A Project Report

On

Enhancing and Optimizing Computing using Elastic Load Balancing and Auto Scaling

Submitted by

Sai Vivek Rambha

Roll no: 1605049

Utkarsh Shivam

Roll no: 1605414

Hritik Sahu

Roll no: 1605360

1. Introduction

Cloud computing is a newly progressing technique which offers online computing resources, storage and permits users to organize applications with enhanced scalability, availability and fault tolerance. Cloud computing is about storing the stuff on remote servers instead of on own computers or other devices. This information can be retrieved using the internet with any device, everywhere in the world as long as that device can support cloud computing systems. The cloud computing system is comprised of a front-end, which is the client side and a back-end which is a collection of the servers and computers owned by a third party which stores the data. A central server which is a fragment of the back-end follows protocols and uses middleware to communicate between networked computers. Cloud computing accumulates all the computing resources and manages them automatically. Its characteristics describe a cloud computing system on-need selfservice, pooling of resources and access to the internet. The elasticity of service availability and measurement of services utilized by individual users. Cloud computing is everywhere with tools like Google Drives replacing Microsoft Office, Amazon Web Services replacing traditional enterprise data storage, banking websites replacing branch offices and Dropbox storing all our data and files.

2. Load Balancing

Load balancing is a serious concern in cloud computing. With the increase in attractiveness of cloud computing among users, the load on the servers and the quantity of processing done is surging drastically. There are multiple nodes in the cloud, and due to the random allocation of a request made by the client to any node, the nodes become unevenly loaded. So to avoid the condition where some nodes are either severely loaded or under loaded, the load balancer will evenly divide the workload among all the nodes. Thus load balancing will equally distribute the workload among the nodes, and it can help in minimizing delays in communication, maximizing the throughput, minimizing execution time and maximizing resource utilization.

2.1 Advantages of Load Balancing

1. It should possess fault tolerance.
2. It should be capable of modifying itself according to any changes.

3. It should also maintain system stability.

2.2 Issues of Load Balancing

1. Load balancing becomes critical because, in the middle of execution, the processes may shift amongst nodes to ensure equal workload on the system.
2. For a load balancing scheme to be good it should be scalable, general and stable and should add minimal overhead to the system. These requirements are interdependent.

3. Auto Scaling

Auto Scaling is a method used in cloud computing, whereby the amount of computational resources in a server farm, typically measured in terms of the number of active servers, scales automatically based on the load on the farm. It is closely related to, and builds upon, the idea of Load Balancing.

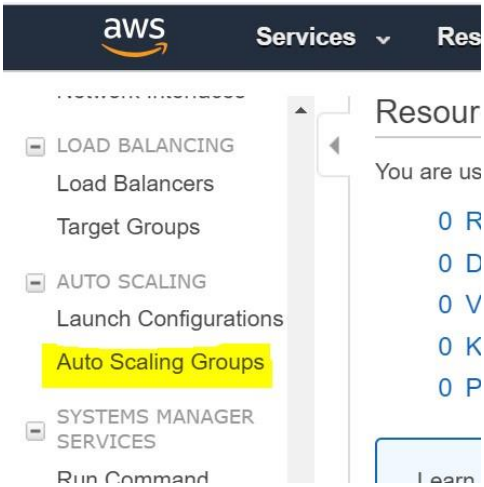
3.1 Advantages

1. For companies running their own web server infrastructure, auto scaling typically means allowing some servers to go to sleep during times of low load, saving on electricity costs (as well as water costs if water is being used to cool the machines).
2. For companies using infrastructure hosted in the cloud, auto scaling can mean lower bills, because most cloud providers charge based on total usage rather than maximum capacity.
3. Auto scaling can offer greater uptime and more availability in cases where production workloads are variable and unpredictable.
4. Auto scaling solutions, such as the one offered by Amazon Web Services, can also take care of replacing unhealthy instances and therefore protecting somewhat against hardware, network, and application failures.

