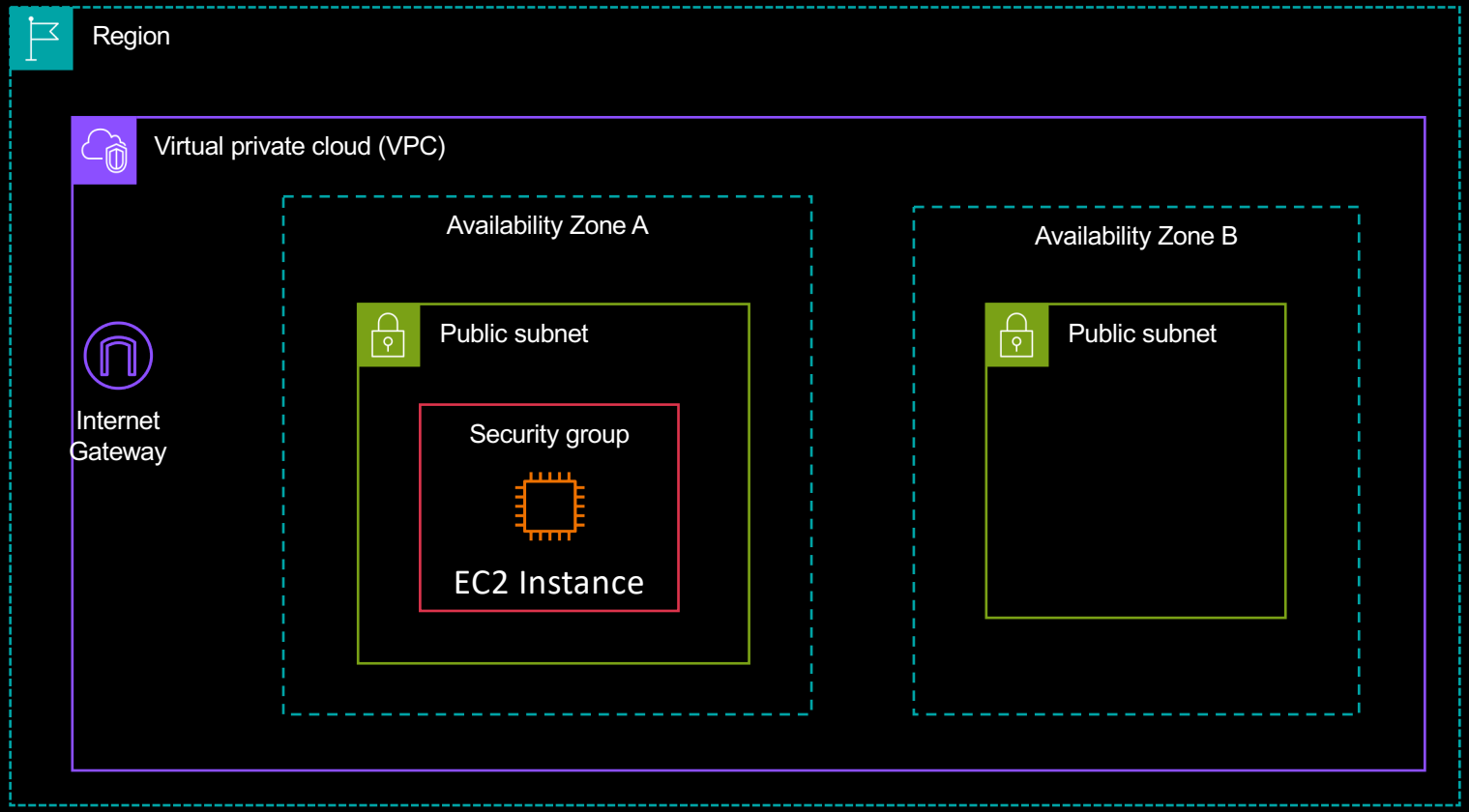
# AWS Solution Architect Training

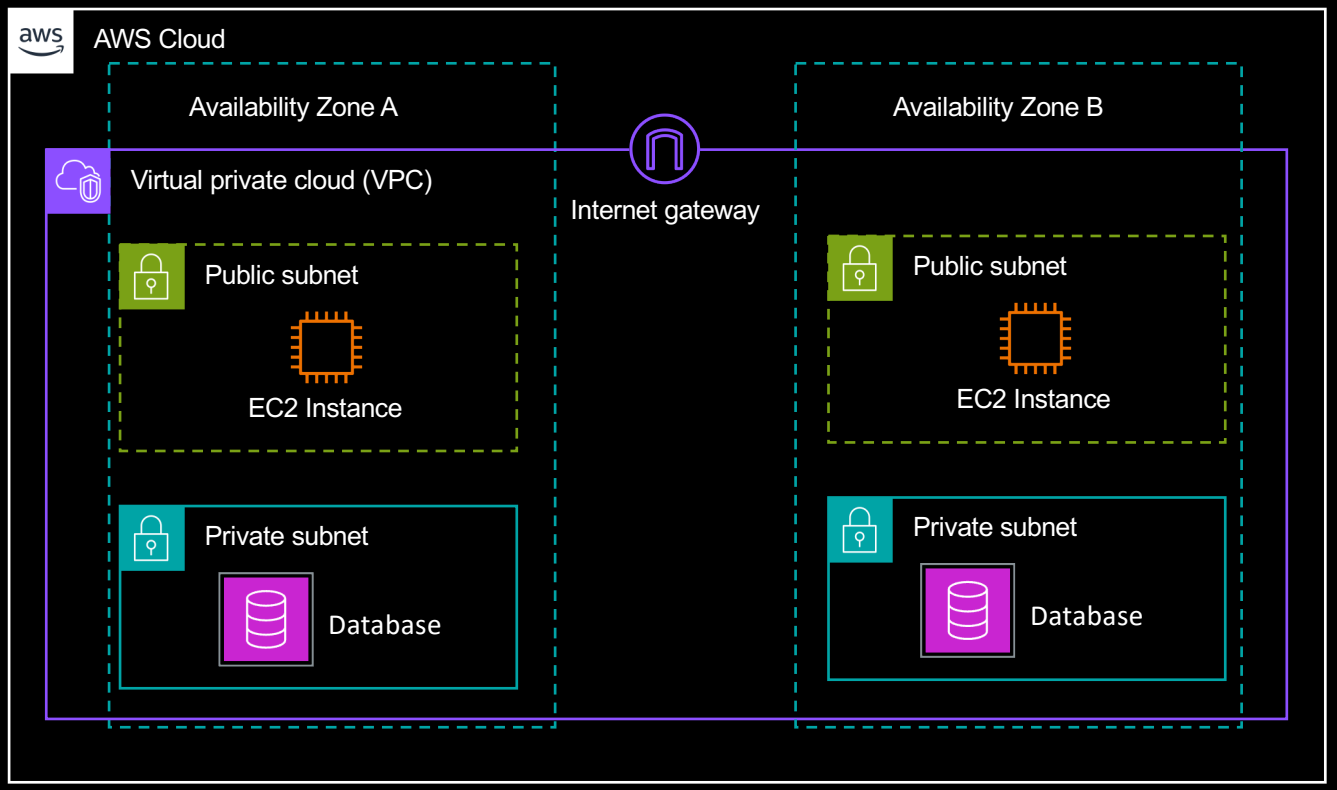
# Module 04
# High Availability and Scalability

Instructor: Tim Platt, Cloud Solution Architect

# How many single points of failure do you see?



# This is better…




## Why?

- Redundancy for critical components
- The EC2 instances are “clones” of one another and are in separate data centers (possibly up to 100 Km apart) – RIGHT?
- Database is redundant too

But is TWO servers enough? Might we have a lot of traffic and need 3 or 4 or 10 or 100?

