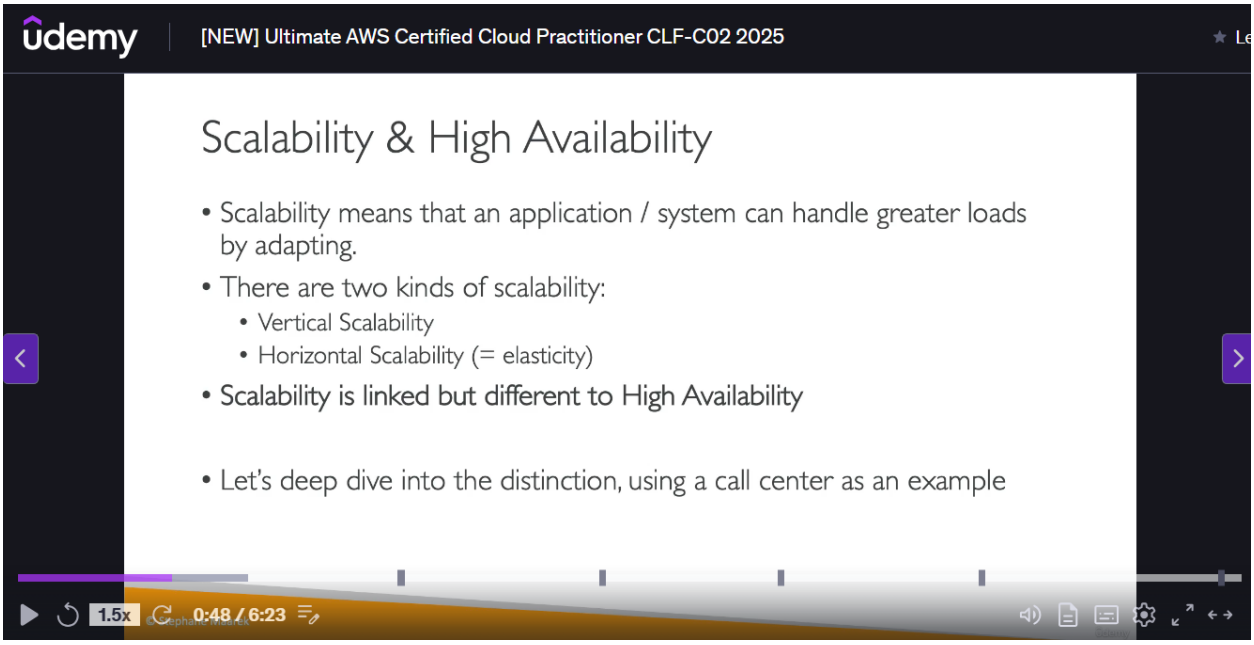
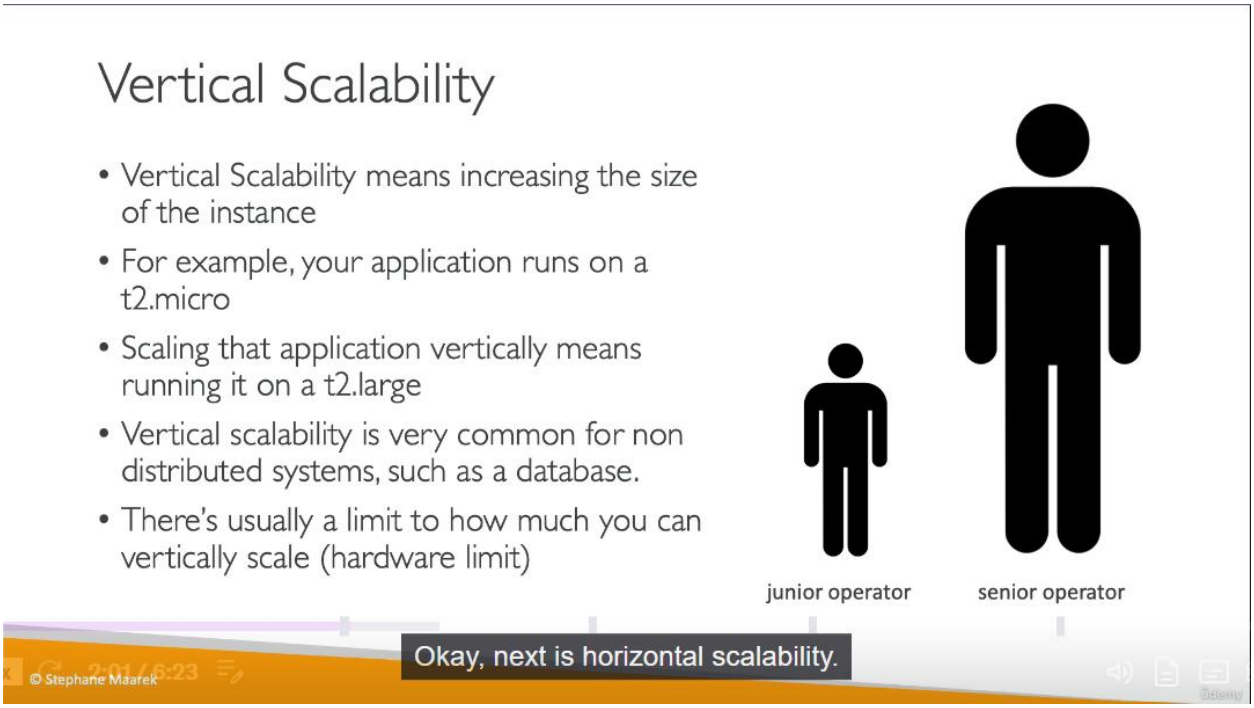
