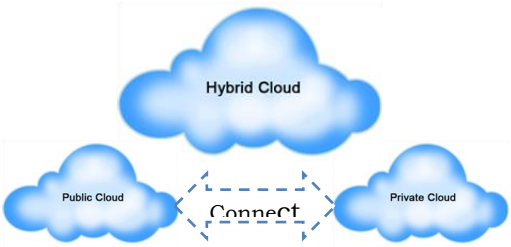

Fig. 5. Hybrid Cloud

III. CLOUD SCALABILITY

Cloud scalability has two dimensions, namely horizontal cloud scalability and vertical cloud scalability [2].

A. Horizontal Cloud Scalability

Horizontal cloud scalability is the ability to connect multiple hardware or software entities, such as servers, so that they work as a single logical unit. It means adding more individual units of resource doing the same job. In the case of servers, you could increase the speed or availability of the logical unit by adding more servers. Instead of one server, one can have two, ten, or more of the same server doing the same work. Horizontal scalability is also referred to as scaling out, which is shown in Fig. 6.

Fig. 6. Horizontal Scalability

B. Vertical cloud scalability:

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

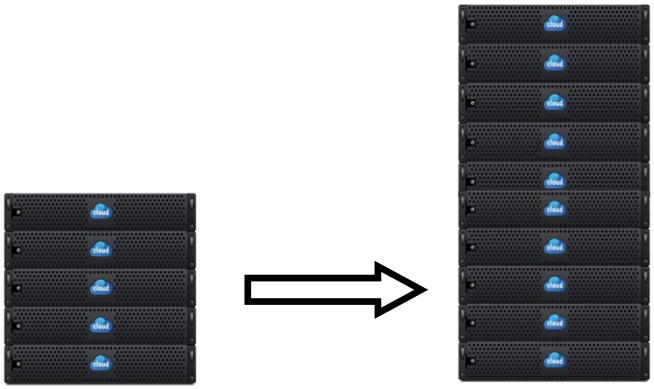

Fig. 7. Vertical Scalability

IV. CLOUD AUTO SCALING: OVERVIEW

Auto Scaling is the ability to scale up or down the capacity automatically according to conditions of the user define. With Auto Scaling ensure that the number of instances is increasing seamlessly during demand spikes to maintain performance, and decreases automatically during demand reduce to minimize costs [3]. The auto scaling in cloud infrastructure is shown in Fig. 8.

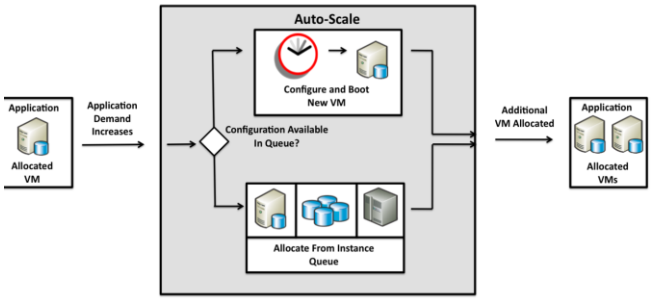

Fig. 8. Auto Scaling in a Cloud Infrastructure [4]

Auto scaling has wonderful features,

- Scale out instances seamlessly and automatically when demand increases.
- Shed unneeded cloud instances automatically and save money when demand subsides.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

- Replace unhealthy or unreachable instances to maintain higher availability of your applications.
- Run On-Demand or Spot instances, including those inside your Virtual Private Cloud (VPC) or High Performance Computing (HPC) Clusters.

V. RELATED WORKS

Harish Ganesan et al [5], present the uses of auto scaling in Amazon cloud. They proposed an architecture how the auto scaling technique works and the tools which are used to identify the cloud peak situations in Amazon cloud. To face the peak situation, in Amazon cloud using the elastic load balancer. Here they found out the problem that the time taken to start the auto scaling and the valid malicious traffic.

Ming Mao et al [6], presented an approach whereby the basic computing elements are virtual machines (VMs) of various sizes/costs, jobs are specified as workflows, users specify performance requirements by assigning deadlines to jobs, and the goal is to ensure all jobs are finished within their deadlines at minimum financial cost. The ultimate aim is to dynamically allocating/deallocating VMs and scheduling tasks are the most cost-efficient instances. To evaluate their approach in representative cloud workload patterns and shows the cost saving from 9.8% to 40.4% compared to other approaches.

