# Network Security at Scale with AWS Gateway Load Balancer

**NOTE: This content will be available as a video course soon! We are making it available as a Learning Path resource in the meantime to help you prepare for the newly updated AWS Certified Solutions Architect - Associate (SAA-C03) exam.**

## Introduction

Hello and welcome to this Cloud Academy presentation.  A little bit about myself, I'm Jorge Negrón and I'm part of the AWS Content development team here at Cloud Academy.

In this course, we introduce the latest addition to the AWS Elastic Load Balancer Family, the AWS Gateway Load Balancer.  It solves the problem of scaling third party virtual network appliance deployments to match the scalability of your applications.

Network appliances examine network traffic both inbound and outbound usually for network security reasons like intrusion detection systems.  Before the introduction of the Gateway Load Balancer scaling a network appliance deployment was difficult to implement for a number of network technical issues.  We are going to clarify the basic architecture for using AWS Gateway Load Balancer and discuss a number of advanced Amazon VPC concepts.

If you have any questions about the material being discussed, please, feel free to contact me at jorge.negron@cloudacademy.com. As an alternative, you can always get in touch with us here at Cloud Academy by sending an email to support@cloudacademy.com where one of our Cloud experts will reply to your question.

### Who should attend this course?

This course is intended for architects, and network engineers looking to understand the basic function and operation of AWS Gateway Load Balancer in Amazon VPCs. This course also covers some of the objectives for both the solutions architect professional certification exam and the AWS Networking Specialty certification exams.

## Learning Objectives

This course is an advanced level AWS Networking course.You will learn about the gateway load balancer and how it's implemented to support virtual network appliance deployments in Amazon VPCs.

## Prerequisites

To get the most out of this course you will need to meet the requirements for any of the associate level certifications by AWS or the equivalent experience.

This session expects that you are fluent with the fundamentals of networking using AWS including VPC, Subnets, Route Tables, and VPC Endpoints among other advanced features. We will have a review of these fundamental ideas for the sake of context in the explanations. For details about these items you can refer to this  course:

**Working with AWS Networking and Amazon VPC**
https://cloudacademy.com/course/amazon-vpc-networking

For details about using elastic load balancing & EC2 Auto Scaling to support AWS Workloads you can refer to this course:

**Using Elastic Load Balancing & EC2 Auto Scaling to Support AWS Workloads**
https://cloudacademy.com/course/using-elastic-load-balancing-ec2-auto-scaling-support-aws-workloads

## Feedback

Feedback on our courses here at Cloud Academy is valuable to us as trainers and any other students looking to take the same course in the future. If you have any feedback, positive or otherwise, Please, share it with us by sending an email to support@cloudacademy.com.

At the time of writing this content, all course information was accurate.  AWS implements hundreds of updates every month as part of its ongoing drive to innovate and enhance its services.

As a result, minor discrepancies may appear in the course content over time.  Here at Cloud Academy, we strive to keep our content up to date in order to provide the best training available.

If you notice any information that is outdated, please contact support@cloudacademy.com.  This will allow us to update the course during its next release cycle.

# Intro to Gateway Load Balancer

Many of AWS services and innovations are customer driven. The AWS Gateway Load Balancer Service is no exception. Many customers have relied on virtual appliances from AWS Partners in the AWS Marketplace.

However, the deployment process and scaling for virtual appliances was difficult to say the least.

First, we would need to be able to direct all traffic inbound and outbound from an Internet Gateway or Virtual Private Gateway to an elastic network interface of a specific EC2 instance in a VPC.

This feature is essential and happens to be implemented using VPC Ingress Routing. Using a VPC Ingress Routing we can forward traffic to a Gateway Load Balancer by updating the route tables in a VPC.

The next feature needed is to deal with IP tunneling such as not to incur errors or conflicts with IP addressing. In a nutshell, we need to be able to grab all traffic inbound and outbound for the VPC and redirect it to a virtual network appliance for security processing AND not interrupt the normal flow and interactions of the request and the reply. The process needs to be transparent.

The AWS Gateway Load Balancer uses a single point of access for all inbound and outbound traffic and allows you to scale your virtual appliances with demand as done with other ELBs like the application load balancer. The Gateway Load Balancer routes traffic through to healthy virtual appliances and stops sending traffic if an appliance becomes unhealthy.