4. Creating an Application Load Balancer and Auto Scaling

1. Open EC2 console.
2. Creating Auto Scaling group.

2.1 Go to Auto Scaling Groups.



2.2 Create Auto Scaling group and Create launch Configuration.

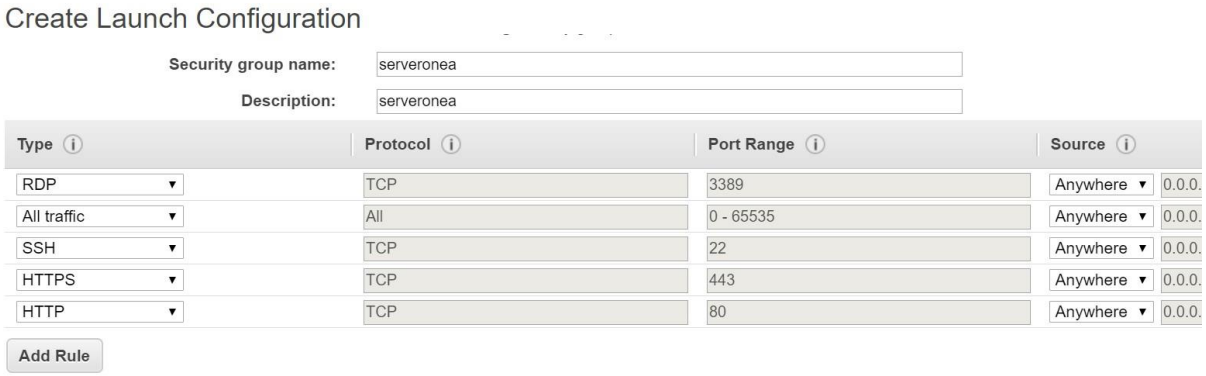
2.3 Select Microsoft Windows Server 2012 R2 Base.

2.4 In create Launch Configuration select t2 small with 2GB RAM.

2.5 Configure details (give your server name).

2.6 Add Storage to your Server.

2.7 In configure Security Group (give security group name and description) and add Rules as per your need.



2.8 Create Launch Configuration, download the new key pair and create launch.

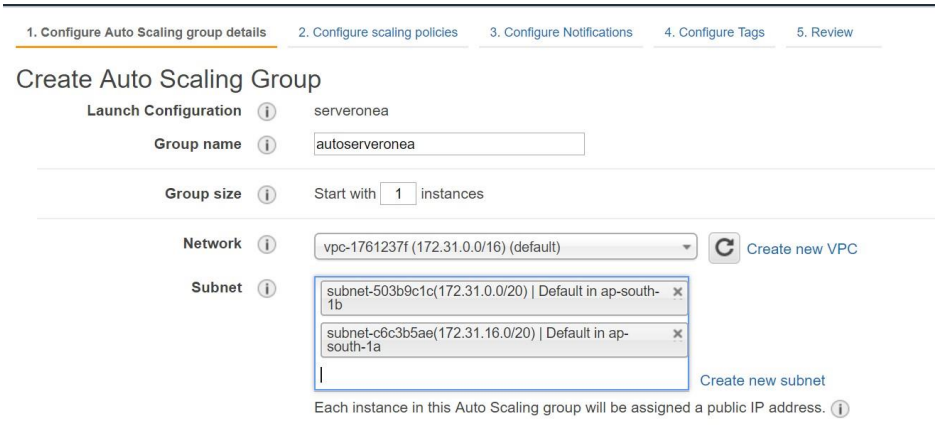
2.9 Configure Auto Scaling group details.