Brian et al [7], presented a model-driven engineering approach to optimizing the configuration, energy consumption and operating cost of cloud auto-scaling infrastructure to create greener computing environments that reduce emissions resulting from superfluous idle resources. They provided four contributions to the study of model-driven configuration of cloud auto-scaling infrastructure by i) explaining how virtual machine configurations can be captured in feature models, ii) describing how these models can be transformed into constraint satisfaction problems (CSPs) for configuration and energy consumption optimization, iii) showing how optimal auto-scaling configurations can be derived fromthese CSPs with a constraint solver, and iv) presenting a case-study showing the energy consumption/cost reduction produced by this model-driven approach.

Ruiqing et al [8], proposed a global performance-to-price model based on game theory, in which each application is considered as a selfish player attempting to guarantee QoS requirements and simultaneously minimize the resource cost. They applied the idea of Nash equilibrium to obtain the appropriate allocation, and an approximated solution is proposed to obtain the Nash equilibrium, ensuring that each player is charged fairly for their desired performance. Each player maximizes its utility independently without considering the placement of virtual machines. Then based on the initial allocation, each player reaches its optimal placement solely without considering others' interference.

Roy et al [9], made three contributions to overcome the general lack of effective techniques for workload forecasting and optimal resource allocation. Firstly, it discusses the challenges involved in auto scaling in the cloud. Secondly, it develops a model-predictive algorithm for workload forecasting that is used for resource auto scaling. Finally, the empirical results are provided that demonstrate that resources can be allocated and deallocated by our algorithm in a way that satisfies both the application QoS while keeping operational costs low.

Ching et al [10], developed an auto-scaling system, WebScale, which is not subject to the aforementioned constraints, for managing resources for Web applications in data centers. They also compared were the efficiency of different scaling algorithms for Web applications, and devise a new method for analyzing the trend of workload changes. The experiment results demonstrate that WebScale can keep the response time of web applications low even when facing sudden load changing.

Thepparat et al [11], proposed to simulate feasibility of using virtualization technology to auto-scaling problem in cloud computing. It uses ARENA simulation software to build two different models. There are auto-scaling without server virtualization and auto-scaling with server virtualization. The results of this experiment show that employing virtualization technology increases both life time of servers and CPU utilization.

Ciciani et al [12], proposed that the key design choices underlying the development of Cloud-TM's Workload Analyzer (WA), a crucial component of the Cloud-TM platform that is change of three key functionalities: aggregating, filtering and correlating the streams of statistical data gathered from the various nodes of theCloud-TM platform, building detailed workload profiles of applications deployed on the Cloud-TM platform,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

characterizing their present and future demands in terms of both logical and physical resources, triggering alerts in presence of violations (or risks of future violations) of pre-determined SLAs.

Venugopal et al [13], introduced a system that uses the Amazon EC2 service to automatically scale up a software telephony network in response to a large volume of calls and scale down in normal times. They demonstrate the efficacy of this system through experiments based on real-world data.

Gandhi et al [14], presented the design and implementation of a class of Distributed and Robust Auto-Scaling policies (DRAS policies), for power management in compute intensive server farms. Results indicate that the DRAS policies dynamically adjust server farm capacity without requiring any prediction of the future load, or any feedback control. Implementation results on a 21 server test-bed show that the DRAS policies provide near-optimal response time while lowering power consumption by about 30% when compared to static provisioning policies that employ a fixed number of servers.

VI. CHALLENGES AND ISSUES

Even though cloud computing is an emerging technology, the research on cloud computing is an early stage. New challenges keep on rising in cloud computing. In this section, we address some of the emerging research challenges in cloud computing that relates to auto scaling.

- The time taken to start the auto scaling is up to 3 minutes.
- The provider cannot differentiate the valid and malicious traffic.
- Auto scaling won't apply for all applications.
- Badly configured auto scaling will increase the cost of infrastructure and creates unnecessary capacity.
- Cloud workload patterns shows the cost saving only 9.8% to 40.4% compared to other approaches.
- Auto Scaling mainly focused to reduce the Cost, Energy, High availability and QoS.
- The auto scaling concepts are used in various aspects related to cloud computing not in the auto scaling problem.

VII. CONCLUSION

In this paper we discussed various issues of auto scaling. Considering auto scaling needed to know the present mechanism and the tolls used in auto scaling. Here we have discussed the auto scaling techniques and the related works. We found that in auto scaling there are more ways to do research in different levels. A deeper study on auto scaling approaches to deal with different scalability issues related to the cloud should be the focused of future work.

REFERENCES

[1] Peter Mell and Timothy Grance, "The NIST Definition of Cloud Computing", *National Institute of Standards and Technology Gaithersburg*, Special Publication 800-145, January 2011.

[2] Lijun Mei, W.K. Chan and T.H. Tse, "A Tale of Clouds: Paradigm Comparisons and Some Thoughts on Research Issues", *IEEE Asia-Pacific Services Computing Conference*, 2008, pp 464-469.

