[ENGLISH]

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.

[SPANISH]

Hola, Cloud Gurus, y bienvenidos a esta conferencia.

donde vamos a explorar Elastic Load Balancer.

Entonces, ¿qué es un balanceador de carga?

Bueno, un balanceador de carga distribuye el tráfico de red

a través de un grupo de servidores.

Así que si piensas en un sitio web simple

que consta de 3 servidores web detrás de un balanceador de carga,

el usuario navega al sitio web

y la solicitud llega al balanceador de carga

y luego se enruta a uno de los servidores web.

Y dependiendo de cómo lo haya configurado todo,

el balanceador de carga podría enviar la solicitud al servidor web,

cuál es el menos ocupado,

o podría usar un algoritmo de turno rotativo,

por lo tanto, reenviar solicitudes a cada servidor a su vez.

Y lo bueno de esta arquitectura

es que si uno de los servidores web falla,

entonces el balanceador de carga notará

y dejará de enviar solicitudes a este servidor web

hasta que vuelva a estar en línea.

Y podemos aumentar fácilmente la capacidad cuando sea necesario.

Entonces, si nuestro sitio web de repente se vuelve muy popular,

entonces podemos agregar algunos servidores web más

y registrarlos con el balanceador de carga.

Y el balanceador de carga comenzará a enviar tráfico

a nuestros nuevos servidores web también.

Ahora con AWS Elastic Load Balancer,

hay algunas opciones diferentes para elegir.

Así que tenemos el Equilibrador de carga de aplicaciones,

que equilibra la carga de HTTP y HTTPS.

También está el balanceador de carga de red,

que equilibra la carga del tráfico TCP,

y esta es la opción de alto rendimiento.

También tenemos el Classic Load Balancer,

que puede manejar HTTP y HTTPS

y protocolos TCP también.

Y esta es la opción heredada.

Y también está el Gateway Load Balancer.

Y esta es una oferta bastante nueva.

y puede o no salir en el examen,

pero es bueno tener en cuenta a un alto nivel.

Y este te permite equilibrar cargas de trabajo

para dispositivos virtuales de terceros que se ejecutan en AWS,

como dispositivos virtuales adquiridos en AWS Marketplace,

virtuales de empresas como Fortinet ,

Palo Alto, Juniper o Cisco,

y sistemas de detección o prevención de intrusos

de empresas como CheckPoint y Trend Micro, etc.

Así que echemos un vistazo a cada uno de los balanceadores de carga principales

con un poco mas de detalle

comenzando con el Equilibrador de carga de aplicaciones.

Ahora, esto se usa para equilibrar la carga del tráfico HTTP y HTTPS

y los balanceadores de carga de aplicaciones operan en la capa 7

del modelo OSI, por lo que son conscientes de la aplicación.

Y si no has oído hablar del modelo OSI,

se llama el modelo de 7 capas.

¿Así que qué es lo?

Bueno, es un marco conceptual,

que describe las funciones de una red,

comenzando con la capa de aplicación,

que atiende directamente al usuario final

hasta la capa física y todo lo demás.

Tenemos la capa 7, que es la capa de aplicación,

y esto es todo lo que ve el usuario final.

Entonces HTTP opera en esta capa

y también lo hace su navegador web.

Luego tenemos la Capa 6, que es la capa de presentación,

y esta capa se asegura

que los datos están en un formato utilizable

y protocolos como cifrado y SSH

opera en esta capa.

La capa 5 también se conoce como la capa de sesión.

y esto se trata de mantener conexiones y sesiones.

Luego está la Capa 4, que es otra capa importante,

y también se conoce como la capa de transporte.

Y todo esto se trata de transmitir datos usando protocolos.

como TCP y UDP.

Y, por supuesto, TCP es uno de los principales protocolos

de internet

La capa 3 también se conoce como la capa de red.

Y esta capa está preocupada

con paquetes de red de enrutamiento lógico

basado en direcciones IP.

La capa 2 es la capa de enlace de datos,

y se ocupa de la transmisión física de datos

basado en direcciones MAC.

Y luego está la Capa 1, que es la capa física,

y esto se trata de transmitir bits y bytes

sobre dispositivos físicos,

como los cables y concentradores que componen la red.

Entonces, el balanceador de carga de aplicaciones está operando en la capa 7,

y es en esta capa donde está operando HTTP.

Y en el examen, no se evaluará en el modelo OSI,

pero recuerde que el balanceador de carga de aplicaciones

es compatible con HTTP y está operando en la capa 7

o la capa de aplicación de la pila de red.

Ahora, dado que Application Load Balancer reconoce las aplicaciones,

admiten el enrutamiento de solicitudes avanzadas.

Eso significa que puede enrutar solicitudes a servidores web específicos

basado en el encabezado HTTP.

Así que echemos un vistazo a un ejemplo.

Así que imagínese el sitio web de un concesionario de automóviles,

y quiero que piensen en el tipo de servicios

que puedas encontrar.

Así que definitivamente estarán vendiendo autos,

pero también pueden ofrecer préstamos o contratos de crédito

y servicio y reparaciones.

Y mediante el uso de un balanceador de carga de aplicaciones,

podemos enviar solicitudes a servidores web específicos

dependiendo de lo que el usuario esté buscando.

Podemos enviar consultas de ventas a un conjunto de servidores web,

de préstamo pueden ir a otro,

y si está reservando un servicio o reparación,

que puede ir a otro conjunto de servidores web.

