Scalable Computing in AWS

Ryan Lewis
CLOUD ENGINEER
@ryanmurakami ryanlewis.dev

Meet User Demands with Scaling

Overview

Scalability !== Elasticity

Launching a launch configuration

Balancing requests with an ELB

Scale in your sleep

When to scale?

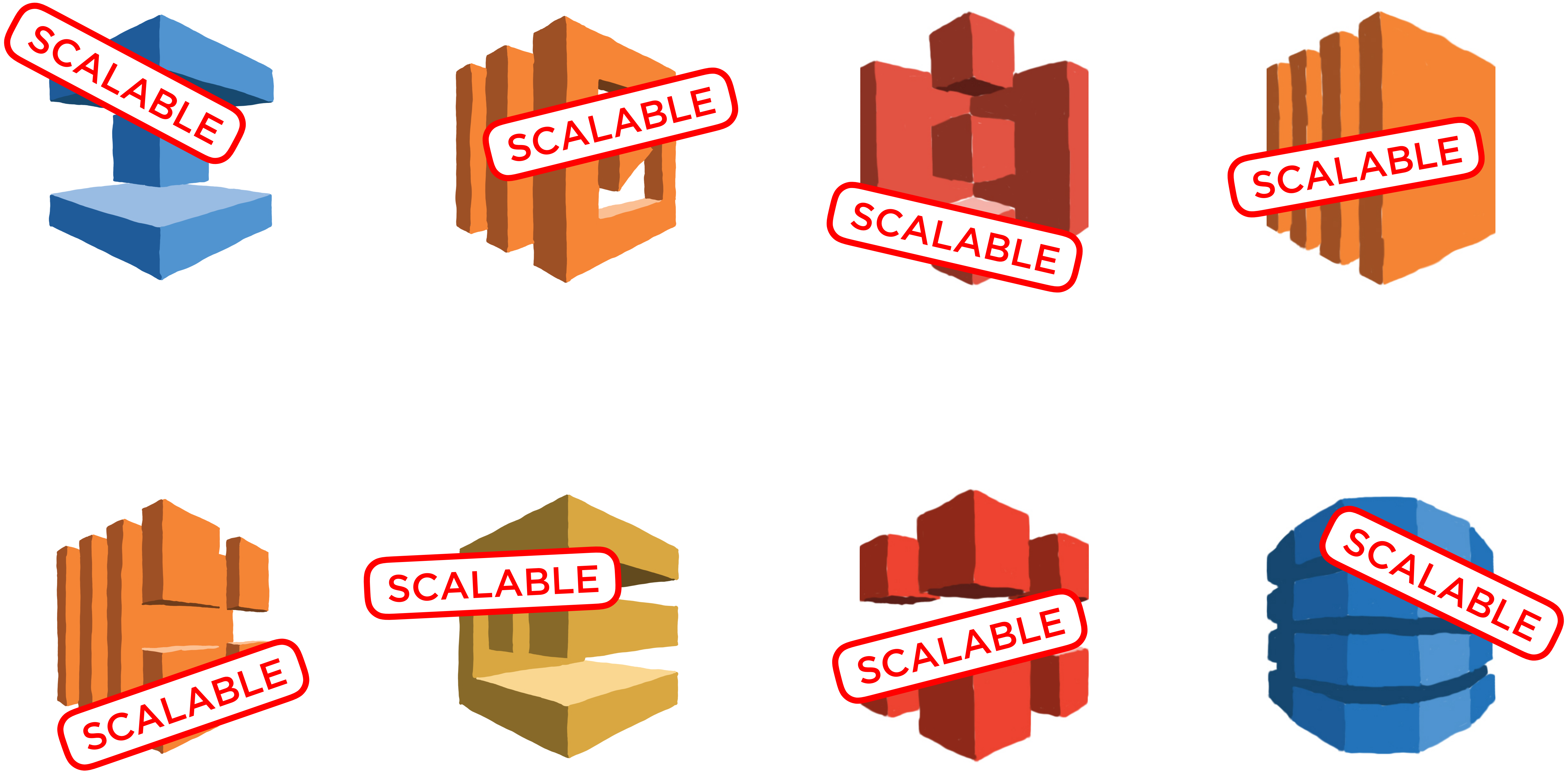
Limits to the scaling magic

Understanding Scalability and Elasticity with AWS

Scalability !== Elasticity

Scalability

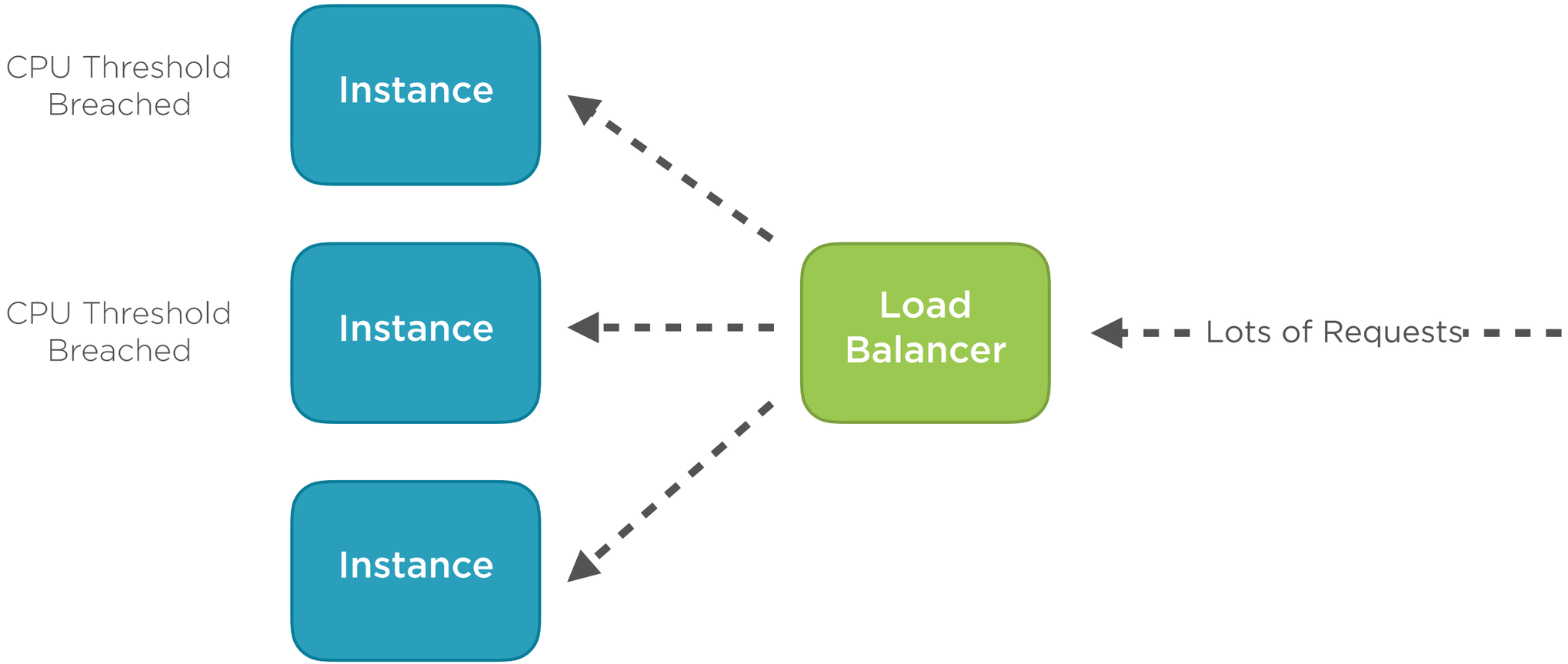
