aws

DevOps 90%

إحمد مروي

METWALLY LABS

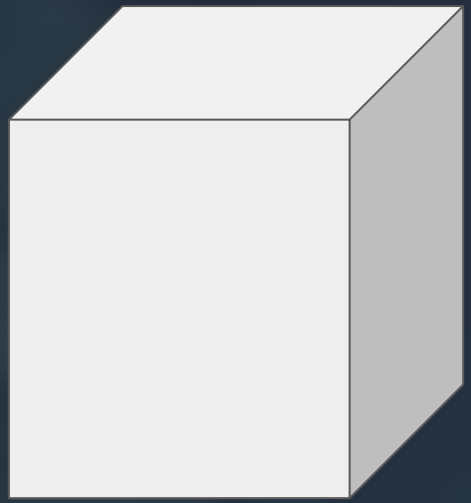
CLOUD NATIVE BASE CAMP

EC2

Auto Scaling Group

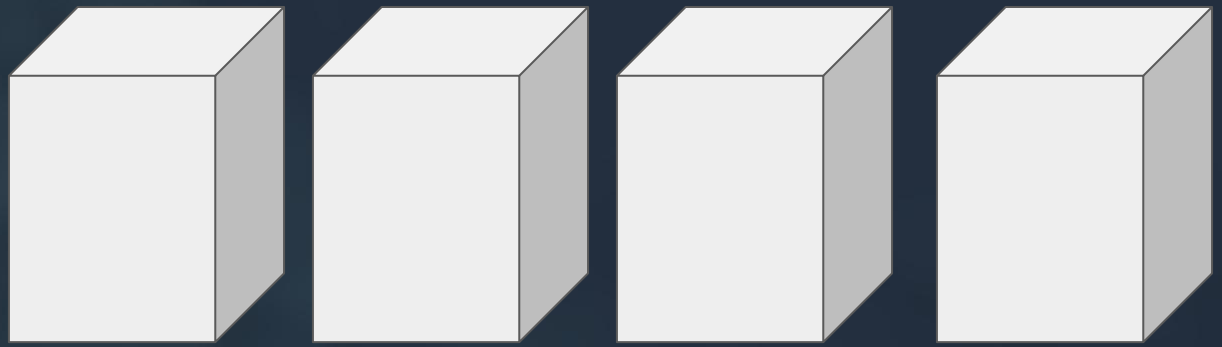
Scalability

is a property of a system that can grow according to current load



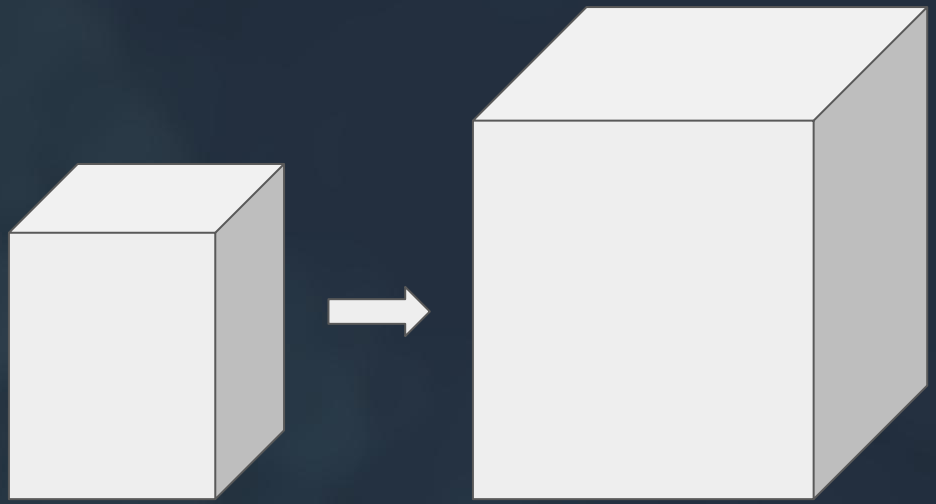
Scalability

is a property of a system that can grow according to current load



Scalability

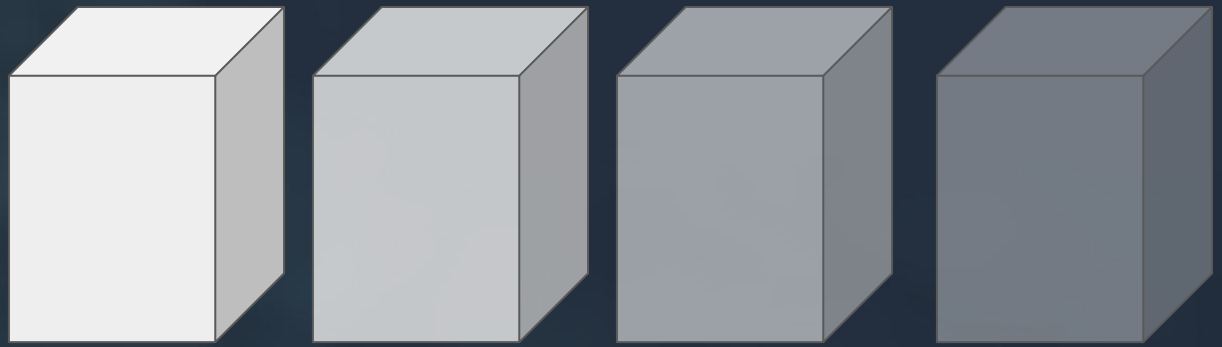
is a property of a system that can grow according to current load