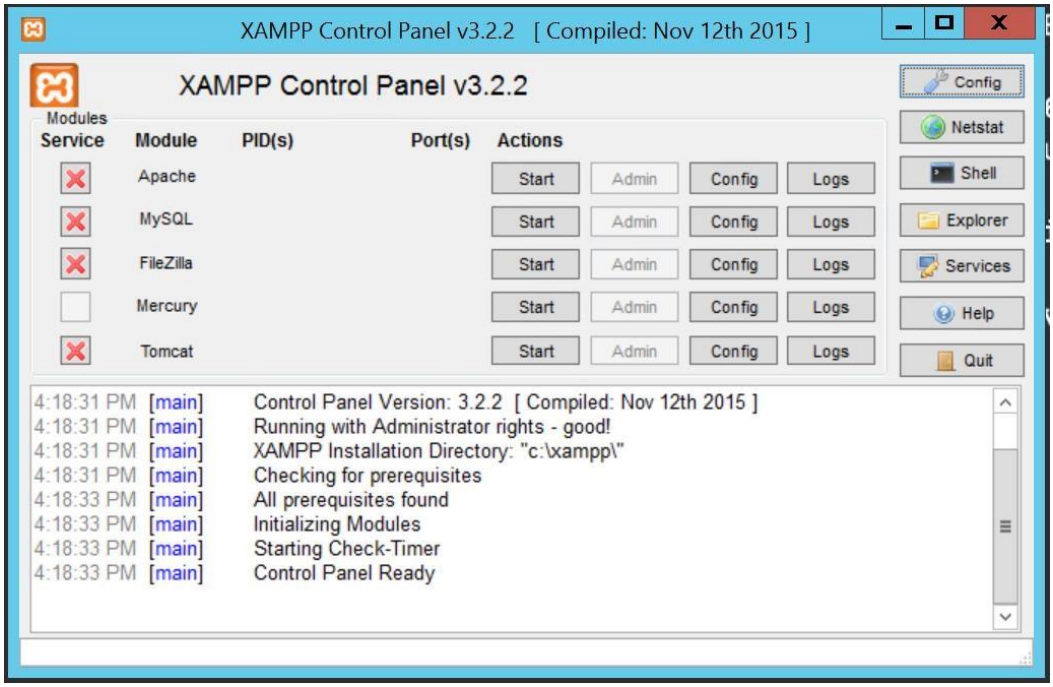
2.10 Now create auto scaling group and check whether the instance is running in EC2.

2.11 Now create another auto scaling group using the same process and name the instances.

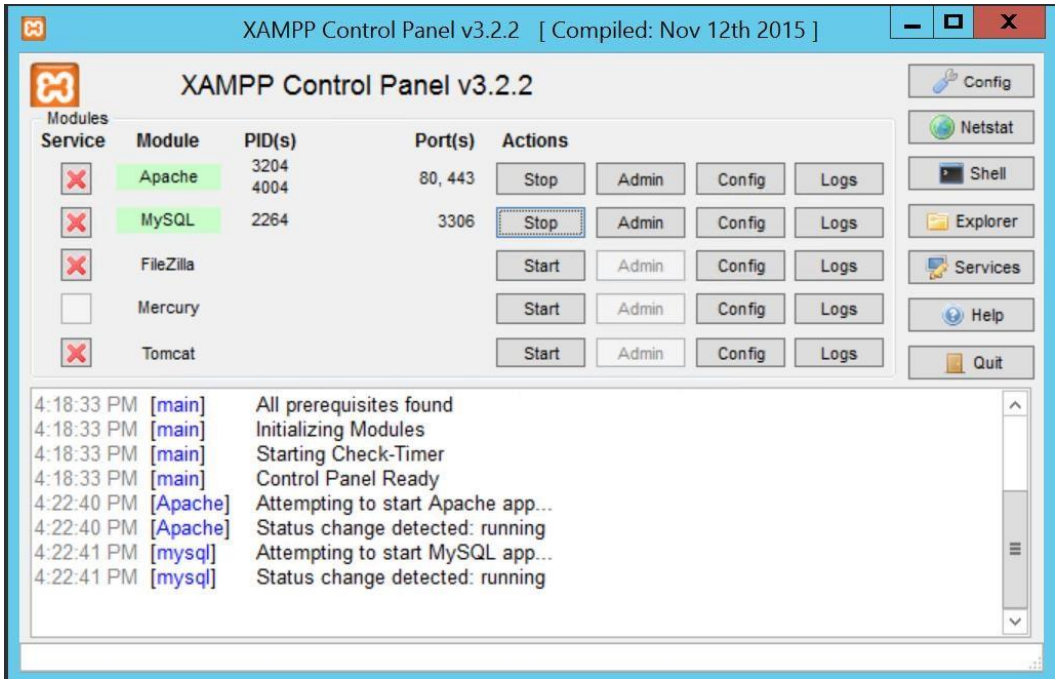
2.12 Connect both the instances and disable the firewall.