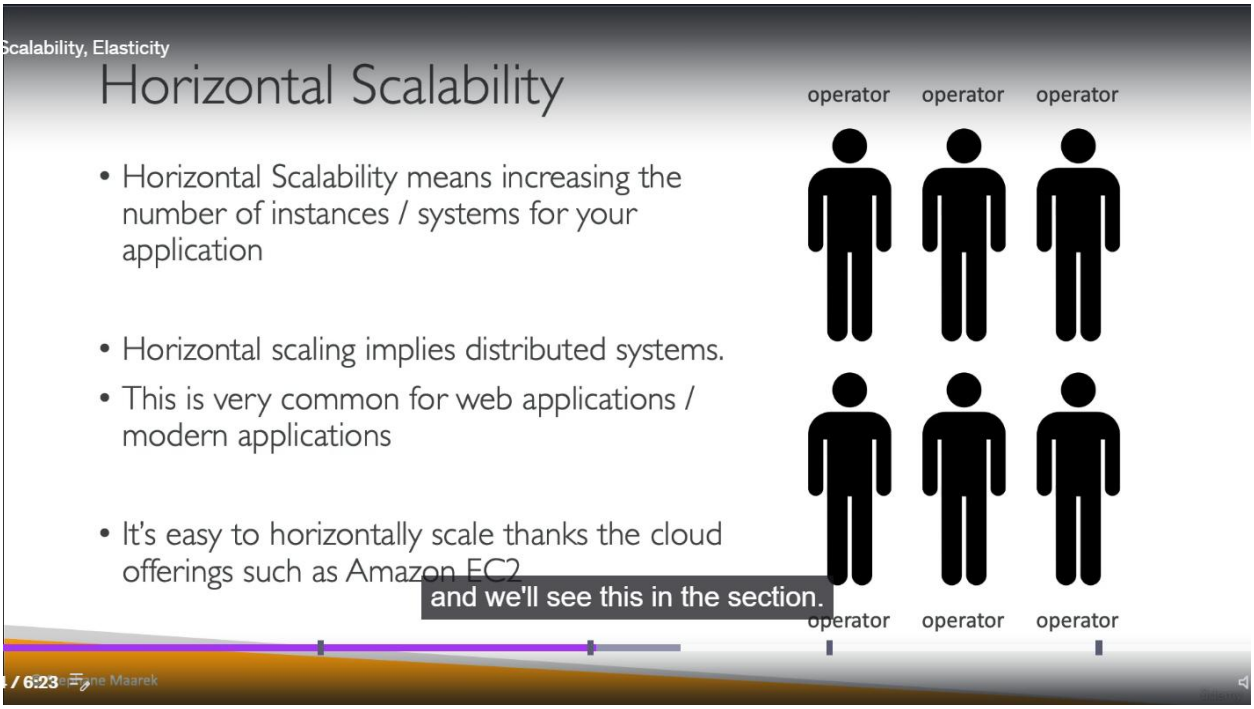