# Auto Scaling Group (ASG)

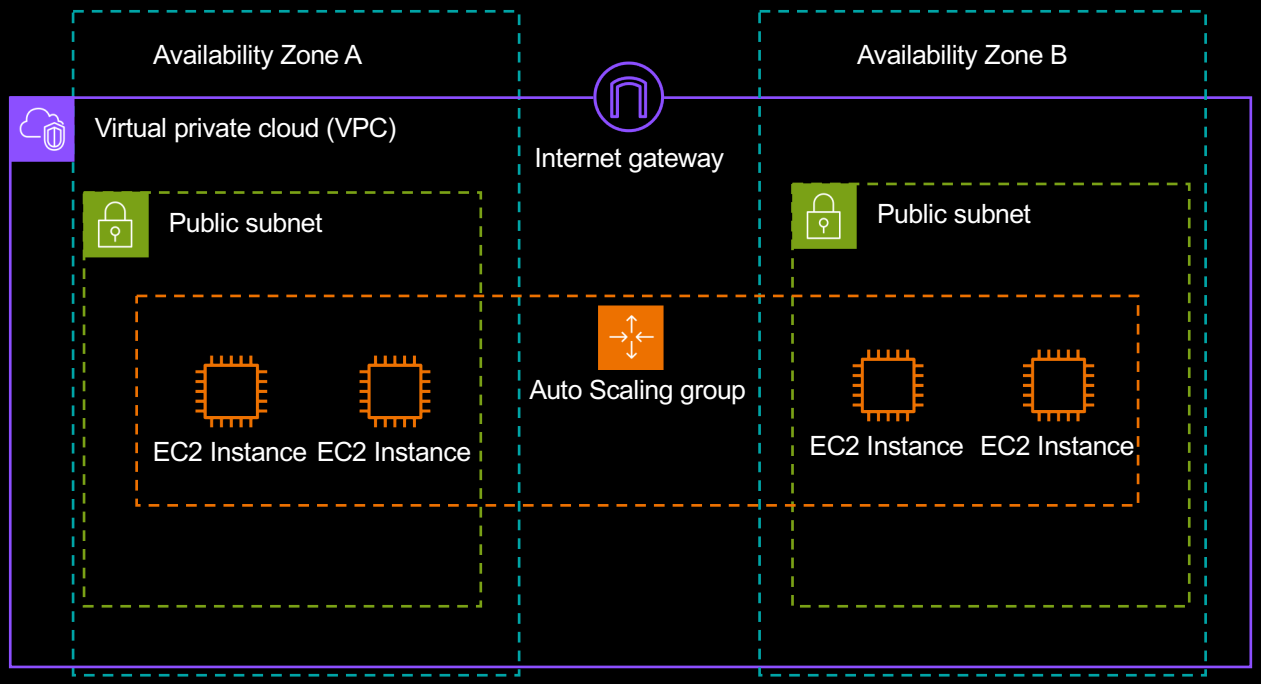
## EC2 Feature

An Auto Scaling Group lets you have a DYNAMIC number of instances (all copies of one another)

## Key Points

- Define a LAUNCH TEMPLATE – a blueprint – to use for creating IDENTICAL (or nearly so) servers
- Set a MINIMUM capacity – perhaps TWO
- Set a MAXIMUM capacity – so you keep $$$$ under control
- You can set a DESIRED capacity manually or using dynamic Scaling Policies (based on CPU%, etc.)
- SUPERPOWER: The Auto Scaling Group can span Subnets (AZs) and will SPREAD the new instances across all the AZs in use!

# This is better…




NOTE: Database not shown for simplicity

## Why?

- We can HORIZONTALLY scale – which means MORE servers when there is more work to do and LESS servers when there is less work to do.
- With 4 t3.medium instances that will be 8 vCPUs total of processing power (each t3.medium has two vCPUs)

**But…**

How will the end user know which server to connect to ???

# Elastic Load Balancer (ELB)

## EC2 Feature

An Elastic Load Balancer is a virtual LOAD BALANCER. It gives a single point of access that will distribute incoming requests to the servers

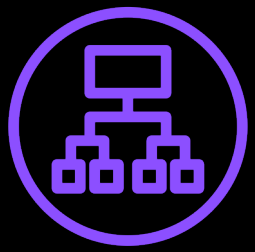
## Key Points

- An Application Load Balancer (ALB - a type of ELB) will have a single un-changing DNS name associated with it.
- The ALB will distribute the incoming HTTPS (Layer 7) requests to a “TARGET GROUP” of servers
- SUPERPOWER 1: The ELB can span Subnets (AZs) and is therefore not a single point of failure
- SUPERPOWER 2: The “Target Group” can be an ASG
- SUPERPOWER 3: The ELB can check the HEALTH of the targets and STOP sending traffic to failed EC2 instances

# Elastic Load Balancer (ELB)

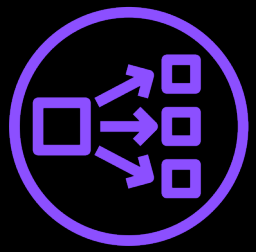
There are three main types of ELB

## Application Load Balancer (ALB)



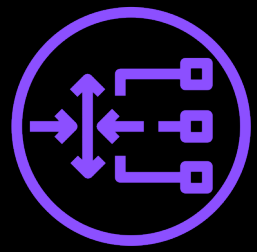
- HTTP/HTTPS
- Layer 7
- Best for Web Apps and REST API
- Has a single DNS name

## Network Load Balancer (NLB)



- TCP (Transmission Control Protocol) and UDP (User Datagram Protocol)
- Layer 4
- Most scalable option
- DNS Name, but also a STATIC IP address