Increasing or decreasing the size or quantity of a resource in AWS.


SCALABLE
SCALABLE
SCALABLE
SCALABLE
SCALABLE
SCALABLE
SCALABLE
SCALABLE

Elasticity

Scaling in response to preset rules, often triggered by CloudWatch alarms.

Elasticity Example

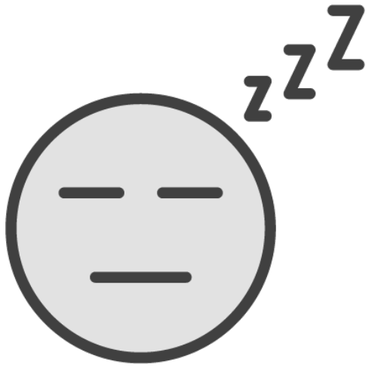


Scalability

\+

Elasticity

=

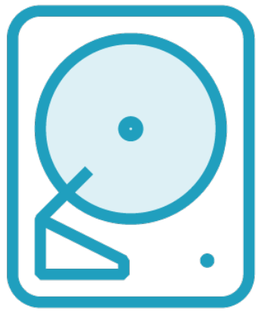


Creating a Launch Configuration

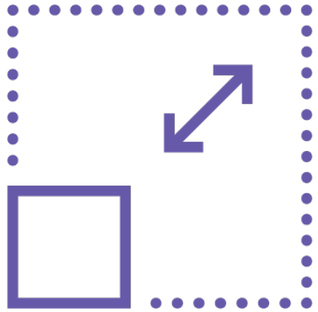
Launch Configuration

Blueprint for creating an EC2 instance.

