International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Analysis of Scalability Factor in Cloud Computing

Deepali Saini*, Anamika Pandey

Department of MCA, IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

ABSTRACT

Cloud computing is a modern technology which makes available resources from the large data centers. Cloud Computing offered to the users as a service. Now a days, cloud computing is a well known IT which aids in data mining. Different domain utilizes this service to start a business or to utilize the resources without any capital investment. Over the internet, cloud services are "pay-per-use". Microsoft Corporation, Google, Amazon, IBM, VMware, Oracle Corporation, Dell, Rackspace, Salesforce, and HP are the eminent service providers. Cloud services are chargeable. It is an on demand service policy where users charged as per the use. In order to provide the excellent service, one of the major areas of improvement is scalability. In latest 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 on the factor of auto scaling.

Keywords: Cloud Computing, Reliability, Scalability, Auto Scaling, Virtualization.

I. INTRODUCTION

"Cloud computing" is the future of data mining. It is the advancement in on-demand information technology services and products. Considerably, the foundation of cloud computing will be based on the virtualized resources. The term cloud computing [1] was globally accepted in October 2007 when IBM and Google founded a partnership in this field, followed by the IBM's announcement of the "Blue Cloud" effort. Cloud computing faces several challenges and issues like interoperability, security and Privacy, freedom, performance, auto-scalability, reliability, costing model. This paper discusses the concept of cloud computing and a major issue of auto-scalability and its implementation available today.

With the [2] swift enlargement of processing and storage technologies along with the success of the Internet, computing resources have happened to be cheaper, more potent and more universally available than ever before. This technological trend has allowed the recognition of a new computing model called cloud computing, where resources made available as general utilities and users can lease and release through the internet in an on demand fashion. In a cloud-computing environment, the traditional role of service provider are divided into two: the infrastructure providers who manage cloud platforms and lease resources according to a usage-based pricing model, and service providers, who rent resources from one or many infrastructure providers to serve the end users. Since the advancement of cloud computing, a tremendous impact on the Information Technology (IT) industry was globally witnessed, where large companies such as Google, Amazon and Microsoft attempt to offer more powerful, reliable, secure and cost-competent cloud platforms, and business enterprises look for new designs of their business models to expand the list of benefit from this new model. Certainly, cloud computing provides numerous compelling attributes which attracts business owners.

These days, companies are spreading their business globally. They do not limit themselves in conducting business in one country. They need dynamic, elastic, scalable cloud computing platform that operates around-the-clock. Full functionality, adaptability, non-stop availability and reduced cost are the major requirements that were expected from cloud computing services.

In cloud computing environment, the virtualization technology plays an important role by providing the physical resources like disk storage, processor, and broadband network. Virtualization generally refers to platform virtualization, or the abstraction of physical resources for users. In the cloud, these physical resources are regarded as a "resource pool", and can be allocated on demand by user.

Cloud computing [3] is driven by tangible and powerful benefits. The following are the features of cloud environment:

1) Computing power is elastic only if workload is
2) Parallelizable.
3) Data is stored at an untrusted host.
4) Data is replicated across large geographic distances.
5) Hard to maintain ACID while data is replicated over large geographic distances.

Key Benefits of Cloud Computing:

1) Reduced IT operating costs (25%)
2) Shifting Capital Expenses to Operating Expenses
3) Increased efficiency (55%)
4) Agility

II. SCALABILITY

Scalability is a vital property of a system, which signifies its capability to handle growing amounts of work in either an elegant manner or its ability to get better throughput when additional resources (typically hardware) added. If the system performance improves after adding hardware in proportion of the capacity, it is said to be scalable. Similarly, an algorithm is said to scale if it is suitably efficient and practical when applied to large situations (e.g. a large input data set or large number of participating nodes in the case of a distributed system). If the algorithm fails to perform when the resources increase then it does not scale.

A system can be scaled by adding hardware resources. Cloud computing can be categorized into two types: namely vertical cloud scalability and horizontal cloud scalability [1].

The first type is when the system scales *vertically* and is known as *scale-up*. In vertical scaling the resources are added to a single node in a system, typically involving the addition of processors or memory to a single computer. As a result of such vertical scaling the existing systems becomes able to use virtualization technology more effectively, as it provides more resources for the hosted set of operating system and application modules to share. For example, adding processing power to a server to make it faster. Vertical scaling is limited by the fact that you can only get as big as the size of the server.

Airbnb, the popular peer-to-peer online marketplace and home stay network infrastructure serves 50M users with over 800,000 listings in 33,000 cities and 192 countries. According to Mike Curtis, VP of engineering at Airbnb, we learned that the Airbnb site and internal machine learning services are operated by a pool of 5000 EC2 instances which scale according to demand.

Startups and even individual developers all start out with a small service running from their laptops, hoping that one day the services and applications they offer will conquer the world and become the most popular thing on the Internet. On meeting that day a single machine or even a single cluster simply won't be able to handle such a large workload. If the business plans to run the application on an increasingly large scale, they need to think about scaling in cloud computing from the start, as part of the planning process.

The other type of scaling a system is *horizontally* known as *scale-out*, achieved by adding hardware resources. More nodes *are added* to a system to scale horizontally (or scale out). By adding a new computer to a distributed software application we can scale out the system. An example might be scaling out from one web-server system to a system with three web servers.
As computer prices drop and performance demand continue to increase, low cost "commodity" systems can be used for building shared computational infrastructures for deploying high-performance applications such as Web search and other web-based services.

Issues and Challenges:

Cloud computing is a rising technology and growing rapidly. But the field of cloud computing still faces several issues and challenges and demands the research. New challenges keep on rising in cloud computing. One of the emerging challenges is auto scaling. Generally the time taken by the system to initiate the auto scaling is up to 3 minutes. The provider can not

differentiate and reach on decision to understand the valid and malicious traffic. If auto scaling in cloud computing is badly configured, it will result in increasing cost of infrastructure and unnecessary capacity is created. The focus of auto scaling is on reducing the cost, energy, high availability and QoS.

III. RELATED WORK

Harish Ganesan et al [4], 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 [5], 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 [6], 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 from these 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 [7], 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 [8], 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 [9], 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 [10], proposed to simulate feasibility of using virtualization technology to autoscaling 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 [11], 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, 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 [12], 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.

IV.CONCLUSION

The ability to scale on demand is one of the biggest advantages of cloud computing. In the area of cloud computing, we review the major challenges of scalability. Planned or unplanned business system outages are the enemies of the successful business in cloud computing environment. Therefore, it demands highly available, reliable and auto-scalable cloud computing systems. Due to growth of internet and fast speed research in IoT, different sectors are set globally and growing intensively. Hence, this paper provides a base for research in the field of auto-scalability. Auto-scalability will be the next strong pillar for the cloud computing to stand.

V. REFERENCES

[1]. "Amazon Auto Scaling in Cloud Computing", http://aws. amazon. com/autoscaling/30. 05. 2012

[2]. Qi Zhang, Lu Cheng, and Raouf Boutaba, "Cloud computing:state-of-the-art and research challenges", Springer, 20 April 2010.

[3]. Dr. K. Chitra, B. Jeeva Ran. "DES:Dynamic and Elastic Scalability in Cloud Computing Database Architecture. " IJACSA, Volume 5 Issue 1.

[4]. "Amazon Web Services Auto Scaling", http://www. slideshare. net/8KMiles/cloud-computing-autoscaling-amazonec2-4829409.

[5]. 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.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. 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.

[10]. 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.

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

[12]. 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.