**Load Balancing**

High input request 🡪 Many instances running

Vice Versa

Elastic Load Balance

|

|

\/

Auto Scaling Group

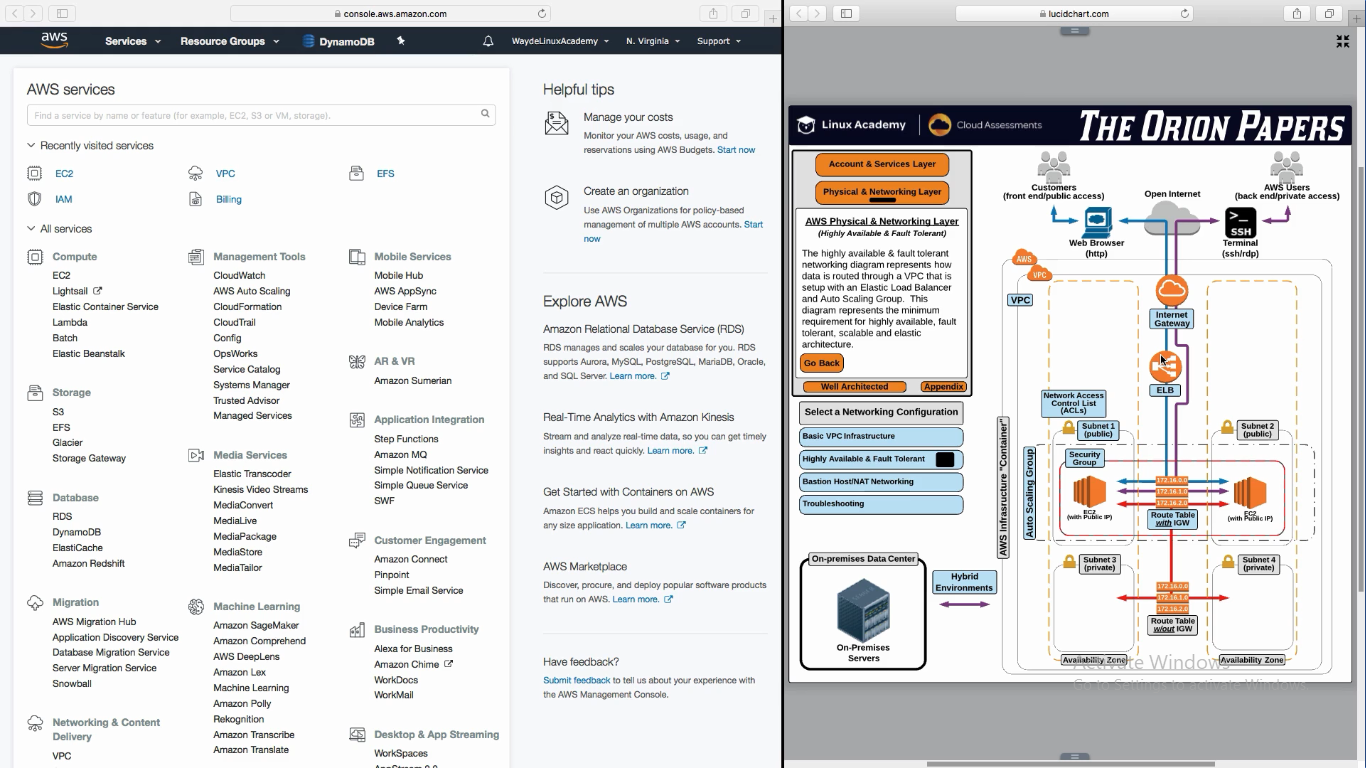
**Auto Scaling**

Three main components

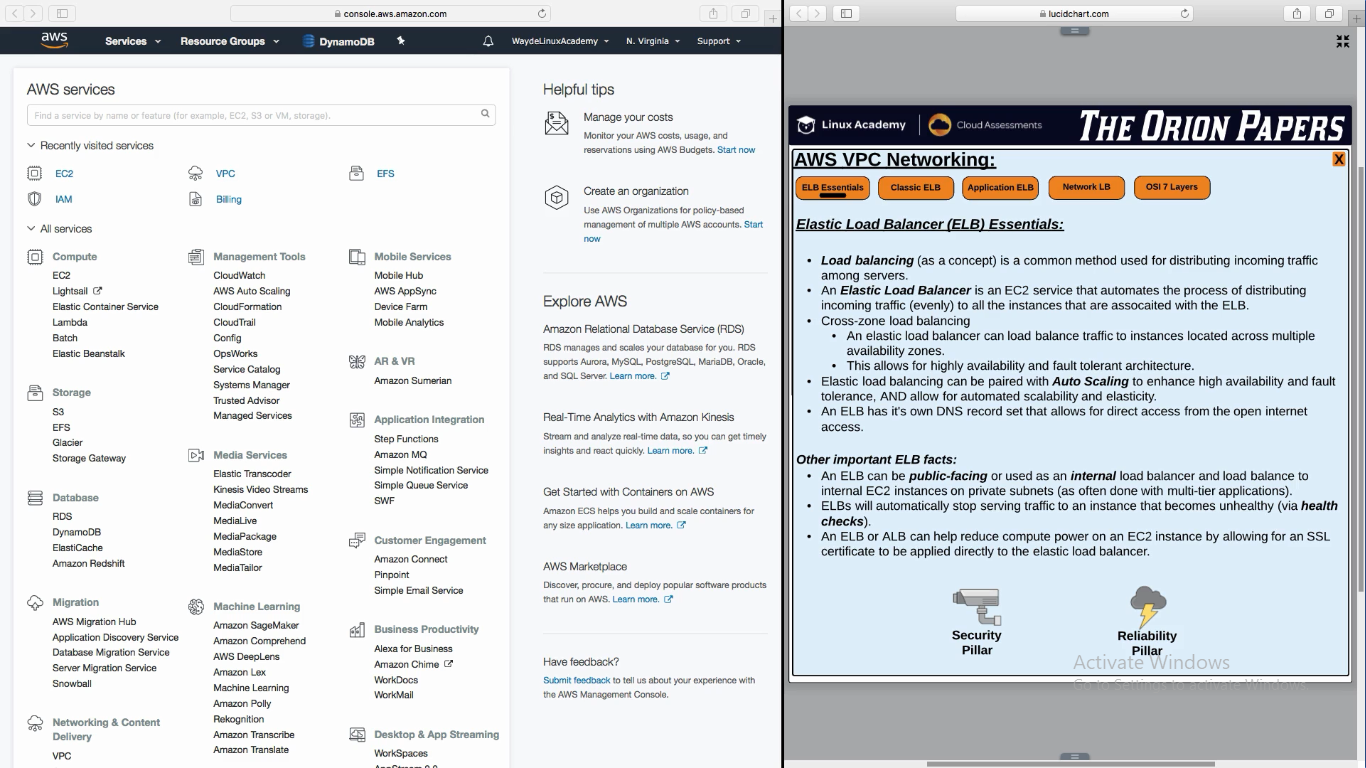
1. Elastic Load Balancing

**ELB – Elastic Load Balancing**

1. Managed service provided by aws to distribute traffic across multiple instance or servers
2. Ability to shrink and grow
3. It is self scaling
4. Adaptive
5. **Using load balancing we can achieve auto scaling in AWS**
6. Load balancer will be notified by the auto scaling group each time a node is added or terminated



Load Balancing get information from “Auto Scaling Group” Whenever a new instance is been added/removed



**Health Check**

Load Balancer sends a request to an instance and validates if it get a valid response 200. If not it determines that the instance is not healthy

