Evaluation of Auto Scaling and Load Balancing Features in Cloud

Ashalatha R
PhD Scholar, Computer Science & Engg Department, P.D.A. College of Engineering, Kalaburagi, India.

Jayashree Agarkhed
Professor, Computer Science & Engg Department, P.D.A. College of Engineering, Kalaburagi, India.

ABSTRACT
Cloud computing is a latest technology that uses internet and centralized servers to maintain data and various types of applications. Cloud computing allows consumers and business people to use applications without any installation of either hardware or software and accessing their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing storage, memory, processing. The cloud computing system is the newer version of utility computing which has replaced its area at various data centers. The Load balancer determines when to start or end any virtual machine in the Cloud. The auto scaling feature along with the load balancing technique makes anyone easy to automatically increase or decrease back-end capacity to meet traffic fluctuation levels.

Keywords: Cloud computing, Auto scaling, Load balancing.

1. INTRODUCTION
Cloud computing has grown extremely well in business by effectively providing the world class services to all its users. The latest technology has shifted from existing ones to cloud computing. Due to the less resources investment and maintenance cost, the companies are moving towards the cloud. The cloud which operates through the Internetwork protocol has the features of virtualization, grid computing, autonomic and utility computing. It is a general term for anything that involves delivering hosted services over the Internet. It is a pay-go-use model wherein the clients pay for the requested resources. Cloud computing customers have complete access to information technology capabilities and services which is provided through Internet. Cloud computing has brought tremendous change in operations of IT industries. It has greater benefits to the IT industries with less infrastructure investment and maintenance costs.

This article deals with the auto scaling and load balancing features various cloud providers. The remainder of this article is organized as follows. Section 2 presents the related work. Section 3 presents the comparison between different load balancers and auto scaling techniques with respect to different cloud platforms. Section 4 represents the conclusion and future work.

1.1 Infrastructure as a Service:
Infrastructure-as-a-Service (IaaS) is the capability to provision processing, storage, networks, and computing resources [1]. It provisions processing, storage and networks benefits. The major services includes server hosting, web servers, storage, computing hardware, operating systems, virtual instances, load balancing, internet access and bandwidth provisioning. The characteristics of IAAS include resource distribution and dynamic scaling capability. IAAS providers offer Load Balancing technique by automatic scaling facility which sets conditions for scaling up and down of applications. This service requires high Internet bandwidth capacity, low-latency, Reliable and low cost communication.

1.2 Load Balancing:
Load Balancing is a technique to redistribute the load across the nodes. The decision to balance load is made locally by a node, based on its current utilization. Each node continuously measures its resource utilization of CPU, memory, network consumption and disk space.

1.3 Auto Scaling:
Auto scaling technique provides on-demand resources availability based on certain workloads in cloud computing systems. The Auto scaling service allows the configuration of capacity management policies applied to dynamically decide on acquiring or releasing resource instances for a given application.

2. RELATED WORK
Dynamic resource provisioning can be solved by the method of fine grained scaling solution for energy efficiency in cloud data centers. Optimal configuration is taken care to minimize the energy consumption and satisfies various performance objectives [2]. A flexible load balancing traffic grooming strategy is maintained for system optimization. A Traffic Engineering optimization strategy in the overlay layer is used to optimize the overall performance of the system [3]. The resource utilization can be achieved and response time of tasks is reduced using Max-Min Task scheduling algorithm which consists of Task Status Tables and Virtual Machine Status Tables and its update and task allocation algorithm in elastic cloud [4]. A QoS-Aware Resource Elasticity (QRE) framework makes assessment of application behavior and develops mechanisms for dynamic scalability of cloud resources which hosts application components. The cloud hosted multi-tier web applications consists of three tiers which are known as presentation or web-tier, business or application tier and database tier. The experiments and analysis show that algorithm known as multi tier performance model called as ‘MT-Perf Mod’ and per tier resource elasticity called as ‘MT-ResElas’ Models [5].