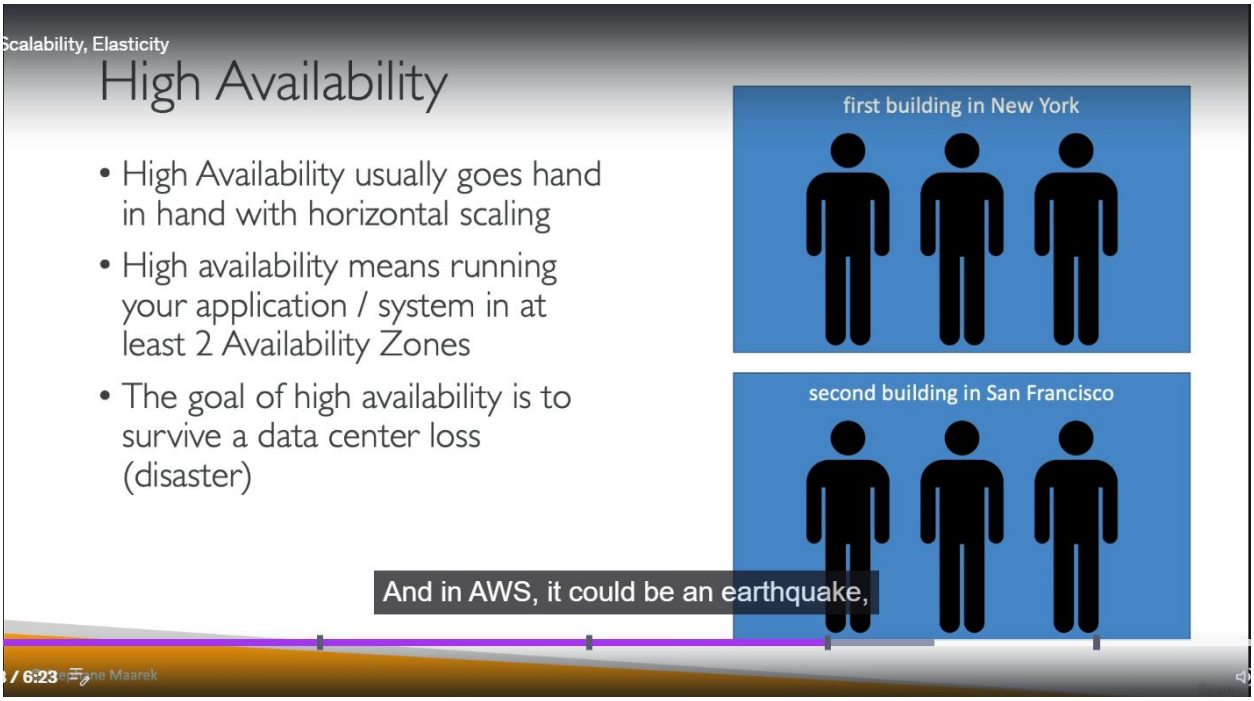


Junior upgraded to senior, senior can handle more calls.

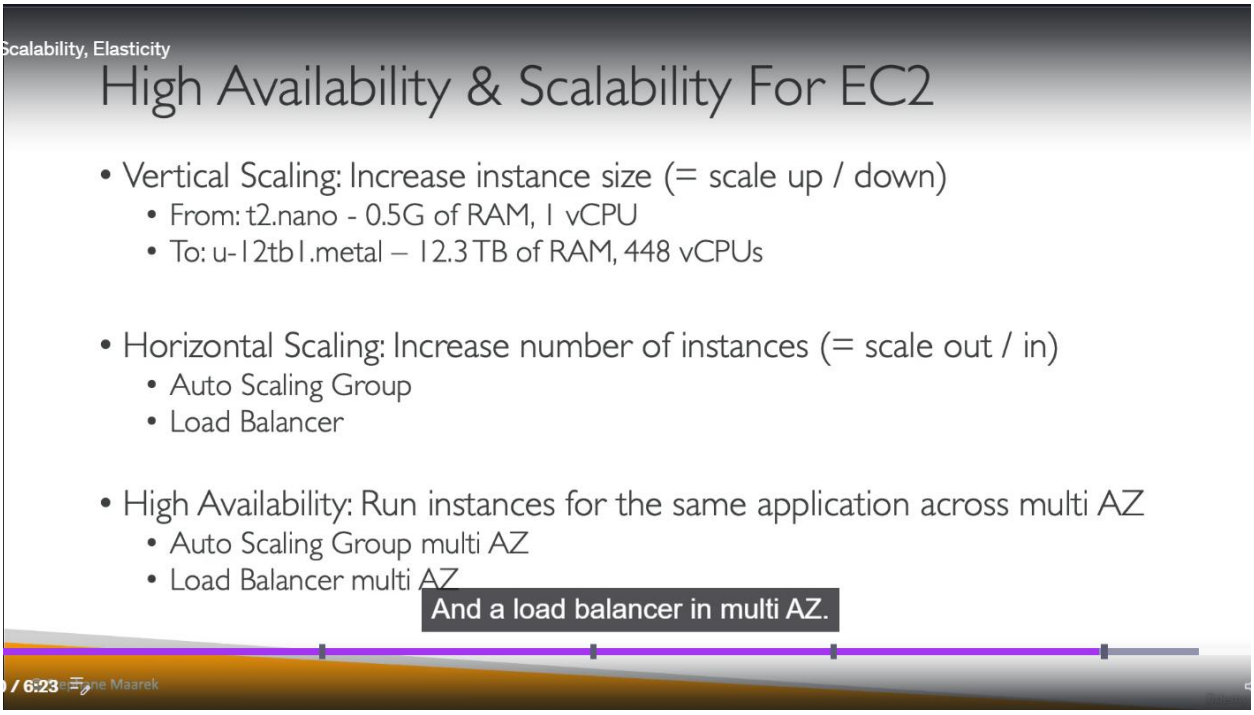
If want more performance to DB, just increase the size of DB