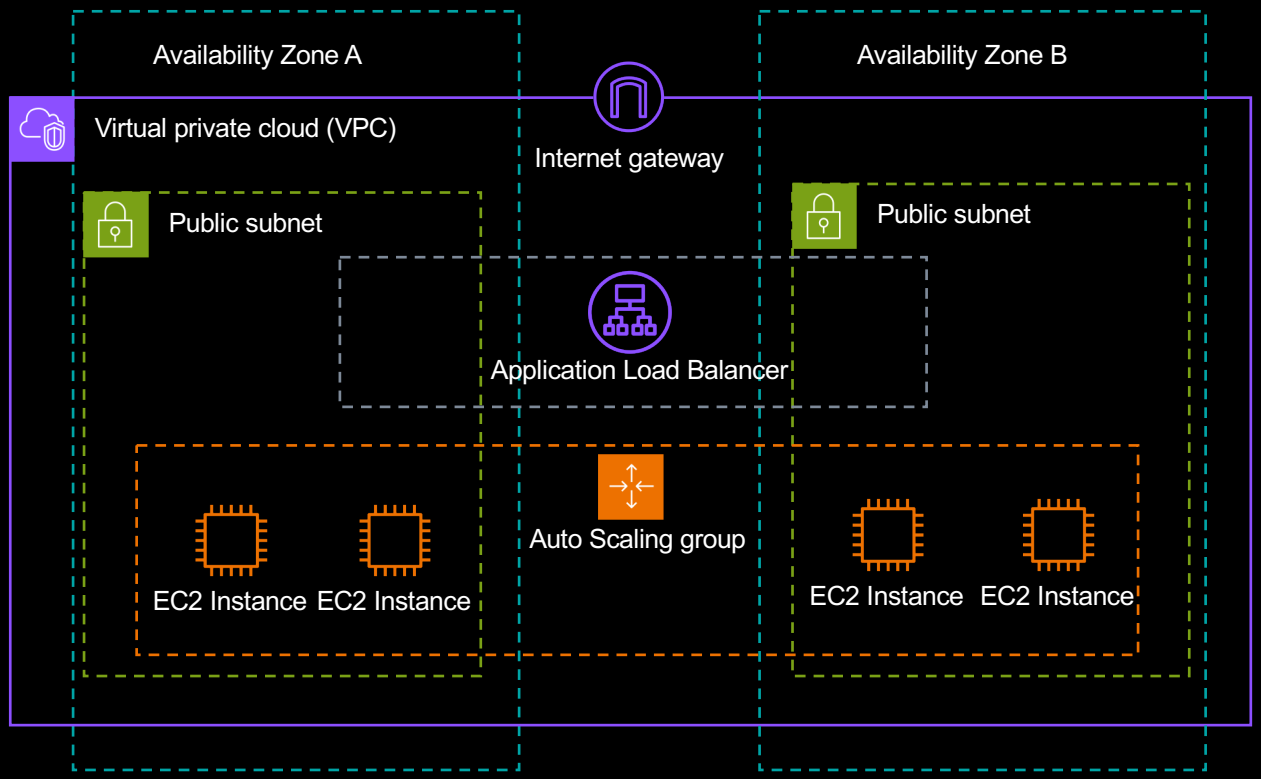
## Gateway Load Balancer (GLB)



- IP (Internet Protocol)
- Layer 3
- Use for security appliances – such as deep packet scanning

NOTE: The Classic Load Balancer (CLB) is still available, but using one of these newer types is always recommended

# Even better…




## Why?

- We have a single , stable endpoint (DNS name) we can direct the users to that will balance the load across the HEALTHY servers available at any point in time

# CloudTrail

Enables auditing, security monitoring, and operational troubleshooting by tracking user activity and AWS API usage

## Key Points

- An Audit Trail of AWS API calls – good for:
  - Who is changing stuff?
  - Who launched that server?
  - Who deleted that database?
  - Is someone trying to do things that they are not allowed to do? (possible hacker intrusion!)
- Don’t get CloudWatch (Monitoring – Metrics, Traces, Logs) confused with CloudTrail
- SUPERPOWER: Understand what’s going on inside your AWS Account

# Links

- Auto Scaling Groups (ASG): https://docs.aws.amazon.com/autoscaling/ec2/userguide/auto-scaling-groups.html
- Launch Templates: https://docs.aws.amazon.com/autoscaling/ec2/userguide/launch-templates.html
- Ways to scale your ASG: https://docs.aws.amazon.com/autoscaling/ec2/userguide/scale-your-group.html
- ELB Documentation: https://docs.aws.amazon.com/elasticloadbalancing/