An auto scaling method can be used for allocating resources in hybrid cloud environment for all types of user requirements on SLA. The Service Architecture of Auto Scaling framework defines for sub-modules to perform auto-scaling tasks. SLA Driven VM Auto-Scaling Algorithm consists of Run-time Scaling and SLA monitoring and performance-oriented scheduling mechanisms [6]. A novel server-side auto scaling mechanism to allocate virtual resources on cloud for real time tasks has been proposed and the functional concepts in Auto-Scaling Mechanism include Monitor, Analyzer, Planner and Executer [7]. The auto-

scaling method describes the execution of an application within deadline. The Auto-Scaling Algorithm includes Run-time Scaling and Performance-oriented Scheduling Algorithm [8]. Various Auto-Scaling Strategies using log-traces of Google data center clusters include Auto-scaling Demand Index (ADI) metric for auto-scaling strategy. The Adaptive strategy for defining step sizes in auto-scaling operations include two step size configuration strategies which are fixed (for regular metrics) and Adaptive for irregular and peaky system utilization. Various Auto-scaling triggering strategies include Reactive, Conservative and Predictive methods [9].

3. COMPARISON OF AUTO SCALING AND LOAD BALANCING FEATURES WITH VARIOUS CLOUD PROVIDERS

3.1 Auto scaling in commercial cloud:

3.1.1 AMAZON

Amazon Web Service (AWS) provides compute and storage servers with high speed networks for accessing any type of resources. Amazon provides auto scaling service as IaaS EC2 (Elastic compute cloud) public cloud. EC2 provides an elastic IP address with every user account to reduce the instance failures. Auto scaling in AWS allows increasing or decreasing the number of EC2 instances within the application's architecture. With Auto scaling, one can create collections of EC2 instances called as Auto scaling groups. We can also specify minimum and maximum number of instances in each Auto Scaling group. Each auto scaling group contains one or more scaling policies which define when auto scaling launches or terminates EC2 instances within the group. Auto scaling in AWS uses load balancers to distribute traffic across the instances within auto scaling technique along with the elastic load balancing technique [10].

3.1.2 MICROSOFT AZURE

Platform-as-a-Service (PaaS) clouds offer a runtime environment system where users' components can be deployed and executed in a straightforward manner which offers an additional abstraction level when compared to IaaS clouds [11]. The users need not have to handle virtual resources such as machines or networks to start running their systems. Microsoft Windows Azure does not implement any embedded auto scaling solution to its users rather it supports *Paraleap* software which automatically scales resources in Azure to respond to changes on demand [12]. Data storage for application scheduling and rules based on customer performance counters is an added advantage in Windows Azure platform which is not available in other cloud providers [13].

Table 1: Auto Scaling Techniques Used By Various Cloud Providers

Cloud Providers	Auto scaling feature
AMAZON	Automatically scales number of EC2 instances for different applications.
WINDOWS AZURE	Provides auto scaling feature manually based on the applications.
GOOGLE APP ENGINE	Owns auto scaling technology Google applications.
GOGRID	Supports auto scaling technique in programmatic way and does not implement it.
FLEXISCALE	Provides auto scaling mechanism with high performance and availability.
ANEKA	Application management service through cloud peer service.
NIMBUS	Open source cloud provided by resource manager and Python modules.
EUCALYPTUS	Open source cloud which provides wrapper service for various applications.
OPEN NEBULA	Open source cloud which provides OpenNebula Service Management Project.

The survey on auto scaling mechanisms with different commercial cloud providers and open source cloud platforms are shown in Table 1 [23].

3.1.3 GOGRID

Cloud does not implement any auto scaling functionality but does provide an API to remotely command the addition or removal of virtual machines whenever required. It uses *RightScale* software which is a cloud management platform which offers control functionality over the virtual machines deployed in different cloud platforms [14]. GoGrid supports auto scaling functionality based on alerts and associated actions to run each time an alarm is triggered [15].