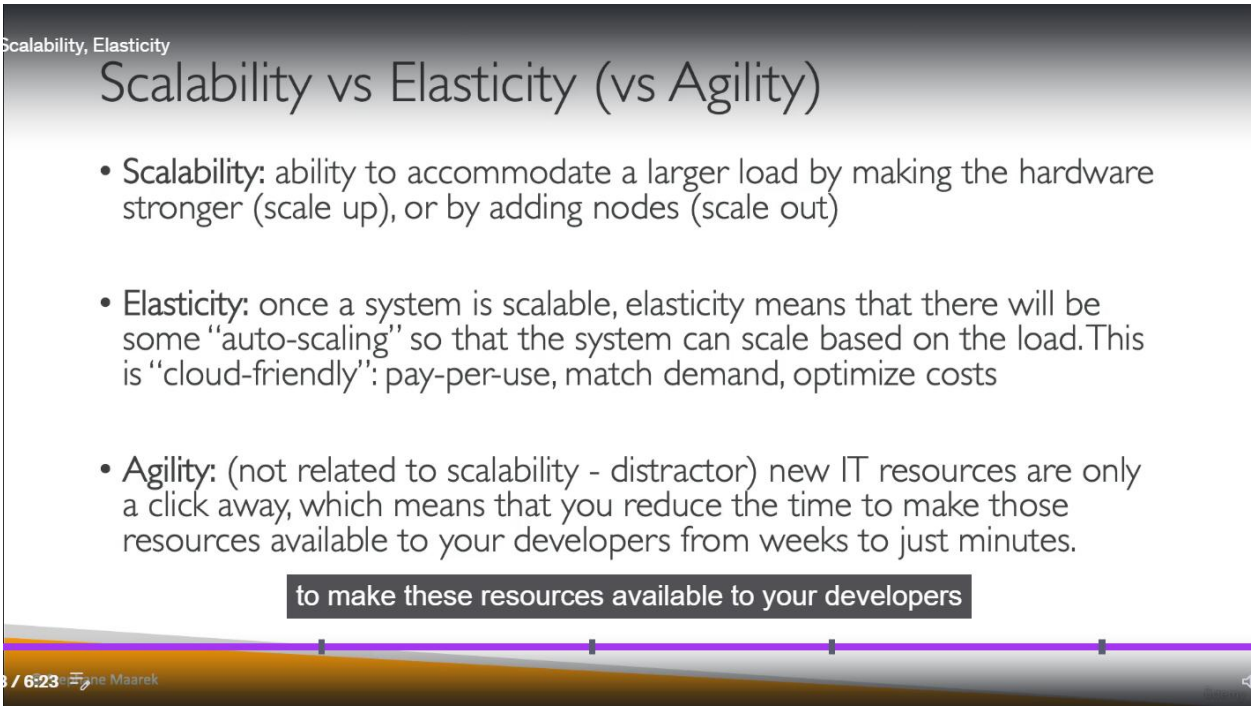
If want to handle more calls, add more operators



If new York fail, still take calls in san Fransisco



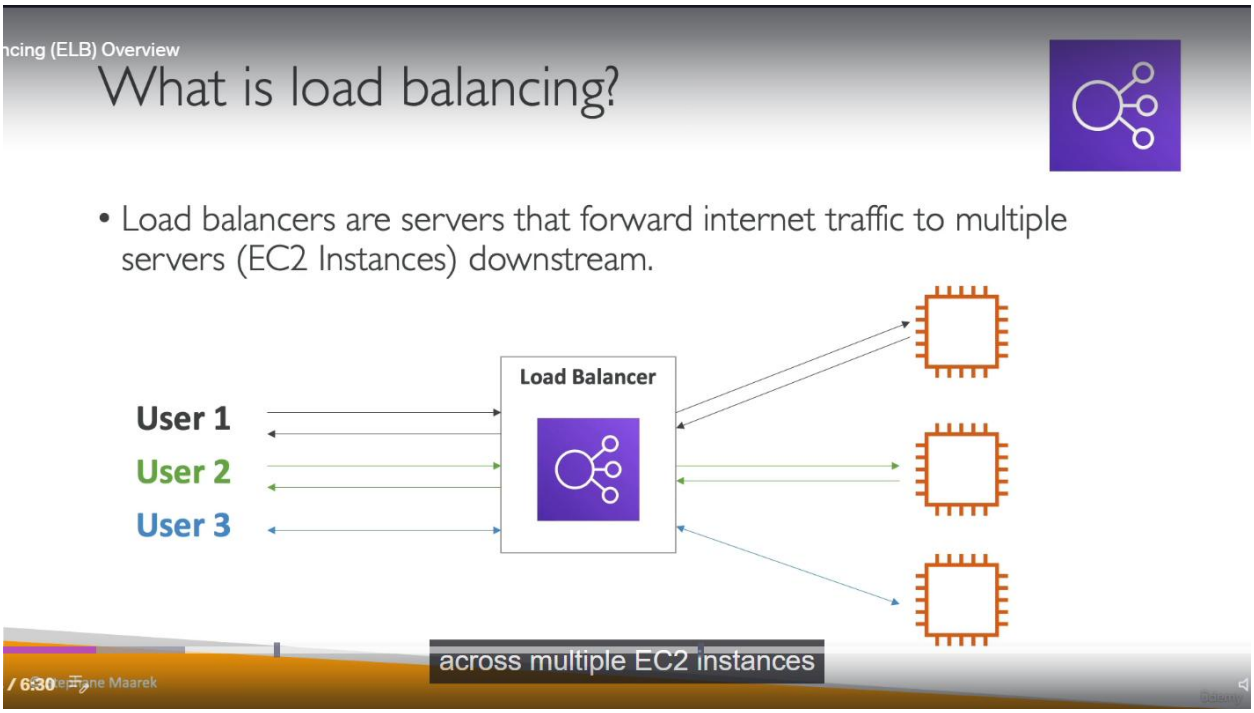
Auto Scaling is a cloud computing feature that **automatically adjusts the number of running instances or resources** in response to real-time demand — **scaling up** when demand increases and **scaling down** when demand drops.



