A survey: Resource Utilization Using Load balancing & scaling in Cloud Computing

Vishva Joshi¹, Prof. Umang Thakkar²
¹,²Silver Oak College of Engineering & Technology., S.G.Highway, Gota Ahmedabad, India.

Abstract—Cloud computing is the pool of the heterogeneous resources aim to achieve maximum resource utilization to users with higher availability at minimized cost. Now, a days demand of cloud resources are increases and demand is change day by day. So, to solve the problem of resource utilization required the load balancing, scale–up, scale-down of the resources as per requirement of cloud service users. And for scaling decision there are need check CPU utilization and current requests of users then gives the prediction decision to scale up and scale down scenario.


I. INTRODUCTION

Cloud computing is a general terms for the delivery of hosted services over the internet. Cloud computing build and maintain the infrastructure of the resources. Cloud computing such as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (eg networks, servers, storage, applications and services) that can be rapidly provisioned. In the simple terms, Cloud computing means storing and accessing data over the internet instead of your computer’s hardware. Everything, you need physically close to you about computer we can access them fast and easily[1].

Cloud computing is aggregating from two words. First, is Cloud and Second term is computing. Cloud is the collection of heterogeneous resources and computing means some criteria to use this resources. Like, In minimum cost high utilization. It provides online platform to handle huge amount of work without affecting execution of framework. It is the technique to preform parallel and distributed computing on several different nodes that provide their virtualized hardware with on-demand service request by customer based on service level agreements (SLAs)[3]. Cloud computing is an emerging commercial infrastructure paradigm and in the last decade its usage has gained a lot of popularity. elasticity is beneficiary in terms of cost,obligation of maintaining Service Level Agreements (SLAs) eads to the necessity in dealing with the cost/performance.

Web-based email services like Gmail and Hotmail deliver a cloud computing service: users can access their email “in the cloud” from any computer with a browser and Internet connection, regardless of what kind of hardware is on that particular computer. The emails are hosted on Google's and Microsoft's servers, rather than being stored locally on the client computer. These examples illustrate the different types of cloud computing services available today:

Amazon EC2 - virtual IT
Google App Engine - application hosting
Google Apps and Microsoft Office Online - software as a service
Apple iCloud - network storage

Cloud Computing Characteristics are following:

Resource Pooling: A Cloud Services Provider creates resources that are pooled together in a a system that support multi-tenant usages customization.

Rapid Elasticity: Resources are rapidly and elastic provisioned to the users. Add resources by scaling up and remove by scaling down may be automatically.

Customer can purchase resources services any time and in any quantity.

On-Demand Self-Service: Provide the Cloud services to the customer without personal interaction.
Broad Network Access: Cloud services access by the internet using standard method in manner. So, it is platform independently access anywhere.

Measured Service and Billing: Resources access by users audited and measured on meter system. It can be measured the amount of storage used, number of transaction, network input/output, bandwidth processing power etc. client have to be charged only they used the resources.

II. LOAD BALANCING

Load Balancers allow you to distribute the workload and balance it between two or more Cloud Servers. You can therefore shape your infrastructure to allow it to meet activity spikes, optimize the allocation of resources and ensure a minimal response time. Using a load balancer is recommended in all cases, whether you require one or more of the following:

- Guaranteed service continuity
- Handle high traffic
- Be prepared for sudden request spikes

When using such a setup, the Load Balancer intercepts the network requests and sorts them between two or more Cloud Servers. Thanks to this redundancy, you benefit from high-reliability and stability thus ensuring continuity of the service. Should one of the load-balanced Cloud Servers stop working for any reason, the load balancer will automatically exclude it and redirect the requests to the other servers: this ensures maximum Business Continuity.

Load Balancing Techniques:

There are load balancing has following types

A. Static Algorithm
In static algorithms decision about load balancing is made at compile time. These are limited to the environment where load variations are few. These algorithms are not dependent upon the present condition of system. A static load balancer algorithm divides the traffic equally among the servers. It does not use the system information while distributing the load and is less complex [6].

B. Dynamic Load Balancer
Dynamic algorithm is based on the different properties of the nodes such as capabilities and network bandwidth. This need constant check of the node and are usually difficult to implement [6]. Dynamic algorithms are well suited in cloud computing environment because they distribute work at run time and assign suitable weights to the servers. A lightest weight server is search in network and preferred by this algorithm. Dynamic algorithms are consider more complicated.

III. AUTO-SCALING

Cloud computing with its dynamic scaling feature which allows one to scale, that is to increase or decrease, the amount of resources depending on the demand has become a great boon for IT professionals everywhere. This is especially true for environments with very unpredictable traffic flow, like the whole internet for instance. With traditional servers that were preconfigured to handle a certain amount of load, a website might go down especially when traffic has suddenly surged to levels above the capacity of the server. This happens when some sort of news or event leads people to a specific web location. The solution in a Cloud context is to allocate more resources, and in this case allocate more server instances.