Y el balanceador de carga de aplicaciones

puede manejar todo esto por ti

porque está operando en la capa de aplicación.

Entonces puede usar el encabezado de solicitud HTTP

para determinar dónde enviar cada solicitud.

Paso al Equilibrador de carga de red

y esta es la opción de alto rendimiento.

Así que quiero que pienses en un auto deportivo súper rápido .

Y esto se usa para equilibrar la carga del tráfico TCP

cuando se requiere un rendimiento extremo.

Y opera en la Capa 4 del modelo OSI,

que es la capa de transporte.

Entonces, si volvemos a nuestro modelo de 7 capas,

es en esta capa de transporte de Capa 4 donde opera TCP.

Por lo tanto, Network Load Balancer equilibra la carga

basado en el protocolo TCP.

Y es capaz de manejar

millones de solicitudes por segundo

manteniendo latencias ultrabajas .

Pero recuerda que es el rendimiento más alto,

es la opción más cara.

Pasando a Classic Load Balancers.

Y cuando piensa en Classic Load Balancer,

Solo quiero que pienses en el auto clásico.

Así que no va a ser el más rápido.

y tampoco va a tener las características más increíbles.

Y esta es la opción heredada,

pero todavía puede aparecer en el examen.

Equilibradores de carga clásicos,

admiten algunas características específicas de la capa 7

como X-Forwarded-For encabezados,

que cubriremos en la siguiente diapositiva,

y sesiones pegajosas que simplemente te permiten

para seguir enviando solicitudes

que se originan en la misma sesión

en el mismo servidor web, lo que hace que la sesión sea pegajosa.

Y también es compatible con el equilibrio de carga de capa 4

para aplicaciones que se basan únicamente en el protocolo TCP.

Así que ahora volvamos al encabezado X-Forwarded-For.

Y este es un encabezado HTTP, que le permite identificar

la dirección IP de origen de un cliente

conectándose a través de un balanceador de carga.

Pero ¿qué significa eso?

Bueno, aquí está nuestro cliente y su dirección IP es 124.12.3.231

y aquí está nuestro servidor web

y se está conectando al sitio web a través de un balanceador de carga.

Y digamos el balanceador de carga

tiene una dirección IP privada de 10.0.0.23.

Ahora, cuando la solicitud llega al servidor web,

el servidor web solo verá esta dirección IP privada

del equilibrador de carga, por lo que solo verá 10.0.0.23.

Ahora bien, esto podría causar problemas para su aplicación

porque tal vez quieras saber

dónde vienen todas estas solicitudes.

Es posible que desee asegurarse de que es una red confiable

o que la solicitud proviene de una ubicación

o un país en el que se le permite operar.

Entonces, ¿cómo se puede obtener la dirección IP?

Bueno, es muy fácil

y ahí es donde entra en juego el encabezado X-Forwarded-For.

Así que la dirección IP de origen aparecerá

en este encabezado HTTP X-Forwarded-For.

Entonces, si estamos usando un balanceador de carga,

que admite X-Forwarded-For,

entonces eso significa que podemos identificar

la dirección IP de origen.

Y finalmente,

necesitamos cubrir los errores comunes del balanceador de carga,

y el más común que vas a ver es Error

504, y es posible que vea el mensaje Gateway Timeout.

¿Entonces que significa eso?

Bueno, por lo general significa que el objetivo

o el servidor web o de aplicaciones descendente

ha sabido responder.

¿Asi que que hacemos?

Bueno, lo primero que tenemos que hacer

es comprobar la aplicación.

Podría ser que el Elastic Load Balancer

se pudo establecer una conexión con el objetivo.

Por ejemplo, el servidor web, la base de datos o la función Lambda.

Podría ser que su aplicación esté teniendo problemas.

Y para resolver el problema,

deberá identificar dónde está fallando la aplicación

y solucionar el problema.

Y de hecho experimenté un error 504 recientemente

y yo estaba haciendo un pago en línea

y el pago tardó demasiado en autorizarse,

así que obtuve este error de tiempo de espera de puerta de enlace.

Y no sé qué salió mal,

pero lo que creo que pasó es que mi banco estaba demasiado lento

para autorizar el pago

y obtuvimos un tiempo de espera de puerta de enlace.

Entonces mi transacción falló y tuve que intentarlo de nuevo.

Entonces Gateway Timeout es un error que podrías ver

cuando intenta realizar un pago en línea.

Así que en mis consejos de examen para Elastic Load Balancer.

En primer lugar, tenemos el balanceador de carga de aplicaciones,

que equilibra la carga de HTTP y HTTPS,

y esto proporciona equilibrio de carga inteligente,

que le permite enrutar solicitudes a un servidor web específico

basado en el encabezado de la solicitud HTTP.

Así que piense en el concesionario de automóviles donde están ofreciendo ofertas

y acuerdos de préstamo y reparaciones.

Tenemos el equilibrador de carga de red,

cual es la opcion de alto rendimiento

para tráfico TCP.

El Classic Load Balancer, que es la opción heredada

y admite tanto HTTP como HTTPS y TCP también.

El equilibrador de carga de puerta de enlace,

que es para dispositivos virtuales de terceros.

Y finalmente,

si necesita encontrar la dirección IP IPv4 de su usuario final,

entonces puedes encontrarlo

en el encabezado HTTP X-Forwarded-For.

Así que eso es todo por esta conferencia.

Si tiene alguna pregunta, por favor hágamelo saber.

De lo contrario, te veré en la próxima conferencia. Gracias.