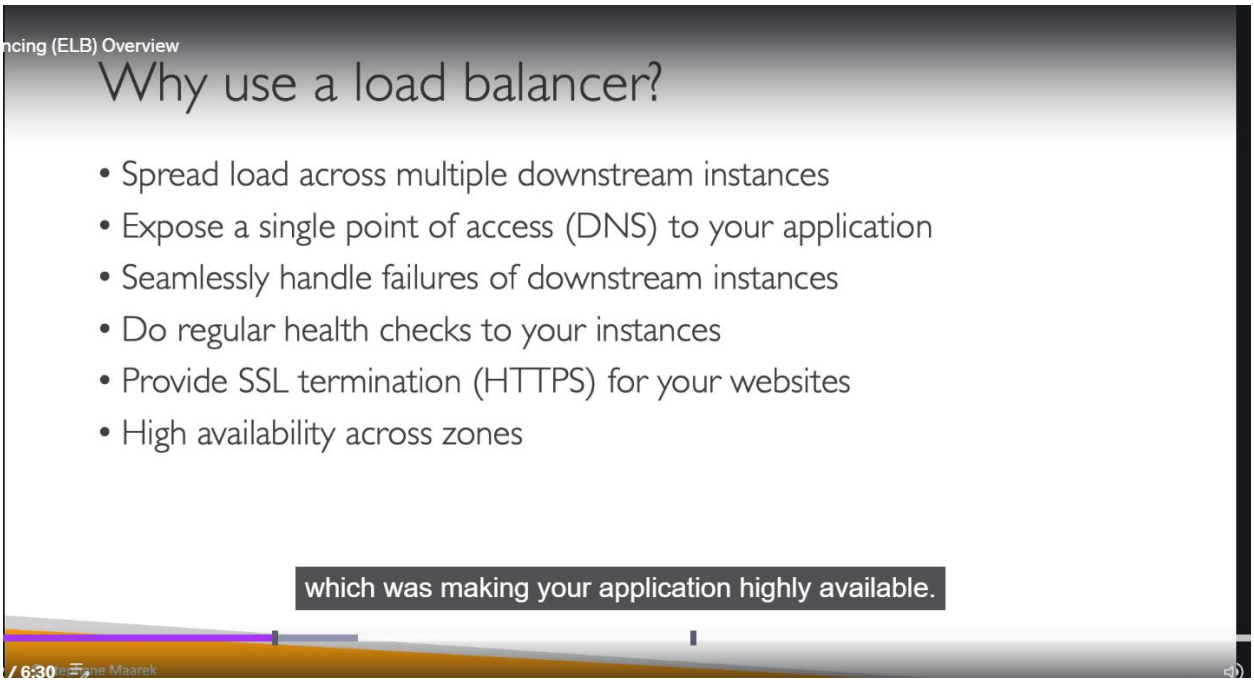
ELASTIC LOAD BALANCING(ELB)