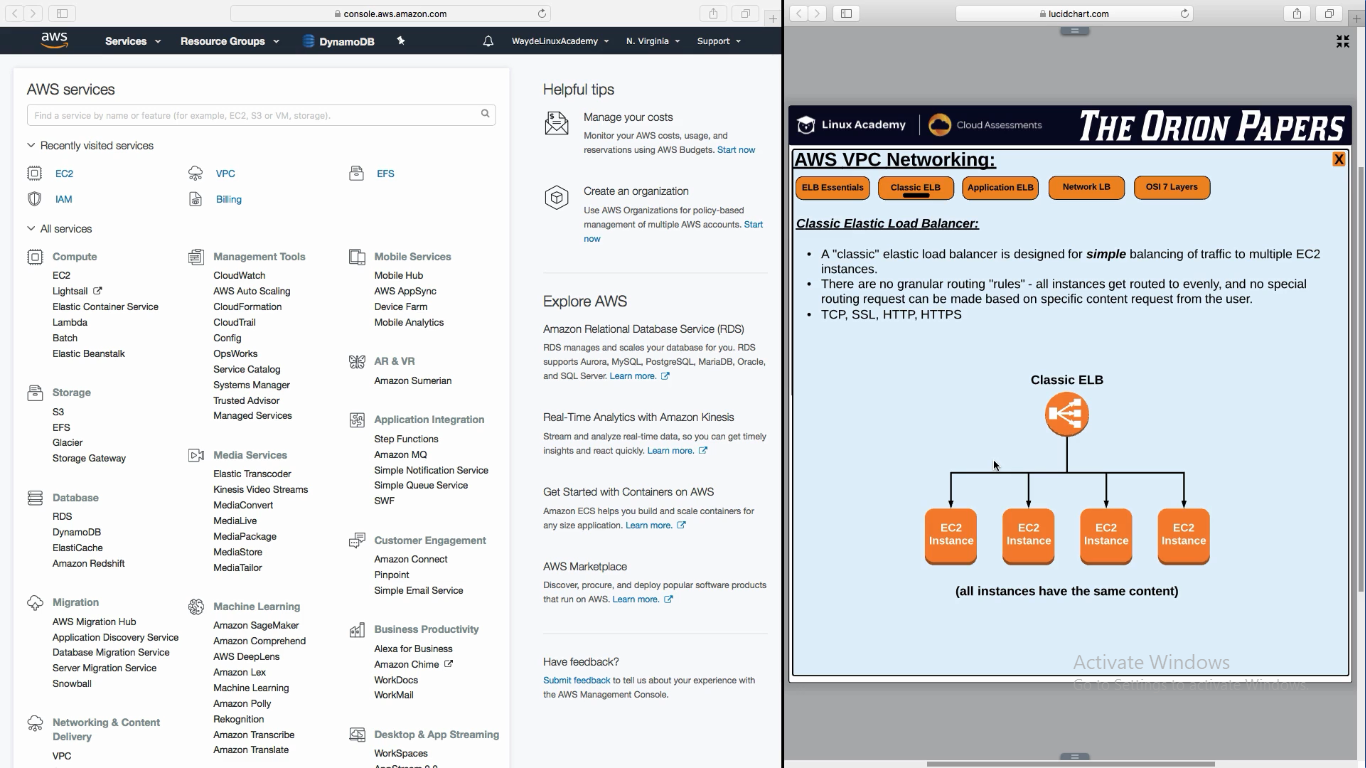
**Three Types of ELB – Elastic Load Balance**

1. Original Load Balancer – Now called as - Classic ELB
2. Application layer 7 ELB
3. Network load balancer

**Classic ELB:**

It wont validate the incoming request. Just send it to instances as how it is received