2.13 Download xampp on one of the server.



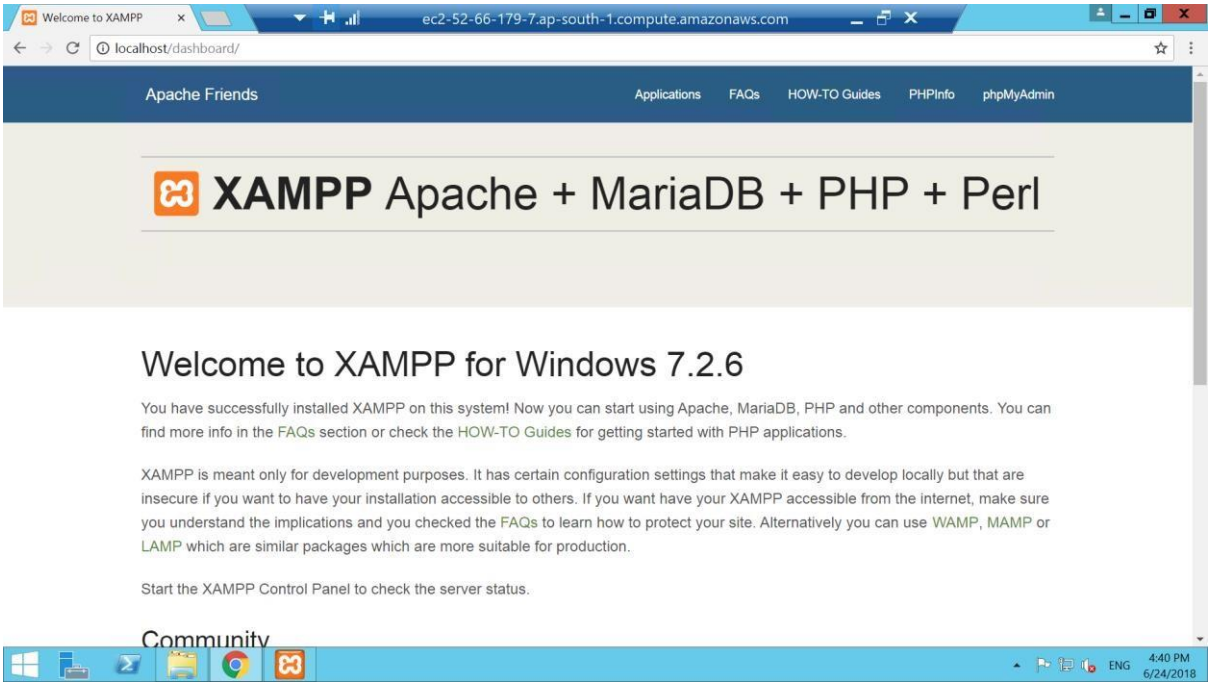


2.14 Start Apache and MySQL.

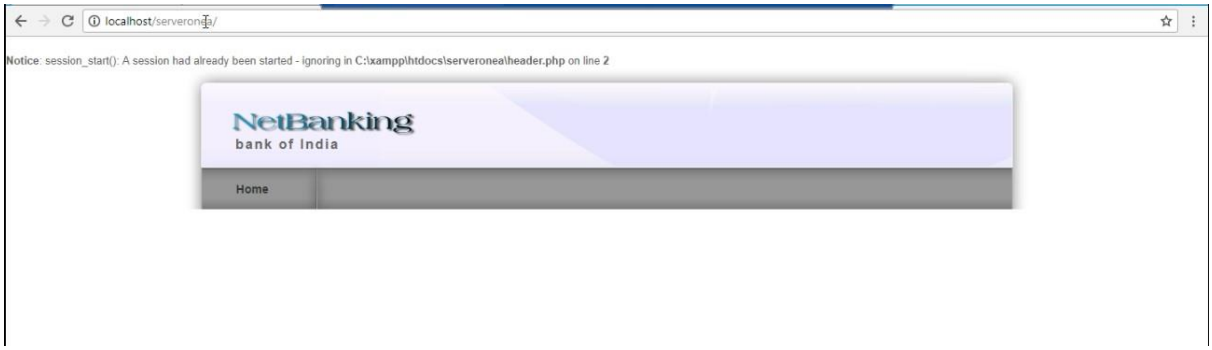


2.15 Copy your website to “C:\xampp\htdocs” by creating a new folder.