Launch Configuration Attributes

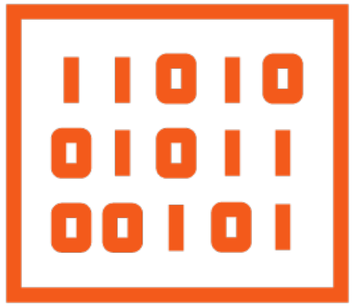


AMI



Instance type

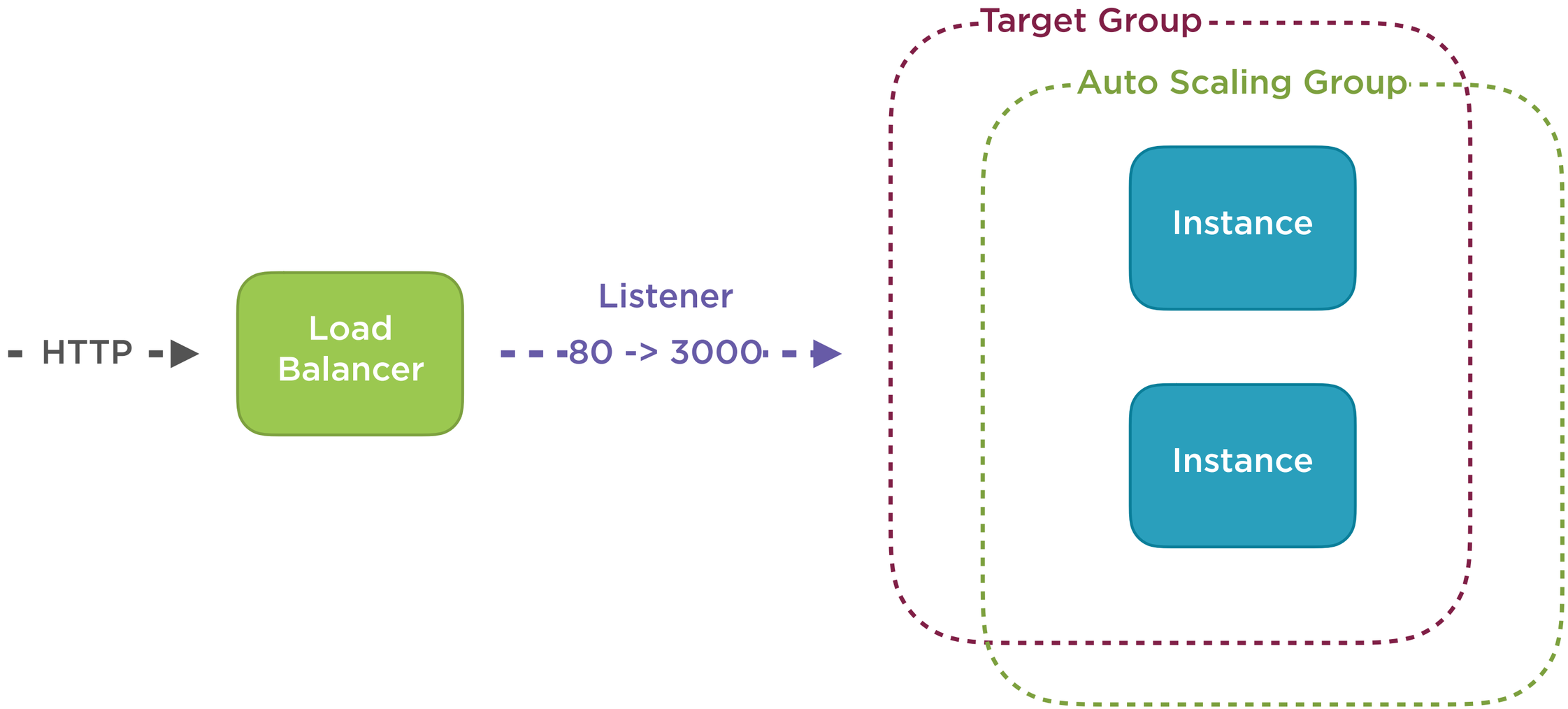
Security groups



User data

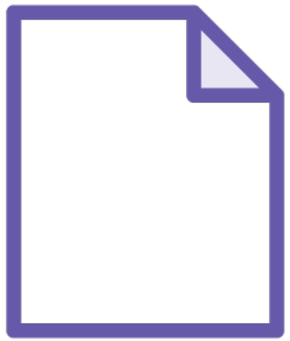
Auto Scaling groups depend on launch configurations

Creating an Elastic Load Balancer


HTTP
Load Balancer
Listener
80 -> 3000
Target Group
Auto Scaling Group
Instance
Instance