3.1.4 RACKSPACE

Rackspace does not support built in auto scaling capabilities but gives an API to its users for remote control of the hosted virtual machines. The user is solely responsible for monitoring the service and taking the scaling decisions as and when necessary. The creation and removal of resources is done through API calls to the remote API's [16]. Rackspace provides *Enstratus* cloud management platform which offers control functionality over the VMs deployed onto different clouds. *Enstratus* software also supports auto scaling feature as that of other cloud platforms does [17].

3.2 Load balancing in commercial cloud:

3.2.1. AMAZON:

Amazon EC2 offers load balancing through Amazon Elastic Load Balancing service (ELB). ELB technique provides high availability of EC2 instances and enhances EC2 applications availability by distributing incoming application traffic across multiple instances [18]. EC2 includes OS such as Linux, Windows, Suse Linux, Fedora, Open Solaris, Red Hat, Open Suse, Ubuntu etc. Any user can interact with EC2 using set of SOAP messages. The elastic load balancer provides high availability of EC2 instances and also enhances EC2 application availability by distributing incoming application traffic across multiple instances. Elastic load balancing also detects unhealthy instances and automatically routes the traffic as necessary. Various metrics evaluation can be done through transactions/second, number of simultaneous users, request latency, performance evaluation, QoS evaluation,

energy efficiency, power saving and cost estimation strategy. Amazon EC2 automatically distributes incoming application traffic among multiple instances using ELB feature and monitoring method using Cloud Watch techniques with high scaling policies.

3.2.2 MICROSOFT AZURE:

In Azure, the load is automatically distributed among available work resources by using a round robin algorithm transparent to the cloud users. Load balancing for applications running under the *AppFabric* service is achieved by using hardware load balancers [19]. The load balancers have redundant copies to reduce failure. Windows Azure gives PaaS cloud platforms to its users where SQL is a cloud based version of SQL servers and *Azure AppFabric* is a collection of services for cloud applications. Windows Azure has three components namely compute, storage and fabric controller. The Fabric Controller ensures scaling, load balancing and memory management and reliability features [20].

3.2.3 GOGRID:

GoGrid cloud has free account usage of the redundant f5 hardware load balancers. The load balancers check the availability of nodes in the balancing pool. If one node becomes unavailable, the load balancer removes it from the pool automatically. Load balancing can disturb client sessions if traffic for the same session is not routed to the same server node that initiated the session throughout the whole duration of the session [21].

3.2.4 RACKSPACE:

Rackspace cloud offers two types of cloud services which are *Cloud Servers* and *Cloud Sites.* The Cloud Servers is a IaaS type service which takes care of auto scaling and load balancing through cloud clients. The Cloud Sites service targets automated scaling, load balancing and daily backups. The algorithm used for distributing the load is round robin. Pricing is done based on the client's platform usage in terms of disk space, bandwidth and compute cycle which is a unit that enables Rackspace to quantify their platform's computational usage. Users are not charged for use of load balancing service. Google provides free of cost to its users the various services such as Gmail, Google Drive, Google Calendar, Picasa, Google groups [22].

Table 2: Load Balancing Techniques Used By Various Cloud Providers

Cloud Providers	**Load Balancing Feature**
AWS	Load Balancing service will allow users to balance incoming request & traffic across multiple EC2 instances.
AZURE	The load automatically distributed among available work resources using round robin algorithm transparent to the cloud users.
GOGRID	Load Balancing algorithm is used as Round Robin, Sticky session, SSL least connect, Source address.
FLEXISCALE	Load Balancing does automatic equalization of server load within clusters using Zeus software.
MOSSO	This service scales with traffic and inherits Load Balancing feature.
ANEKA	Dot Net based service oriented resource management and development platform.