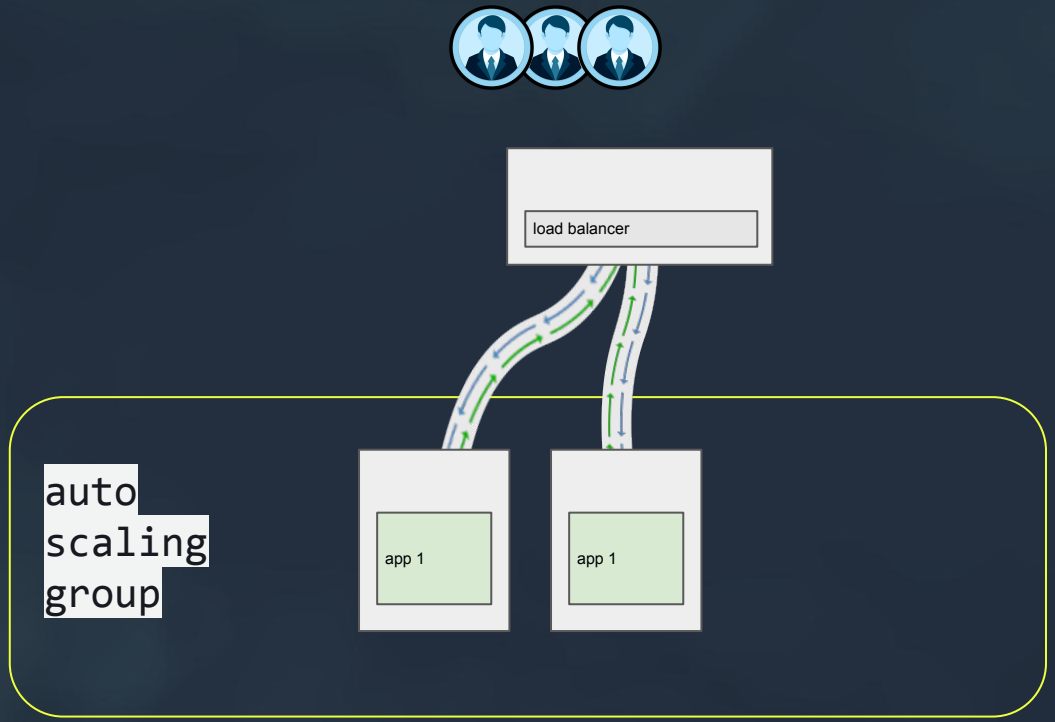
vertical scale in scale up

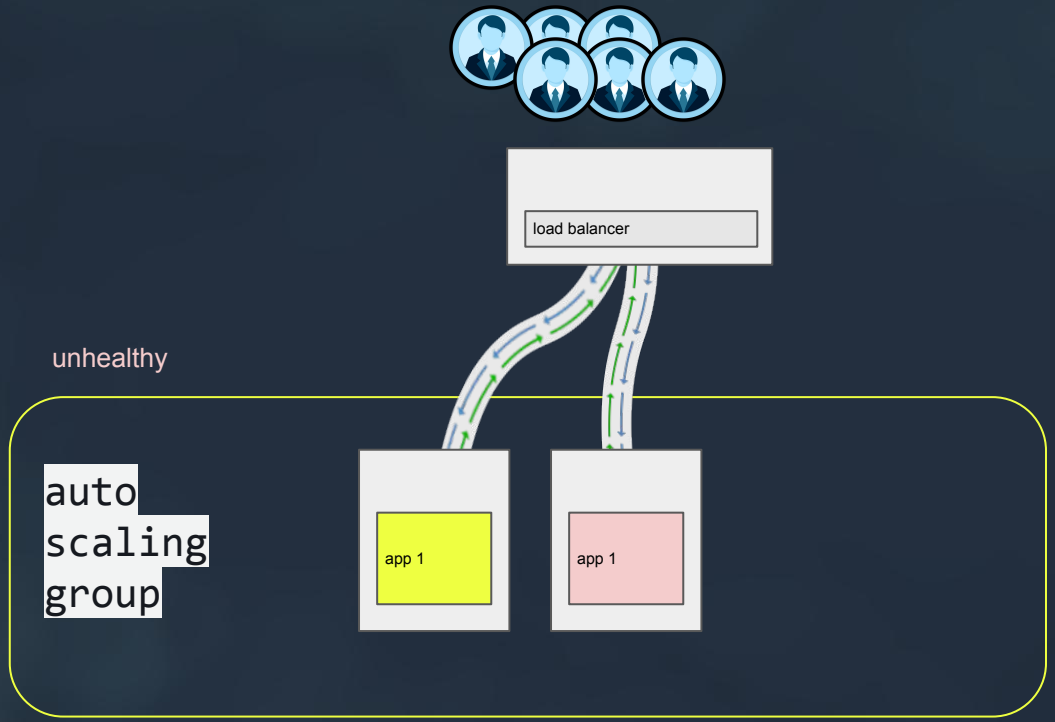