[3] "Amazon Auto Scaling in Cloud Computing", http://aws.amazon.com/autoscaling/30.05.2012

[4] Brian Dougherty, Jules White and Douglas C. Schmidt, "Model-driven Auto-scaling of Green Cloud ComputingInfrastructure", *Institute for Software Integrated Systems*, Vanderbilt University, Nashville, November 1, 2010.

[5] "Amazon Web Services Auto Scaling", http://www.slideshare.net/8KMiles/cloud-computing-autoscaling-amazon-ec2-4829409.

[6] Ming Mao and Marty Humphrey, "Auto-scaling to minimize cost and meet application deadlines in cloud workflows", *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, ISBN: 978-1-4503-0771-0.

[7] Brian Dougherty, Jules White and Douglas C. Schmidt, "Model-driven auto-scaling of green cloud computing infrastructure", *International Journal of Future Generation Computer Systems*, Volume 28 Issue 2, February, 2012, pp 371-378.

[8] Ruiqing Chi, Zhuzhong Qian and Sanglu Lu, "A game theoretical method for auto-scaling of multi-tiers web applications in cloud", *Proceedings of the Fourth Asia-Pacific Symposium on Internetware*, Article No. 3, 2012, ISBN: 978-1-4503-1888-4.

[9] Nilabja Roy, Abhishek Dubey and Aniruddha Gokhale, "Efficient Autoscaling in the Cloud Using Predictive Models for Workload Forecasting", *Proceedings of the 2011 IEEE 4th International Conference on Cloud Computing*, ISBN: 978-0-7695-4460-1, pp 500-507.

[10] Ching-Chi Lin, Jan-Jan Wu, Jeng-An Lin, Li-Chung Song and Pangfeng Liu, "Automatic Resource Scaling Based on Application Service Requirements", *Proceedings of the 2012 IEEE Fifth International Conference on Cloud Computing*, ISBN: 978-0-7695-4755-8, pp 941-942.

[11] Theera Thepparat, Amnart Harnprasarnkit, Douanghatai Thippayawong, Veera Boonjing and Pisit Chanvarasuth, "A Virtualization Approach to Auto-Scaling Problem", *Proceedings of the 2011 Eighth International Conference on Information Technology: New Generations*, ISBN: 978-0-7695-4367-3, pp 169-173.

[12] Bruno Ciciani, Diego Didona, Pierangelo Di Sanzo, Roberto Palmieri, Sebastiano Peluso, Francesco Quaglia and Paolo Romano, "Automated Workload Characterization in Cloud-based Transactional Data Grids", *Proceedings of the 2012 IEEE 26th International Parallel and*

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

Distributed Processing Symposium Workshops & PhD Forum, ISBN: 978-0-7695-4676-6, pp 1525-1533.

[13] Srikumar Venugopal, Han Li and Pradeep Ray, "Auto-scaling Emergency Call Centers use Cloud Resources to Handle Disasters", *IEEE Computer Society*, 2011, ISBN: 978-1-4577-0103-0.

[14] Anshul Gandhi, Mor Harchol-Balter, Ram Raghunathan and Michael A. Kozuch, Distributed, "Robust Auto-Scaling Policies for Power Management in Compute Intensive Server Farms", *IEEE Computer Society*, 2011, ISBN: 978-0-7695-4650-6.

BIOGRAPHY

Kriushanth. M is doing research in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 1.6 years in teaching. He has attended many International and National Conferences, Seminar and Workshops. His area of research is Cloud Computing. He is presently working on Scalability issues in Cloud Computing. His area of interest Computer Networks, Software Engineering and Web Technologies.

Justy Mirobi. G is working as a lecturer in Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. She has 7 years of teaching experience. Her research focuses on many aspects of Cloud Computing, Cloud Scalability, Challenges and Issues and area of interests include Software Engineering, Database and Artificial Intelligence.

Dr. Arockiam. L is working as Associate Professor in the Department of Computer Science, St.Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 24 years of experience in teaching and 17 years of experience in research. He has published more than 140 research articles in the International / National Conferences and Journals. He has also presented 2 research articles in the Software Measurement European Forum in Rome. He has chaired many technical sessions and delivered invited talks in National and International Conferences. He has authored a book on "Success through Soft Skills". His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining and Mobile Networks. He has been awarded "Best Research Publications in Science" for 2010, 2011, & 2012, "Best Teacher Award" for 2012-13 and ASDF Global Awards for "Best Academic Researcher" from ASDF, Pondicherry for the academic year 2012-13.