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.

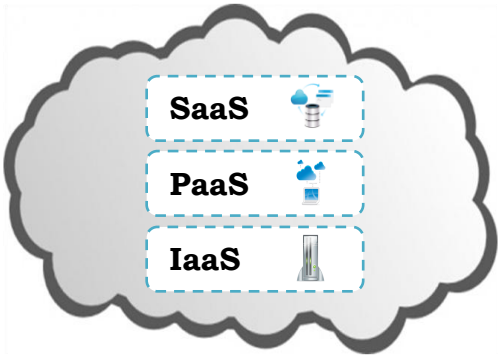

Fig. 1. Cloud Service Models

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