Scalability

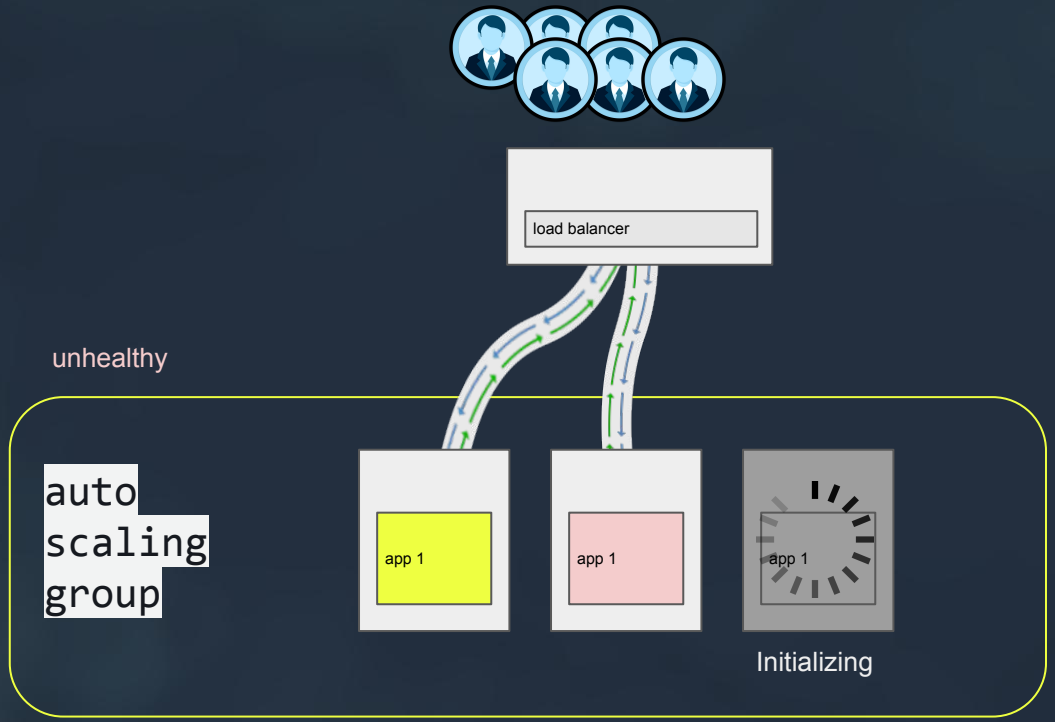
is a property of a system that can grow according to current load