Using Gateway Load Balancer you can also add your own logic into any networking path in AWS when you want to inspect and take action on packets.

The AWS Gateway load balancer sends inbound and outbound traffic transparently over the same consistent route and using the same target.

This implements sticky, transparent and symmetric flow.

# Anatomy of the AWS Gateway Load Balancer

The Gateway Load Balancer consists of two parts.

The first is basically a VPC interface endpoint to the Gateway Load Balancer, Let's call it a VPC Gateway Load Balancer Endpoint. This endpoint is expected to be defined in the VPC where you want to protect the traffic.

The second part is the actual gateway load balancer which sends traffic to a fleet of EC2 instances running third party network appliance software.

The gateway load balancer is required to forward packets without alteration. In order to make this happen the gateway load balancer uses a tunneling protocol called Geneve.

More formally by definition:

*"Geneve is a tunneling mechanism which provides extensibility while still using the offload capabilities of NICs for performance improvement. Geneve works by creating Layer 2 logical networks that are encapsulated in UDP packets"*

If that sounded a bit technical, let's break it down.

A tunnel is created between the gateway load balancer and the fleet of instances on the back end. Traffic is encapsulated and sent through the tunnel to the security appliances implemented in EC2 instances which will examine and act on packets as they're sent or received.

The gateway load balancer encapsulates the packets to the target to provide separation and add some additional information about which "VPC Gateway Load Balancer Endpoint" the packet came from. GENEVE uses port 6081 to get traffic from the GWLB and HTTP - port 80 for health checks:

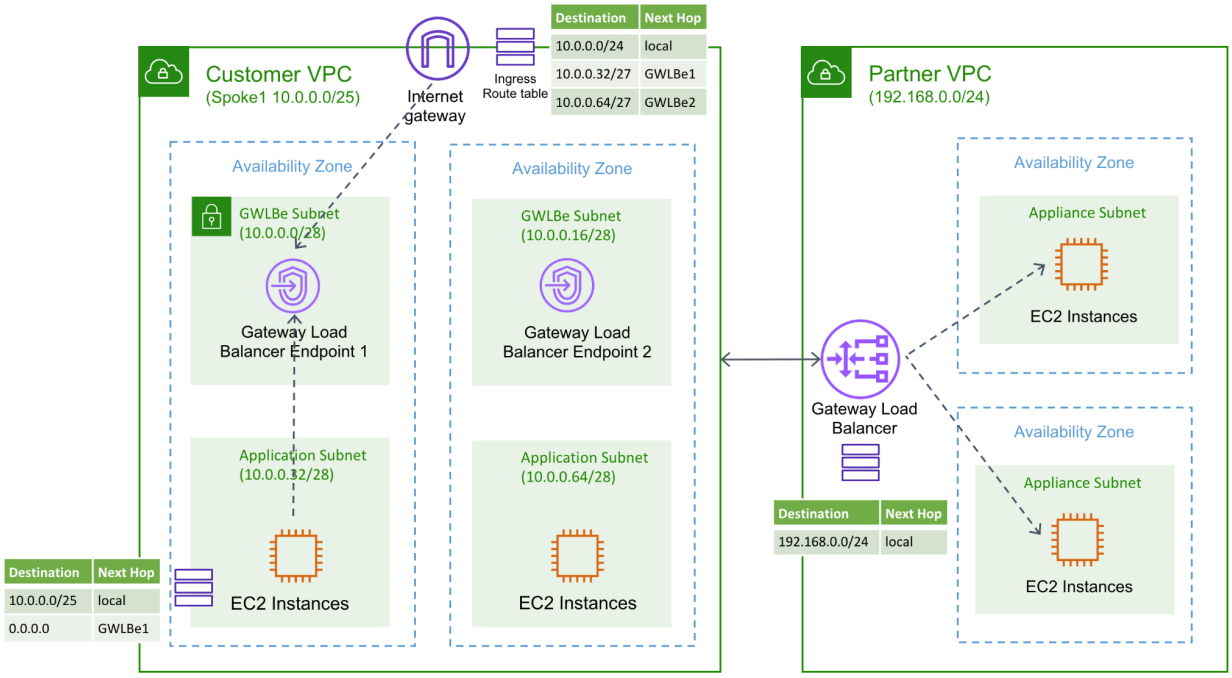



*Image from* [https://d2908q01vomqb2.cloudfront.net/da4b9237bacccdf19c0760cab7aec4a8359010b0/2020/11/09/2020-aws-gwlb-topo-new1.png](https://d2908q01vomqb2.cloudfront.net/da4b9237bacccdf19c0760cab7aec4a8359010b0/2020/11/09/2020-aws-gwlb-topo-new1.png).