Allows to be elastic

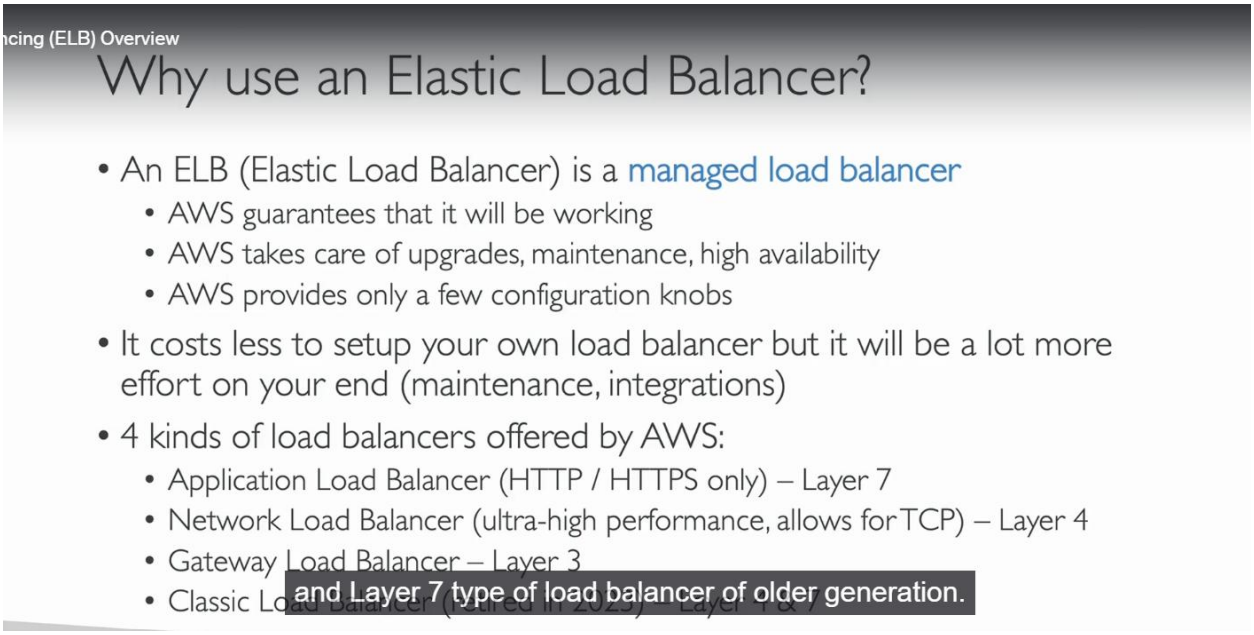
also called backend EC2 instance



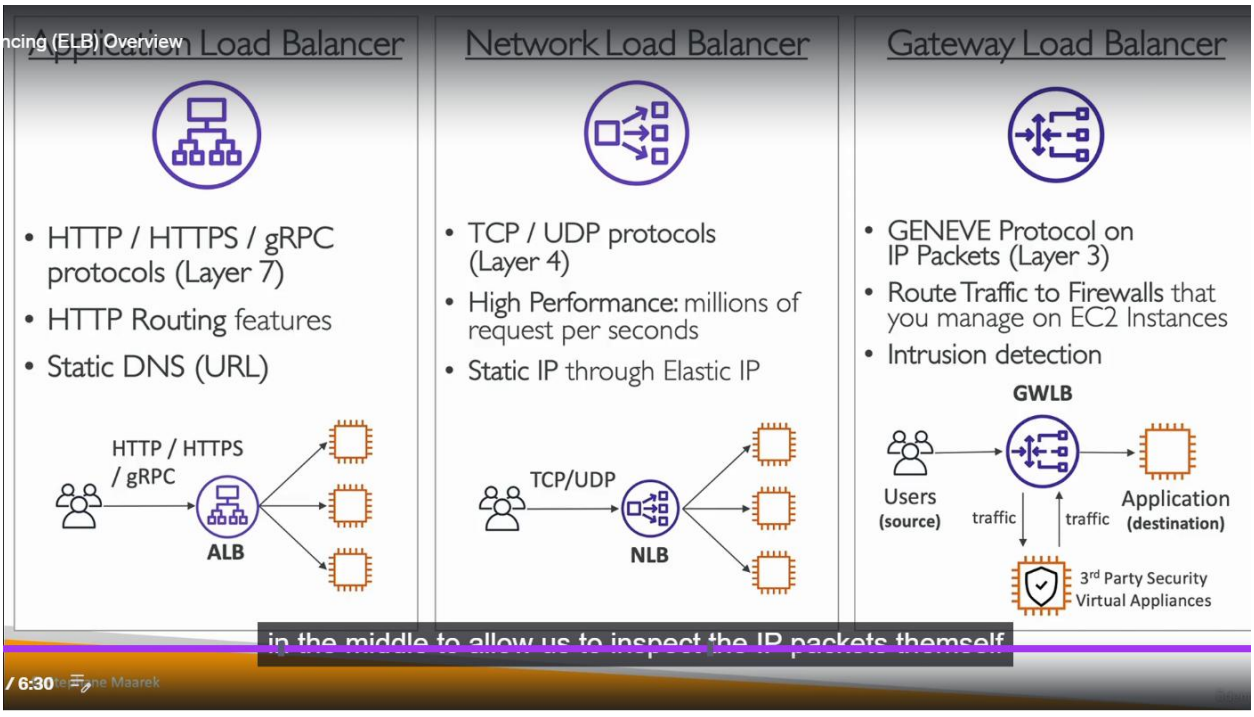
more users, more it will balance the load,..allow us to scale better our backend



If one instance fails, don't send to that instance, so hiding our failure



Only thing we have to do in configure



GLB inspects IP packets and to perform firewall features or intrusion detection or deeppacket inspection

🚀 **What is an Auto Scaling Group (ASG)?**

An **Auto Scaling Group** (used in AWS and other clouds) is a **collection of virtual machines (instances)** that are managed together and **automatically scaled** based on demand.

🔧 **Key Features:**

Feature	Description
Auto-scaling	Adds/removes instances automatically based on CPU, memory, or custom metrics
Minimum and Maximum Size	You define how few and how many instances the group can have
Health Checks	Unhealthy instances are replaced automatically
Integration with Load Balancer	Works with load balancer to route traffic efficiently

---

✅ **Example Use Case:**

You have a website. In the evening, user traffic spikes. The **ASG launches 3 more servers**. Later at night, traffic drops — **ASG terminates those extra servers** to save cost.

🌐 **What is a Load Balancer?**

A **Load Balancer** is a service that **distributes incoming traffic across multiple servers** to ensure:

- No single server is overloaded
- High availability and fault tolerance
- Faster response times