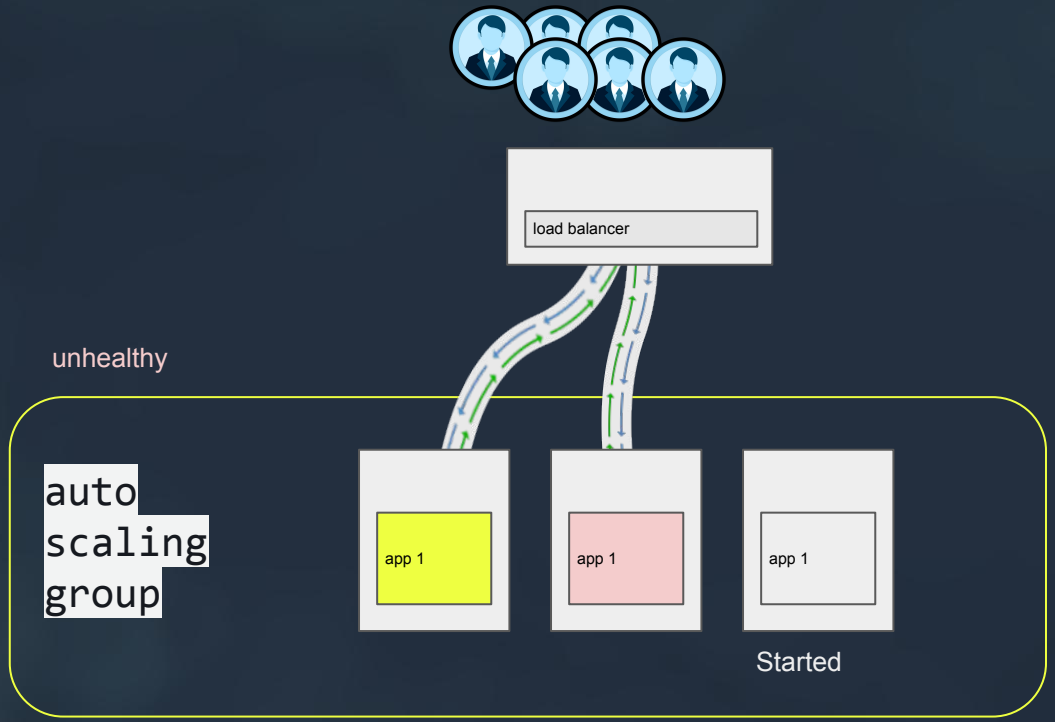


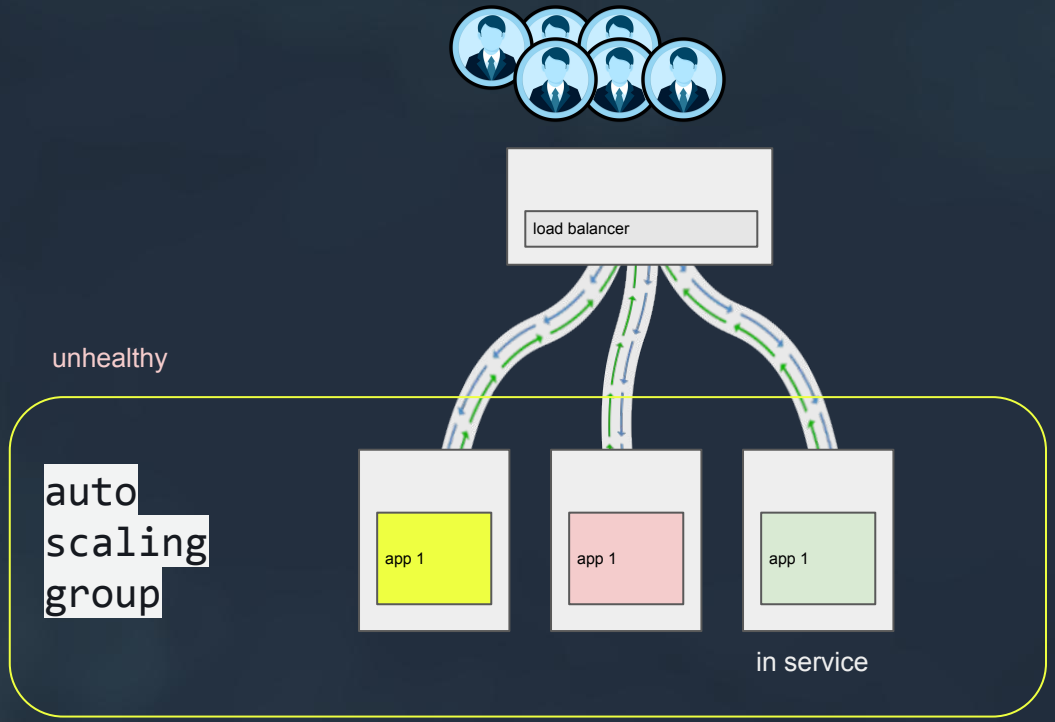
Horizontal scale out

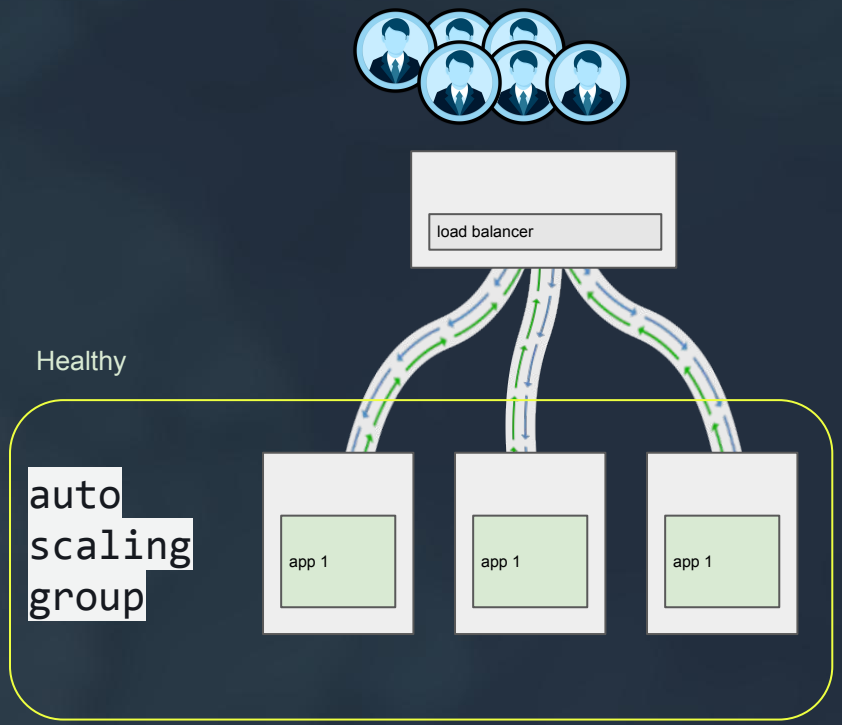

load balancer
auto
scaling
group
app 1
app 1