2.16 Go to the browser and type localhost



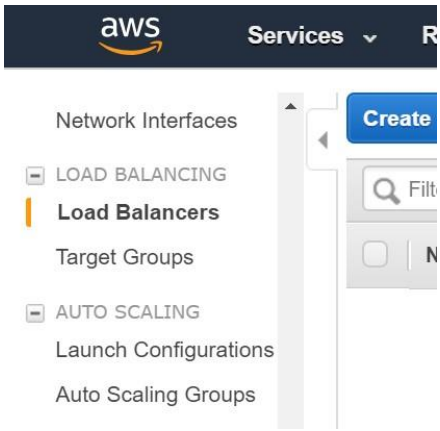
2.17 Now type localhost/'foldername'



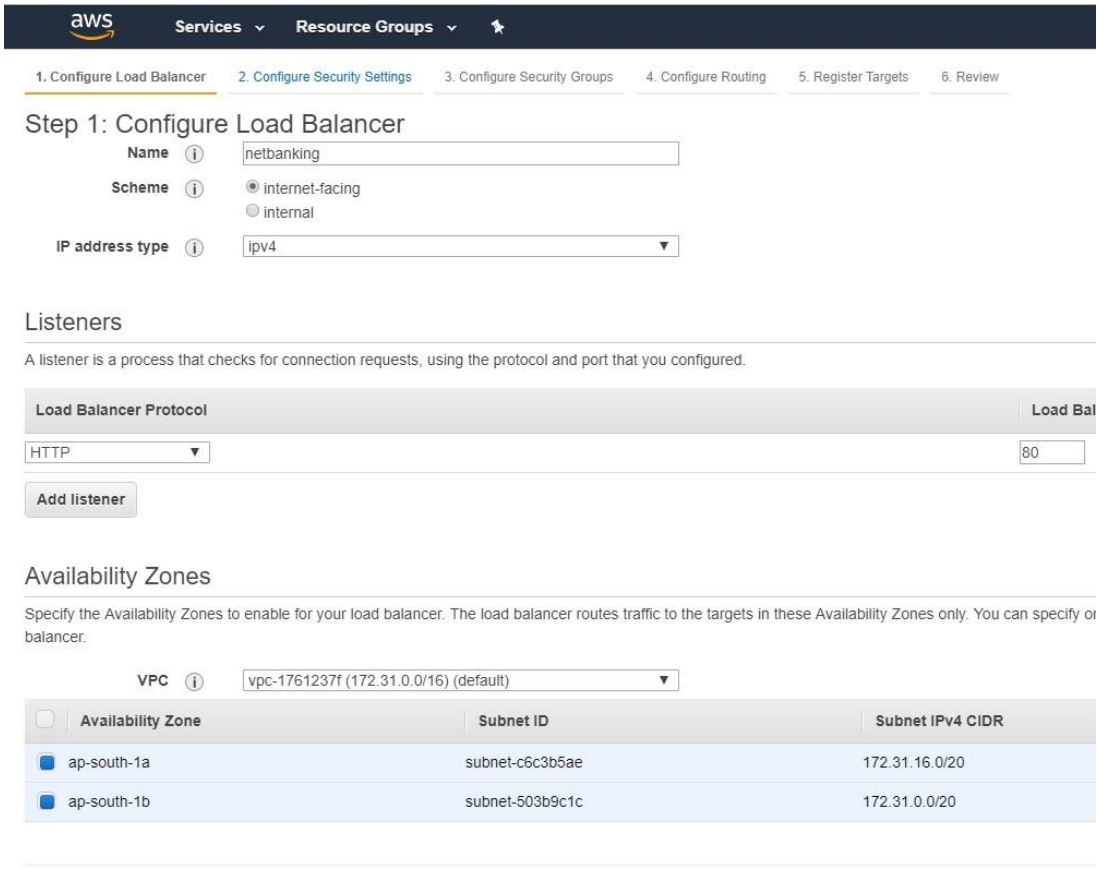
2.18 Do the same thing to the second server and check.