🔧 **Types of Load Balancers:**

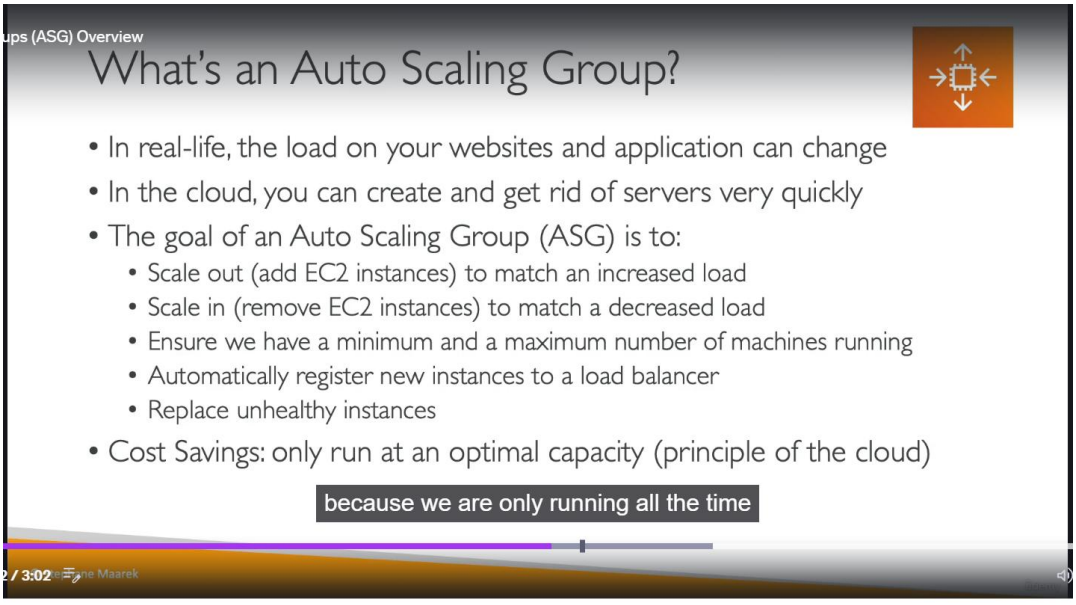
Type	Description
Application Load Balancer (ALB)	Routes based on HTTP/HTTPS **content** (e.g., URL paths, host headers)
Network Load Balancer (NLB)	Routes based on **IP and TCP/UDP ports** — lower level, high performance
Classic Load Balancer (CLB)	Older, basic version for simple traffic distribution

---

🛠️ **How They Work Together:**

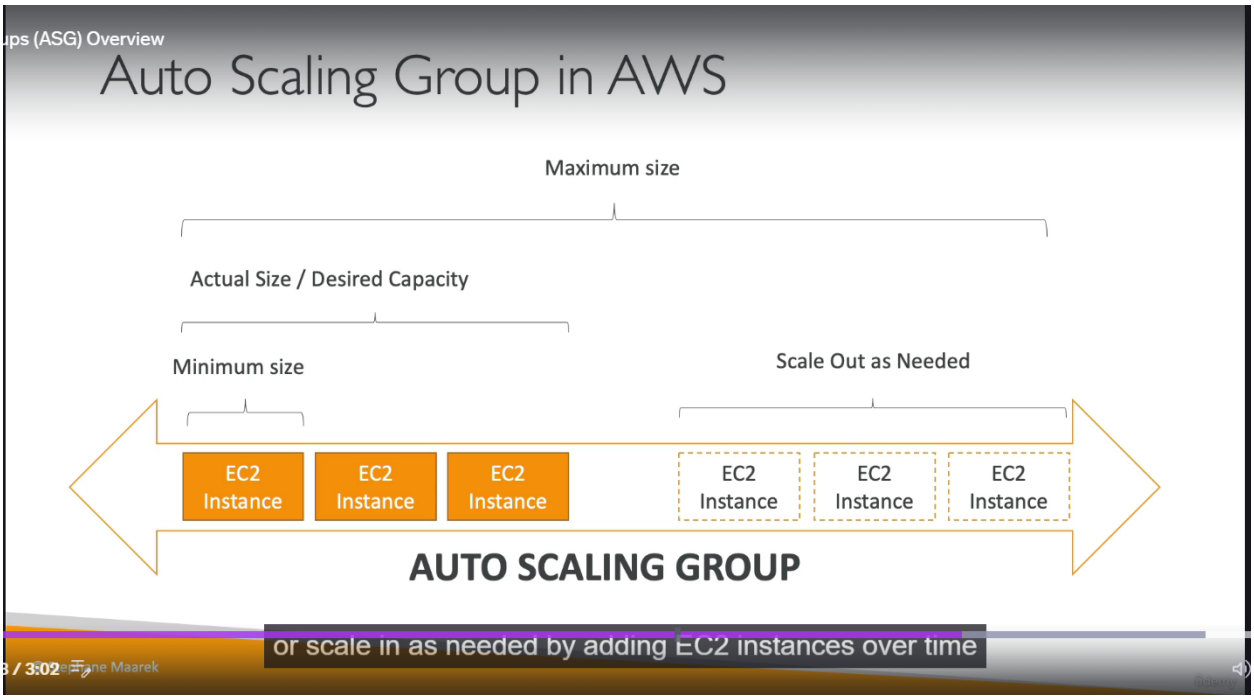
1. **Users** access your web app.
2. Traffic hits the **Load Balancer**.
3. Load Balancer forwards it to one of the **instances in the Auto Scaling Group**.
4. If traffic increases, **ASG adds instances** and Load Balancer includes them automatically.
5. If traffic drops, **ASG removes instances** and Load Balancer stops sending traffic to them.