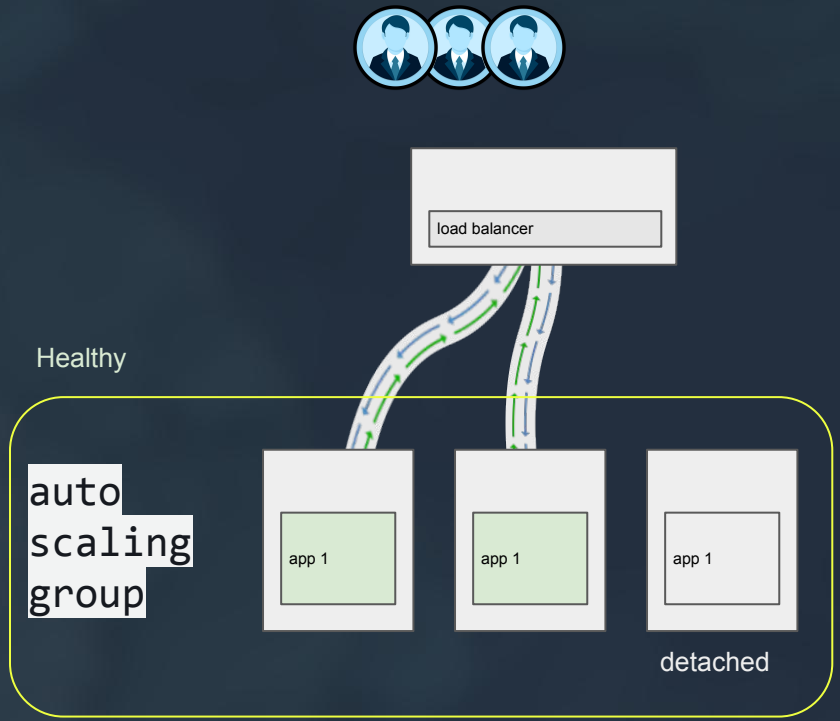

load balancer
unhealthy
auto
scaling
group
app 1
app 1

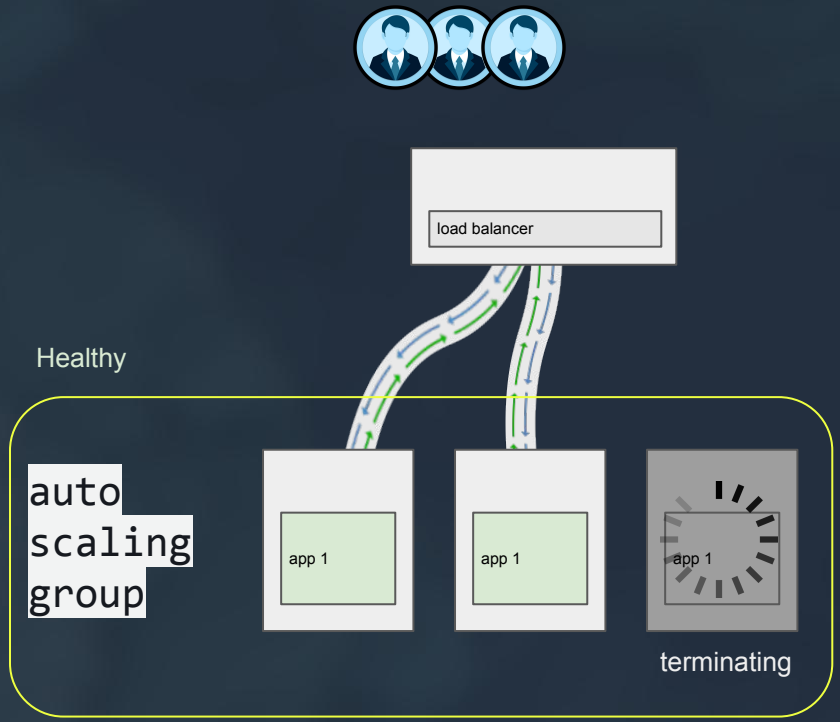

load balancer
unhealthy
auto
scaling
group
app 1
app 1
app 1