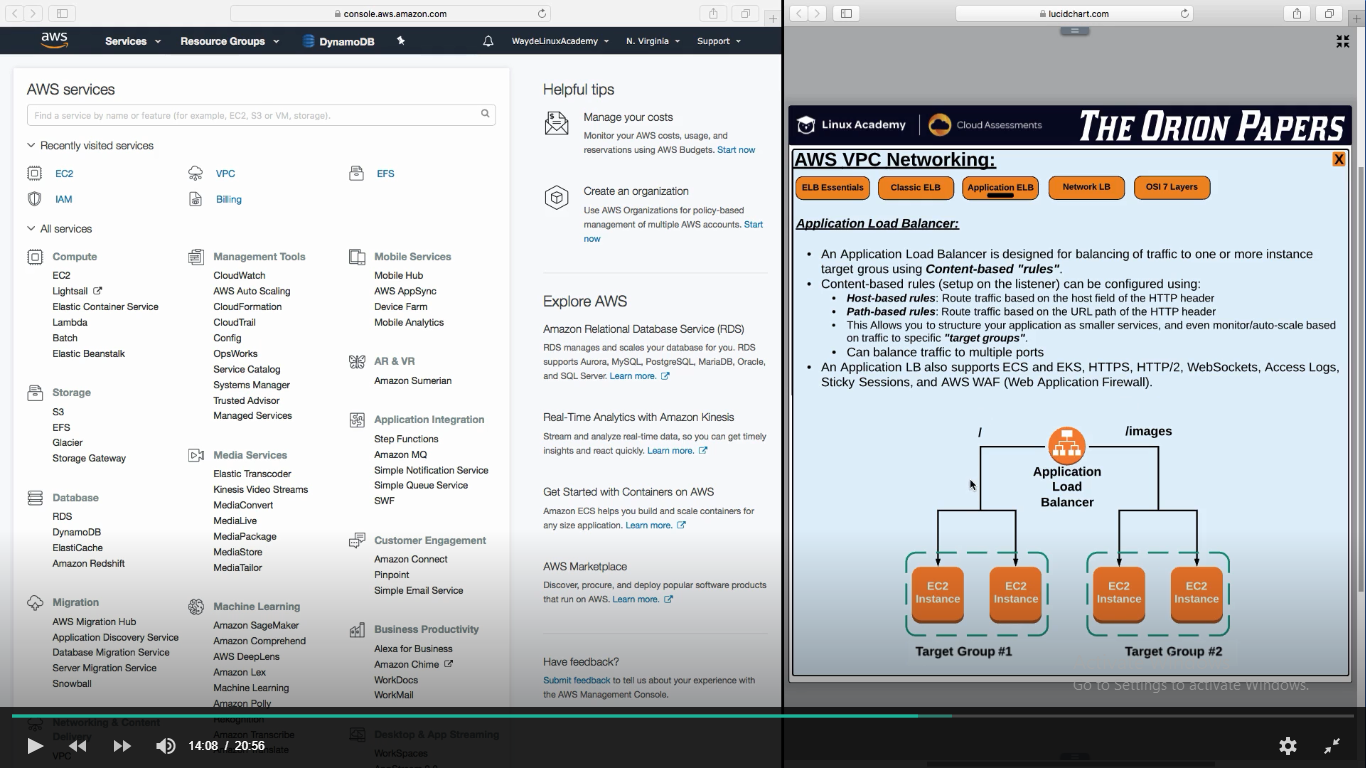
Ssl and tcp are sent to servers in round robin. http and https requests are sent to the instances based on the latency of the response from the instances



**Application ELB**

IT can do all what classic classic ELB Does. WAF – Web Application Firewall can be processed (block cross side scripting attack) and in Application ELB which cannot be done in classic ELB

Application ELB is layer 7

****

**Network Load Balancer**

This is layer 4

It does not launch instances

It is a plain TCP passthrough

It is a plain TCP passthrough

When scale in and out, we have different set of ip’s assigned for the instances. So we cannot route customer http request to one single ip. Hence we have to use DNS. If there is no DNS available, network load balancer offers a static ip (per availability zone) of elastic ip using which users can hit the application

**Thing which network and application LB does which Classic LB does not:**

Route traffic to ip address as target which could be servers back in datacenter. Possible to load balance traffic in ec2 instances in aws and servers back in data center at same time

