Hello, Cloud Gurus, and welcome to this lecture

where we are going to explore Elastic Load Balancer.

So what is a load balancer?

Well, a load balancer distributes network traffic

across a group of servers.

So if you think of a simple website

consisting of 3 web servers behind a load balancer,

the user browses to the website

and the request hits the load balancer

and then gets routed to one of the web servers.

And depending how you've got it all configured,

the load balancer might send the request to the web server,

which is the least busy,

or it might use a round robin algorithm,

so forwarding requests to each server in turn.

And the great thing about this architecture

is that if one of the web servers fails,

then the load balancer will notice

and it will stop sending requests to this web server

until it comes back online.

And we can easily increase the capacity when needed.

So if our website suddenly gets really popular,

then we can add some more web servers

and register them with the load balancer.

And the load balancer will start sending traffic

to our new web servers as well.

Now with AWS Elastic Load Balancer,

there are a few different options to choose from.

So we've got the Application Load Balancer,

which load balances HTTP and HTTPS.

There's also the Network Load Balancer,

which load balances TCP traffic,

and this is the high performance option.

We've also got the Classic Load Balancer,

which can handle HTTP and HTTPS

and TCP protocols as well.

And this is the legacy option.

And there's also the Gateway Load Balancer.

And this one is a fairly new offering

and it may or may not come up in the exam,

but it's good to be aware of at a high level.

And this one allows you to load balance workloads

for third-party virtual appliances running in AWS,

such as virtual appliances purchased on the AWS Marketplace,

virtual firewalls from companies like Fortinet,

Palo Alto, Juniper, or Cisco,

and intrusion detection or intrusion prevention systems

from companies like CheckPoint and Trend Micro, etc.

So let's take a look at each of the main load balancers

in a little bit more detail

beginning with the Application Load Balancer.

Now, this is used for load balancing HTTP and HTTPS traffic

and Application Load Balancers operate at Layer 7

of the OSI model, so they are application-aware.

And if you haven't heard of the OSI model,

it's also called the 7-layer model.

So what is it?

Well, it's a conceptual framework,

which describes the functions of a network,

beginning with the application layer,

which directly serves the end user

right down to the physical layer and everything in between.

So we've got Layer 7, which is the application layer,

and this is everything that the end user sees.

So HTTP operates at this layer

and so does your web browser as well.

Then we have Layer 6, which is the presentation layer,

and this layer makes sure

that the data is in a usable format

and protocols like encryption and SSH

operates at this layer.

Layer 5 is also known as the session layer

and this is all about maintaining connections and sessions.

Then there's Layer 4, which is another important layer,

and it's also known as the transport layer.

And this is all about transmitting data using protocols

like TCP and UDP.

And, of course, TCP is one of the main protocols

of the internet.

Layer 3 is also known as the network layer.

And this layer is concerned

with logically routing network packets

based on IP addresses.

Layer 2 is the data link layer,

and it is concerned with physically transmitting data

based on MAC addresses.

And then there's Layer 1, which is the physical layer,

and this is all about transmitting bits and bytes

over physical devices,

like the cables and hubs that make up the network.

So the Application Load Balancer is operating at Layer 7,

and it's at this layer where HTTP is operating.

And in the exam, you won't be tested on the OSI model,

but just remember that the Application Load Balancer

is HTTP-aware and it is operating at Layer 7

or the application layer of the network stack.

Now, as the Application Load Balancer is application-aware,

they support advanced request routing.

So that means you can route requests to specific web servers

based on the HTTP header.

So let's take a look at an example.

So just imagine a car dealership website,

and I want you to think about the kind of services

that you might find.

So they'll definitely be selling cars,

but they might also offer loans or credit agreements

and service and repairs.

And by using an Application Load Balancer,

we can send requests to specific web servers

depending on what the user is looking for.

We can send sales inquiries to one set of web servers,

loan applications can go to another,

and if you're booking a service or repair,

that can go to another set of web servers.