Initializing

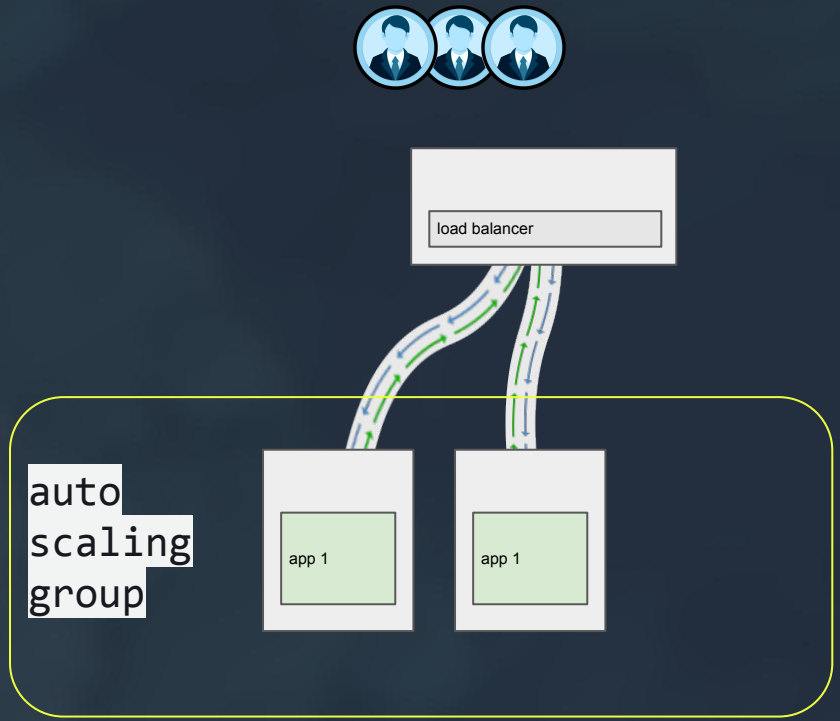

load balancer
unhealthy
auto
scaling
group
app 1
app 1
app 1
Started


load balancer
unhealthy
auto
scaling
group
app 1
app 1
app 1
in service