Auto-scaling systems monitor one or more performance metric. In this work we considered workload (i.e., the number of user requests per time unit) as the performance metric. A, Predictor uses performance metric’s current value from Monitor to forecast future performance metric value. The predicted value is sent to the Decision Maker which generates the final scaling decision by considering Cloud Provider Pricing Model. Since the final scaling decision (decision maker’s output) depends on the prediction result, various research studies have focused on improving accuracy of the Predicator’s output.
Prediction accuracy of predictive auto-scaling systems can be increased by choosing an appropriate time series prediction algorithm based on the incoming workload pattern.

The term workload refers to a number of user requests, together with the arrival timestamp [5]. Workload is the consequence of users accessing the application or jobs that need to be handled automatically. There are four workload patterns:

- Stable workload is characterized by constant number of requests per minute. This means that there is normally no explicit necessity to add or remove processing power, memory or bandwidth for changes in workload.
- Growing workload represents a load that rapidly increases.
- Cycle/Bursting workload represents regular periods (i.e. seasonal changes) or bursts of load in a punctual date.
- On-and-off workload represents the work to be processed periodically or occasionally, such as batch processing.

Prediction Techniques for Scaling:
In contrast to other approaches, threshold-based technique is the only approach, which is widely used in the commercial auto-scaling systems. In order to use time-series analysis in auto-scaling area, a performance metric (such as number of requests per time unit) is periodically sampled at fixed intervals.

A. Support Vector Machine
The purpose of this sub-section is to provide a brief description of the mathematical foundation of SVM. Readers are encouraged to see [17] for more details on SVM. Support Vector Machine (SVM) is used for many machine learning tasks such as pattern recognition, object classification, and regression analysis in the case of time series prediction. Support Vector Regression, or SVR, is the methodology by which a function is estimated using observed data, which in turn “trains” the SVM.

B. Neural Networks
A neural network (NN) is a two-stage regression or classification model, typically represented by a network diagram [19]. Some of the most commonly used neural network classifiers (algorithm) are: feed-forward, backpropagation, time delay, and error correction neural network classifier.

IV. LITERATURE REVIEW
In paper[1] An Autonomic-Computing based Architecture, proposes an elastic architecture for cloud computing based on concepts of Autonomic Computing. Its design based on experiments using micro benchmarks. Automatic elasticity implements by following steps:
1. Collect data from metrics.
2. Implementation of the rule verification mechanism.
3. Implement a mechanism for prediction

Advantage:
Using micro benchmarks applied on both private and hybrid cloud computing, automatic computing work well in elasticity provisioning.

Drawbacks:
Some validation architecture & technologies for autonomic computing.

This paper presents a novel resource provisioning policy that can find the most cost optimal setup of variety of instances of cloud that can fulfill incoming workload. The model takes lifetime of each running instance into account while trying to find the optimal setup using Benchmark experiments.

Advantage:
Find the most optimal setup of instances that fulfils incoming workload and minimizes the resource cost. The presented LP model finds the optimal setup of each task/component in a workflow/SoA application at each run.

Drawbacks:
Network bandwidth can be added in to this model for scaling.

In paper[3] A Novel Hybrid Scheduling Algorithm,
In this paper, propose a hybrid scheduling algorithm for load balancing distributed environment by combining the methodology of Divide-and-Conquer and Throttled algorithms referred to as DCBT.

Advantage:
DCBT is to reduce the total execution time of the tasks and thereby maximizing the resource utilization. Better execution time compared to Modified Throttled algorithm for the same set of inputs & reduces the execution time by 9.972%.
Drawbacks:
DCBT algorithm with wider perspective and can be experimented with machines with different configurations.

In paper[4] Proactive and Reactive Tuning to Scale, In this paper, we show the limitations of a proactive or reactive scaling system. ProRenaTa guarantees a high level of improving the overall resource utilization. The first approach can scale the system with a good accuracy since scaling is based on observed workload characteristics. The latter approach, on the other hand, is able to prepare the instances in advance.

Advantage:
Efficiency of elasticity controller that combine both proactive and reactive for a auto scaling.

Drawbacks:
Extend data migration model to estimate the cost of handling workload.

In paper[5] Autonomic Auto-Scaling Prediction System, This paper investigates the accuracy of predictive auto-scaling systems in the cloud computing. Monitor, Predictor, and Decision Maker are the main components of a predictive auto-scaling system.
1. Monitor one or more performance metric.
3. Decision Maker which generates the final scaling decision by considering Cloud Provider Pricing Model.

Advantage:
Accuracy of auto-scaling systems can be increased by choosing an appropriate time-series prediction algorithm. Predictive auto-scaling algorithms predict future adjust application resources in advance to meet the future needs.

Drawbacks:
This work we focused on the prediction side of predictive auto-scaling system focused on the prediction side of predictive auto-scaling systems, while the exact impact of increasing prediction accuracy on the final scaling decision is unknown.

V. CONCLUSION
A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Now, a day’s cloud computing demand is increases. So to manage the resources between the users needs the load balancing techniques. DCBT load balancing dynamic load balancing algorithm will distributes the user’s load to the server. User’s requirements are not stable, can be increased or decreased. It is managed by the scaling. Based on the current cpu utilization SVM gives the prediction of the scaling. This SVM scaling prediction and current workload of result both checks and take the final decision of scaling. So, it gives the final scaling decision.

REFERENCES