And the Application Load Balancer

can handle all of this for you

because it is operating at the application layer.

So it can use the HTTP request header

to determine where to send each request.

Moving on to the Network Load Balancer

and this is the high performance option.

So I want you to think of a super fast sports car.

And this is used for load balancing TCP traffic

when extreme performance is required.

And it operates at Layer 4 of the OSI model,

which is the transport layer.

So if we come back to our 7-layer model,

it's at this Layer 4 transport layer where TCP operates.

So the Network Load Balancer is load balancing

based on the TCP protocol.

And it is capable of handling

millions of requests per second

while maintaining ultra low latencies.

But just remember as it's the highest performance,

it is also the most expensive option.

Moving on to Classic Load Balancers.

And when you think of Classic Load Balancer,

I just want you to think of the classic car.

So it's not going to be the fastest

and it's not going to have the most incredible features either.

And this is the legacy option,

but it may still appear in the exam.

So Classic Load Balancers,

they support some Layer 7-specific features

like X-Forwarded-For headers,

which we'll cover in the next slide,

and sticky sessions that just allows you

to keep sending requests

which originate from the same session

onto the same web server, making the session sticky.

And it does also support Layer 4 load balancing

for applications which rely purely on the TCP protocol.

So now let's circle back to the X-Forwarded-For header.

And this is an HTTP header, which allows you to identify

the originating IP address of a client

connecting through a load balancer.

But what does that mean?

Well, here is our client and her IP address is 124.12.3.231

and here is our web server

and she's connecting to the website through a load balancer.

And let's say the load balancer

has a private IP address of 10.0.0.23.

Now, when the request reaches the web server,

the web server will only see this private IP address

from the load balancer, so it's only going to see 10.0.0.23.

Now this might cause problems for your application

because you might want to know

where all these requests are coming from.

You might want to be sure that it's a trusted network

or that the request is coming from a location

or a country that you are allowed to operate in.

So how can you get the IP address?

Well, it's really easy

and that's where the X-Forwarded-For header comes in.

So the originating IP address is going to appear

in this X-Forwarded-For HTTP header.

So if we are using a load balancer,

which supports X-Forwarded-For,

then that means that we can identify

the originating IP address.

And then finally,

we need to cover common load balancer errors,

and the most common one that you're going to see is Error

504, and you might see the message Gateway Timeout.

So what does that mean?

Well, it usually means that the target

or the downstream web or application server

has failed to respond.

So what do we do?

Well, the first thing we need to do

is check the application.

It could be that the Elastic Load Balancer

could not establish a connection to the target.

For example, the web server, database, or Lambda function.

It could be that your application is having issues.

And in order to resolve the problem,

you will need to identify where the application is failing

and fix the problem.

And I actually experienced a 504 error only recently

and I was making a payment online

and the payment took too long to authorize,

so I got this Gateway Timeout error.

And I don't know what went wrong,

but what I think happened is that my bank was too slow

to authorize the payment

and we got a Gateway Timeout.

So my transaction failed and I had to try again.

So Gateway Timeout is an error that you might see

when you are attempting to make a payment online.

So onto my exam tips for Elastic Load Balancer.

First of all, we've got the Application Load Balancer,

which load balances HTTP and HTTPS,

and this provides intelligent load balancing,

which allows you to route requests to a specific web server

based on the HTTP request header.

So think of the car dealership where they're offering sales

and loan agreements and repairs.

We've got the Network Load Balancer,

which is the high performance option

for TCP traffic only.

The Classic Load Balancer, which is the legacy option

and supports both HTTP and HTTPS and TCP as well.

The Gateway Load Balancer,

which is for third-party virtual appliances.

And then finally,

if you need to find the IPv4 IP address of your end user,

then you can just find it

in the X-Forwarded-For HTTP header.

So that is it for this lecture.

If you have any questions, please let me know.

Otherwise, I will see you in the next lecture. Thank you.