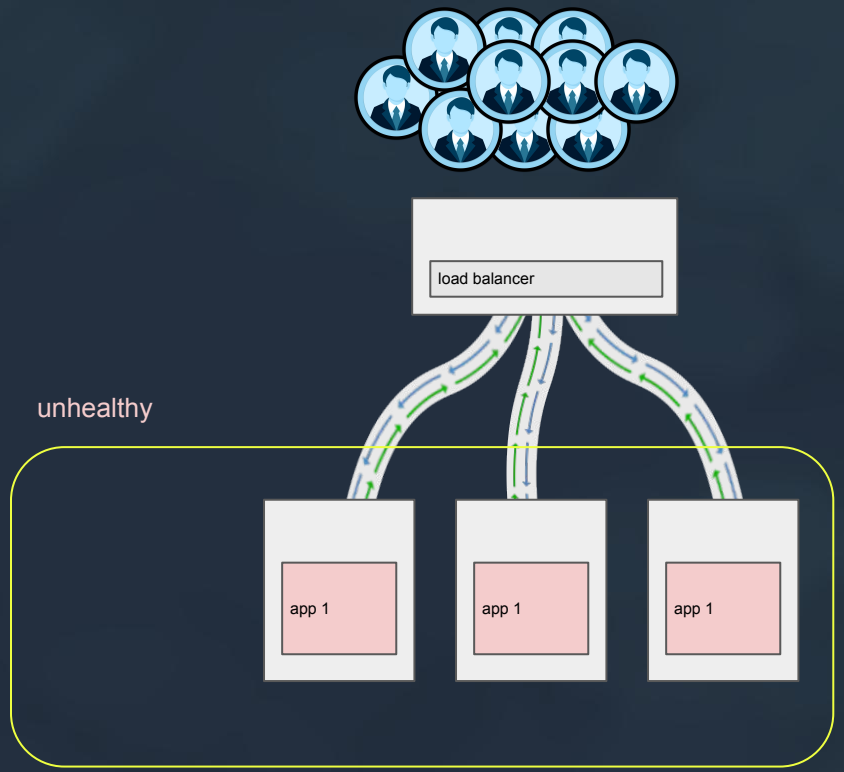

load balancer
Healthy
auto scaling group
app 1
app 1
app 1


load balancer
Healthy
auto scaling group
app 1
app 1
app 1
detached

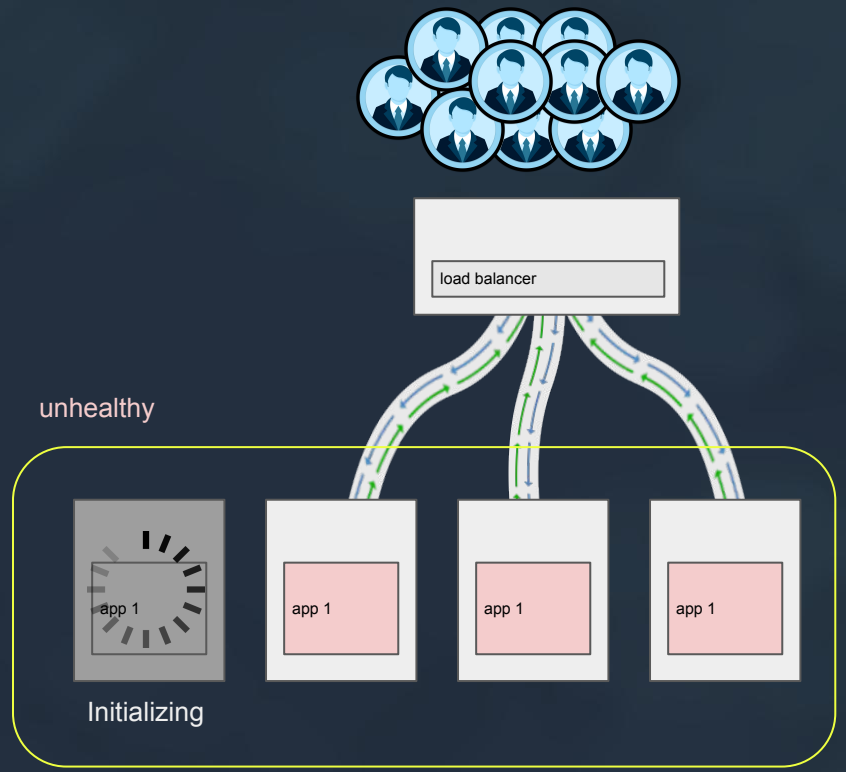