Note: auto scaling is done now load balancer part

2.19 Go to Load Balancer

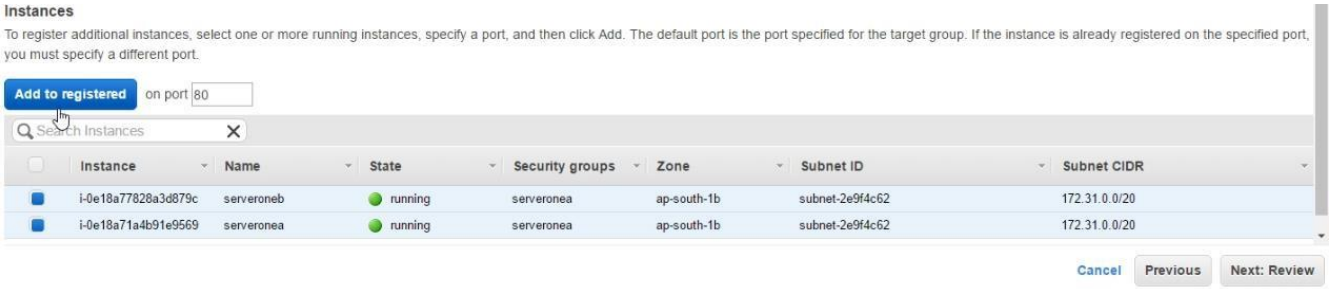


2.20 Create application load balancer and select all the availability zones

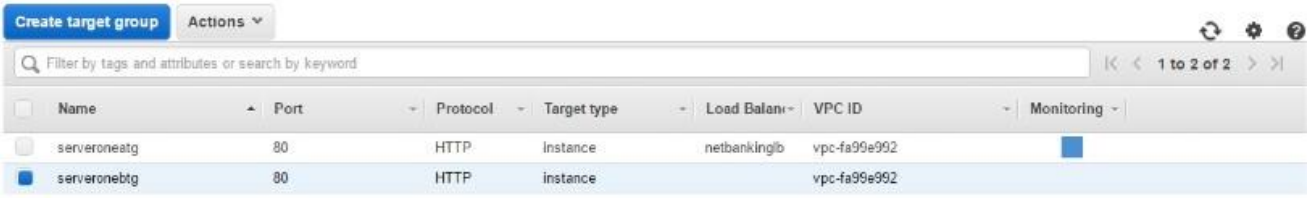


2.21 Select security group.

2.22 In register targets select both the servers and add to register



2.23 Go to Target group and create a target group.



2.24 Go to load balancer listener and select “view edit rules” and add rules.

2.25 Save the rules and copy the DNS and check.

2.26 If you stopped one new instance should start spinning