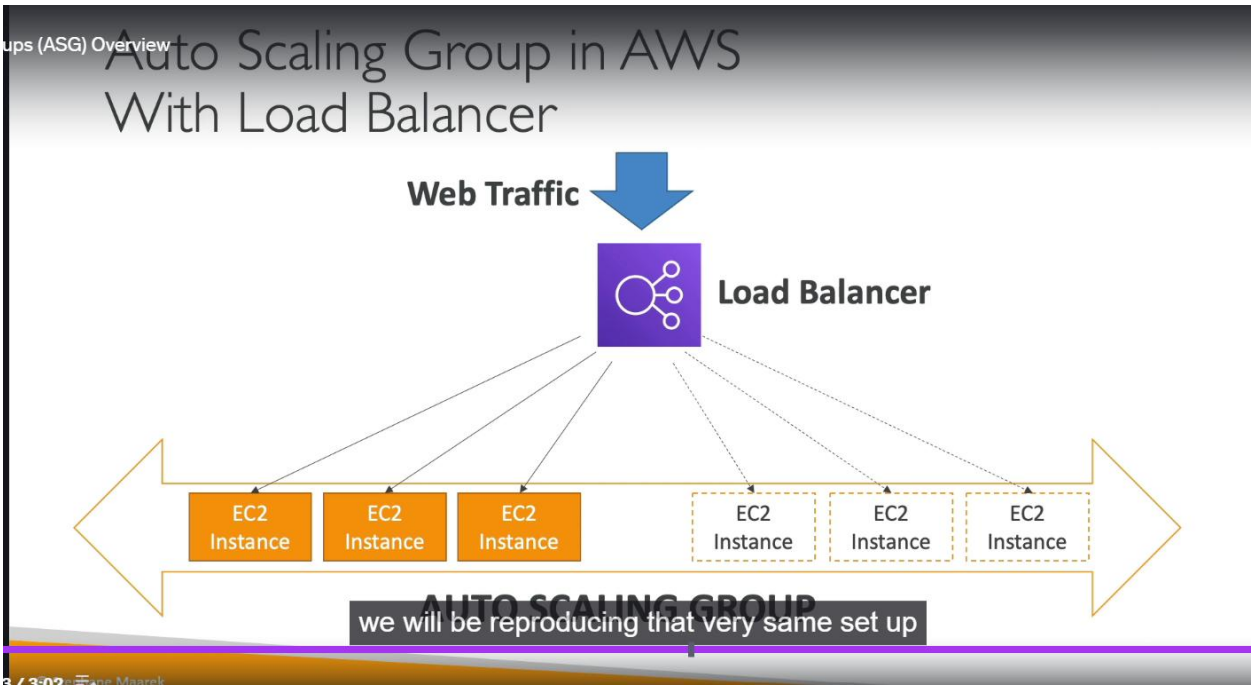
ASG



---

more shopping at day and not at night


ups (ASG) Overview
Auto Scaling Group in AWS
Maximum size
Actual Size / Desired Capacity
Minimum size
Scale Out as Needed
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
AUTO SCALING GROUP
or scale in as needed by adding EC2 instances over time


ups (ASG) Overview
Auto Scaling Group in AWS
With Load Balancer
Web Traffic
Load Balancer
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
AUTO SCALING GROUP
we will be reproducing that very same set up

ps (ASG) Strategies

Auto Scaling Groups – Scaling Strategies

- **Manual Scaling:** Update the size of an ASG manually
- **Dynamic Scaling:** Respond to changing demand
  - Simple / Step Scaling
    - When a CloudWatch alarm is triggered (example CPU > 70%), then add 2 units
    - When a CloudWatch alarm is triggered (example CPU < 30%), then remove 1
  - Target Tracking Scaling
    - Example: I want the average ASG CPU to stay at around 40%
  - Scheduled Scaling
    - Anticipate a scaling based on known usage patterns
    - Example: increase the min. capacity to 10 at 5 pm on Fridays

People are going to do sports betting maybe who knows,

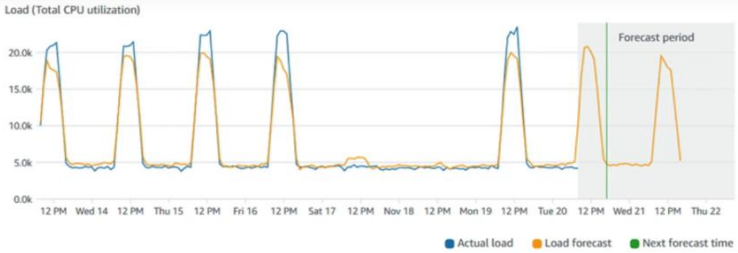
/ 2:54 Stephane Maarek

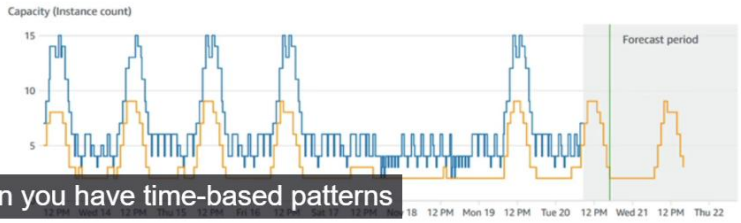
Changing capacity from 1 to 2 or etc

ups (ASG) Strategies

Auto Scaling Groups – Scaling Strategies

- **Predictive Scaling**
  - Uses Machine Learning to predict future traffic ahead of time
  - Automatically provisions the right number of EC2 instances in advance
- Useful when your load has predictable time-based patterns





when you have time-based patterns

/ 2:54 Stephane Maarek

Elasticity is scaling up and down