# Gateway Load Balancer Architecture

One of the simplest architecture diagrams for the gateway load balancer is shown as a way to review some of the details that we just discussed.

As shown in the diagram it's not unusual to see the Gateway Load Balancer endpoint listed as GWLBe and defined each in their own subnet per availability zone. Gateway Load Balancer endpoints can be added to a route table as the next hop and integrate the GWLB into the traffic flow.

A GWLBe is similar to AWS PrivateLink, operating across many accounts and VPCs with centralized control and administration.

Also note the Ingress Route Table associated with the Internet Gateway in the Customer VPC pointing to the GWLB endpoints accordingly.

In general the process of setting up a gateway load balancer entails provisioning a VPC dedicated to the Gateway Load Balancer and the third party virtual appliance software running on EC2 instances. You provision the GWLB and Target groups as you would an Application or Network Load Balancer.

On the VPC where your application lives create GWLB endpoints on their own dedicated subnets and Update the route tables to include the gateway load balancer endpoints for traffic coming from your applications subnets to them and traffic from the internet gateway to them as well in order to integrate the security VPC to the traffic flow.

In summary, the steps are:

1. Locate the partner's virtual appliance software in AWS Marketplace.
2. Launch the appliance instances in your VPC.
3. Create a GWLB and target group with appliance instances.
4. Create GWLB endpoints in the VPC where the traffic needs to be inspected.
5. Update route tables to make Internet gateway move traffic to and from the gateway endpoints and the GWLB endpoint as next-hop.

Let's discuss the traffic flow that includes a Gateway Load Balancer with corresponding Gateway Interface End-points.

# Traffic Flow Steps for Gateway Load Balancers

In the architecture diagram shown we have an application deployed to private subnets in an autoscaling Target Group serviced by an application Load balancer. The Application Load Balancers are deployed to public subnets.

We also have a separate security VPC with a gateway load balancer and security appliances in a target group for auto scaling. This will allow for the appliance fleet to adjust based on application load.

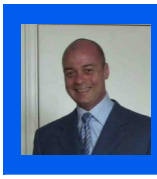
We can follow the steps a packet will travel in this architecture.

1. A Customer accesses your web application.
2. The landing place for public traffic is the application load balancer, however the internet gateway is configured with an ingress route table to direct traffic to a GWLB endpoint.
3. The Gateway Load Balancer endpoint directs traffic to the gateway load balancer in the security VPC.
4. At the gateway load balancer the packets are wrapped using the Geneve tunneling protocol and dispatched through to the security appliance selected.
5. The packet analysis takes place in the security appliance. What actually happens depends on the appliance being used and the configuration defined.
6. After analysis the packets are sent back still encapsulated to the GWLB where the encapsulation is removed and traffic sent to the GWLBe where it originally came from.
7. The corresponding GWLBe will direct traffic to the application load balancer and which will target the application.
8. The response flow is very similar in that the application response will pass through the application load balancer and into the subnet with the GWLBe in the same availability zone. This will send traffic to the GWLB in the security VPC.
9. Packets are encapsulated and sent to the security appliance.
10. Packets are sent back to the GWLB where encapsulation is removed and packets move to the GWLBe.
11. The GWLBe will send traffic out through the internet gateway.
12. The response is received by the customer.

As we get to see from this example flow the gateway load balancer allows you to leverage and horizontally scale third party security appliances from the AWS Marketplace in your VPCs.

## Summary

In this course we introduced the AWS Gateway Load Balancer, its basic architecture in a VPC and the features that allow the AWS Gateway load balancer to send inbound and outbound traffic, transparently, over the same consistent route, and using the same target appliance. This implements sticky, transparent and symmetric security flows for your Amazon VPCs and applications.



**Jorge Negrón, AWS Content Architect**
jorge.negron@cloudacademy.com | linkedin.com/in/jorgetadeonegrondejesus