Cloud computing systems are commercially available through several cloud providers. Extensive survey has been done through several websites documentation. Table 2 gives the comparison survey about the various cloud providers available in the market today [24].

4. CONCLUSION & FUTURE WORK

Auto scaling and load balancing features are the two methods which assure service level objectives in cloud computing era. Various factors affect the cloud services from different cloud providers' point of view. This paper has aimed the best to compare both the feature with respect to leading cloud platforms. The next work includes implementation of load balancing and auto scaling features in real time cloud environments.

5. REFERENCES

[1] Dan C. Marinescu, Cloud Computing Theory and Practice, Morgan Kaufmann, USA, Elsevier, 2013.

[2] S.K. Tesfatsion, E. Wadbro, J.Tordsson, "A combined frequency scaling and application elasticity approach for energy-efficient cloud computing," Future Generation Computer Systems 2014, pp. 205-214.

[3] Qiao hong and Yan Shoubao, "A flexible load-balancing traffic grooming algorithm in service overlay network," In proceeding of the International conference on cloud computing and big data, 2013.

[4] X.Li, Y.Mao, X.Xiao, Y.Zhuang, "An improved max-min task-scheduling algorithm for elastic cloud," In proceeding of the International symposim on computer, consumer and control, 978-1-4799-5277-9/14, IEEE 2014.

[5] P.D. Kaur, I.chana, "A resource elasticity framework for QoS aware execution of cloud applications," Future Generation Computer Systems 2014, pp. 14-25.

[6] H.Kang, J. Koh, Y.Kim, J.Hahm, "A SLA driven vm auto scaling method in hybrid cloud environment," APNOMS IEICE 2013.

[7] Y.W. Ahn, A.M.K cheng, J.Baek, M.Jo and H.chen, "An auto-scaling mechanism for virtual resources to support mobile, pervasive, real-time healthcare applications in cloud computing," 0890-8044/13, IEEE 2013.

[8] Y. Ahn, J.Choi, S. Jeong, Y.Kim, "Auto scaling method in hybrid cloud for scientific applications," IEICE – Asia-Pacific Network Operation and Management Symposium (APNOMS) 2014.

[9] Marco.A.S. Netto, C. Cardonha, R.L.F. Cunha, M.D. Assuncao, "Evaluating auto-scaling strategies for cloud computing environment," In proceeding of the 22nd International MASCOTS, 1526-7539/14, IEEE 2014.

[10] Amazon Web Services. http://aws.amazon.com/

[11] Windows Azure. http://www.windowsazure.com/

[12] Paraleap. https://www.paraleap.com

[13] L.R. Sampaio, "Towards practical auto scaling of user facing applications," LatinCloud, IEEE 2012.

[14] RightScale, http://www.rightscale.com/

[15] GoGrid, http://www.gogrid.com/

[16] Rackspace http://www.rackspace.com/

[17] Enstratus. http://www.enstratus.com/

[18] Amazon Elastic Load Balancing Developer guide 2012. http://aws.amzon.com/elb

[19] E. Caron, L. R. Merino, F. Desprez and A.Muresan, Auto-scaling, load balancing and monitoring in commercial and open-source clouds. [Research Report] RR-7857, 2012, pp.27. <hal-00668713>

[20] Microsoft Azure Appfabric. http://windowsazure.com/appfabric/

[21] R. Buyya, J. Broberg, A.M. Goscinski, Cloud computing: Principles and paradigms, John wiley and sons 2011.

[22] Google App Engine. http:// code.google.com/appengine/

[23] F.L. Ferraris, "Evaluating the auto scaling performance of flexiscale and amazon EC2 clouds", 14th International symposium on symbolic and numeric algorithms for scientific computing, 2012.

[24] R. Ranjan, L. Zhao, X. Wu and A. Liu, "Peer-to-Peer Cloud Provisioning: Service Discovery and Load-Balancing," http://arxiv.org/abs/0912.1905, Dec 2009.