Auto Scaling and Monitoring

Alayna Lee

Elastic load Balancing

Helps redirect traffic from a network or application in availability zones

Scales your load balancer based on the traffic changes
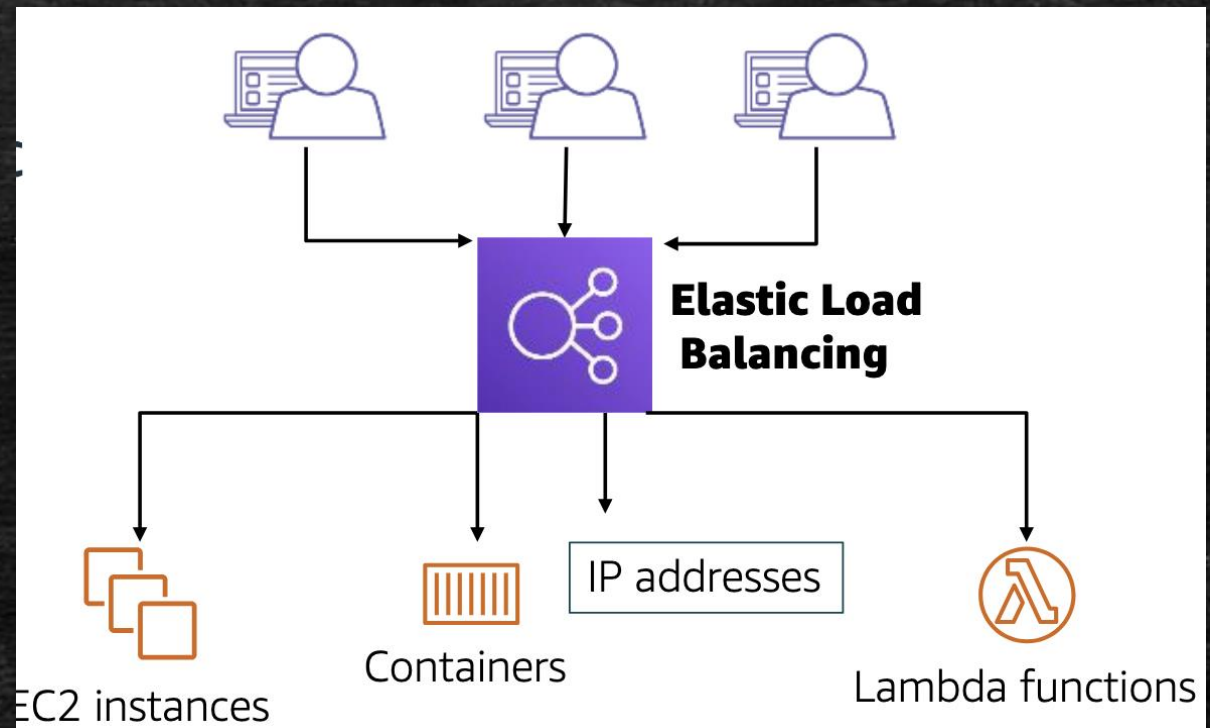
Application load balancer

Load balances HTTP and HTTPS

Network load balancer

Load balances TPC and TLS where high performance is needed

Classic load balancer

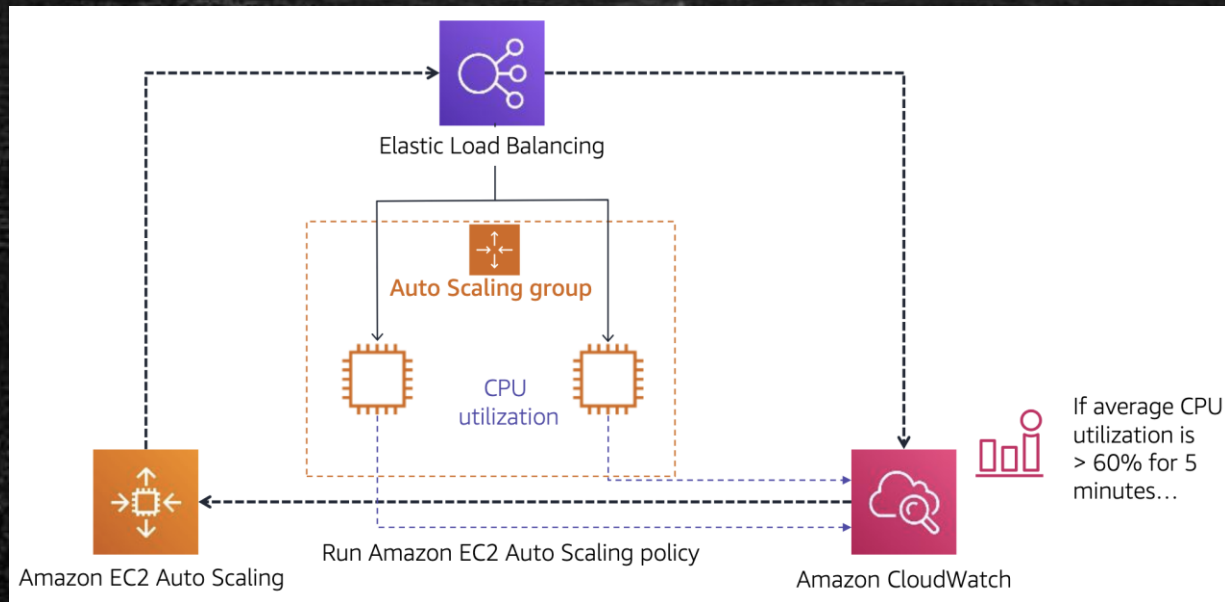Mix of both application and network load balancers

AWS Cloudwatch

- Helps users monitor AWS resources so that you can use the services to their best ability

- Collects and tracks standard metrics

- Performs EC2 actions or scaling

- Allows users to define events to match changes in the AWS environment

- Alarms:
    - Used to detect anomalies
    - Static threshold
    - You must specify:
    - Namespace
    - Metric
    - Statistics
    - Period
    - Conditions
    - Actions

Amazon EC2 Auto Scaling

Elastic Load Balancing

Auto Scaling group

CPU utilization

If average CPU utilization is > 60% for 5 minutes...

Run Amazon EC2 Auto Scaling policy

Amazon EC2 Auto Scaling

Amazon CloudWatch

- Scaling is the ability to increase or decrease the amount of computing capacity for your application

- Helps you maintain application availability

- Detects bad instances and replaces them without affecting the application

- Provides different scaling options

- Scaling groups:
  o Collection of EC2 instances that help auto scale

- Dynamic Scaling
  o Create a CloudWatch alarm that is based on performance and make the event trigger the scaling group to auto scale

AWS auto scaling

Monitors your applications and automatically adjusts to the capacity of computing to maintain steady performance

Simple UI that allows you to plan scaling for resources like EC2 instances, ECS, DynamoDB, and Aurora replicas