Creating an Auto Scaling Group

Configuring Auto Scaling Policies



Simple Scaling Policy

Monitor attributes and perform action

One action per alarm

Scaling up and down requires two policies

Step Scaling Policy

Define multiple actions per alarm

Continuously perform actions

Target Tracking Scaling Policy

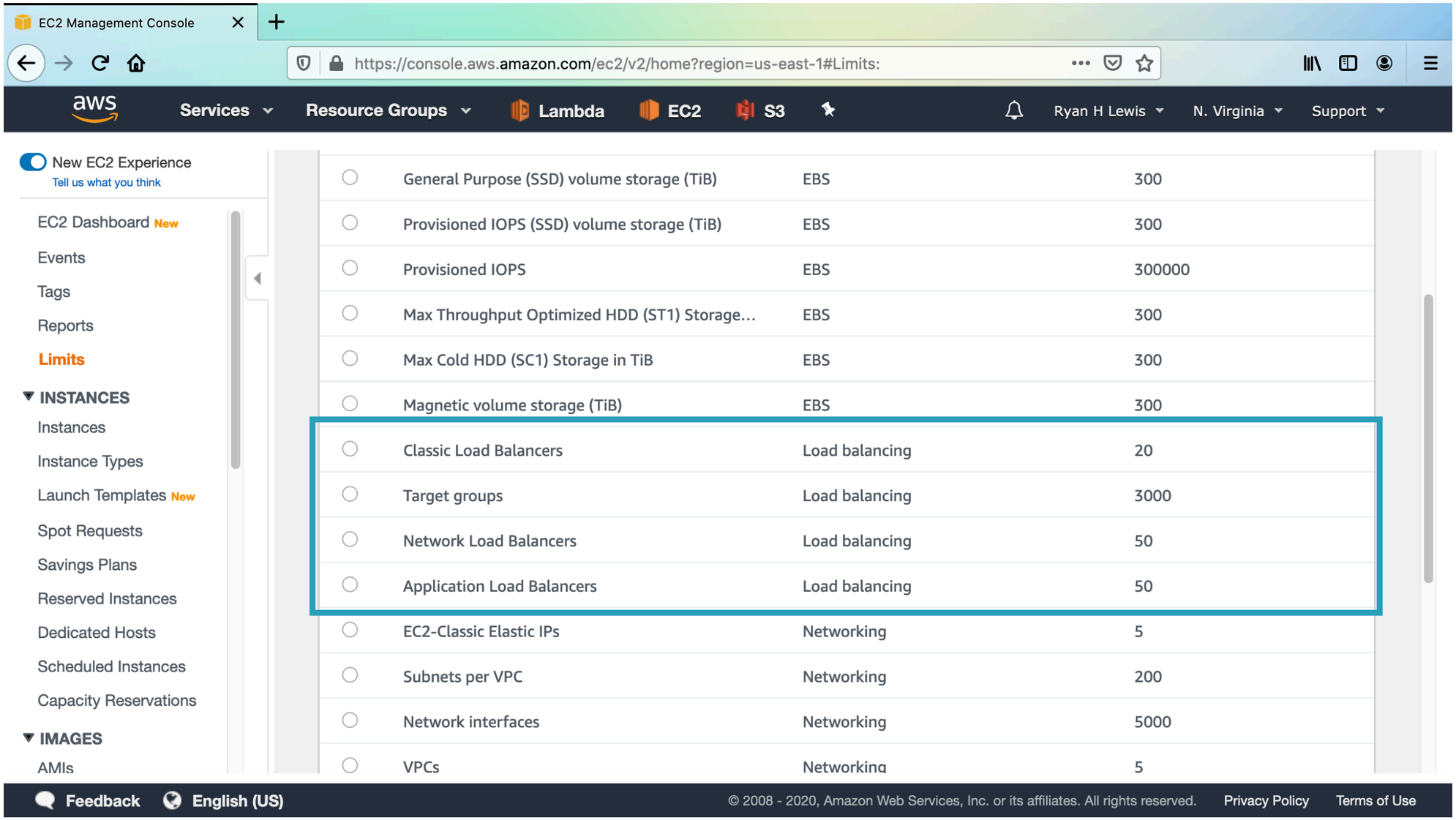
Define metric target

Auto Scaling up and down to achieve target

AWS recommended scaling policy

Limits with Auto Scaling and ELB

EC2 Limits Page



Auto Scaling Group Limit

Soft limit on number of groups and launch configurations

Elastic Load Balancing Limit

Soft limit on number of application/network load balancers

Elastic Load Balancing Limit

One SSL Certificate per load balancer

Elastic Load Balancing Limit

One target group per load balancer

Conclusion

Summary

Scalability or elasticity? Why not both?!

A configuration to launch from

Who balances the load balancers?

Automobile skilling grape

ASG policy dictatorship

Mind the resource limits

Up Next

Storage in AWS