load balancer
Healthy
auto scaling group
app 1
app 1
app 1
terminating

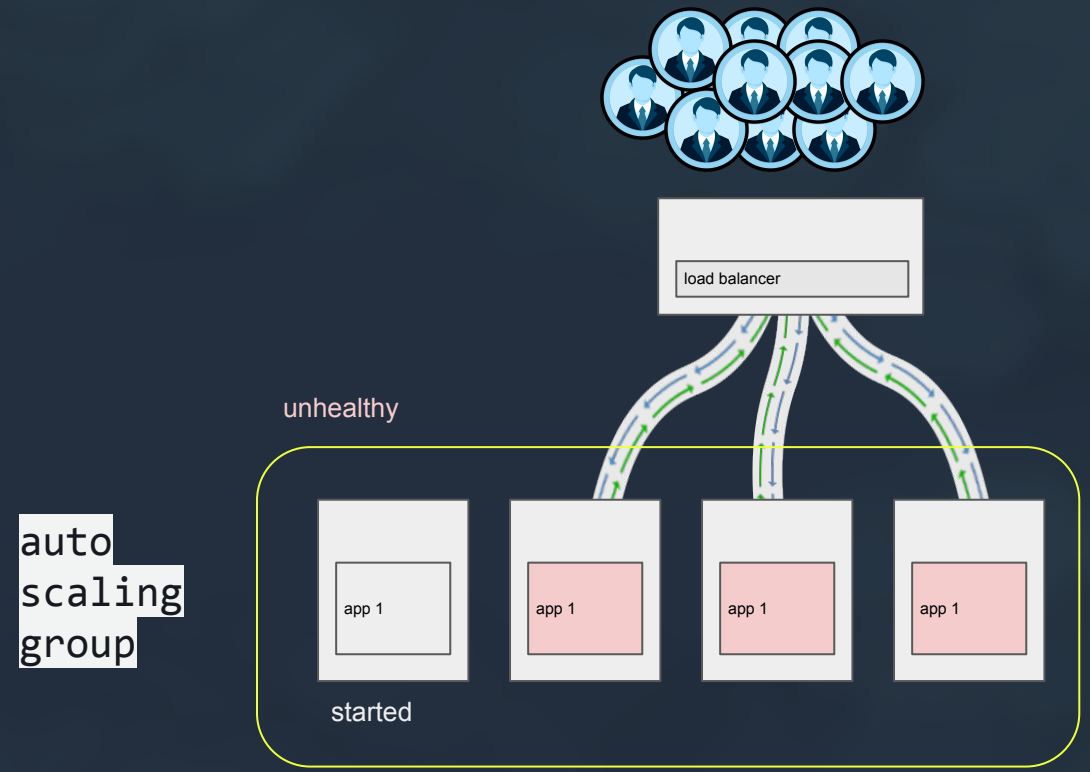

load balancer
auto
scaling
group
app 1
app 1

auto
scaling
group



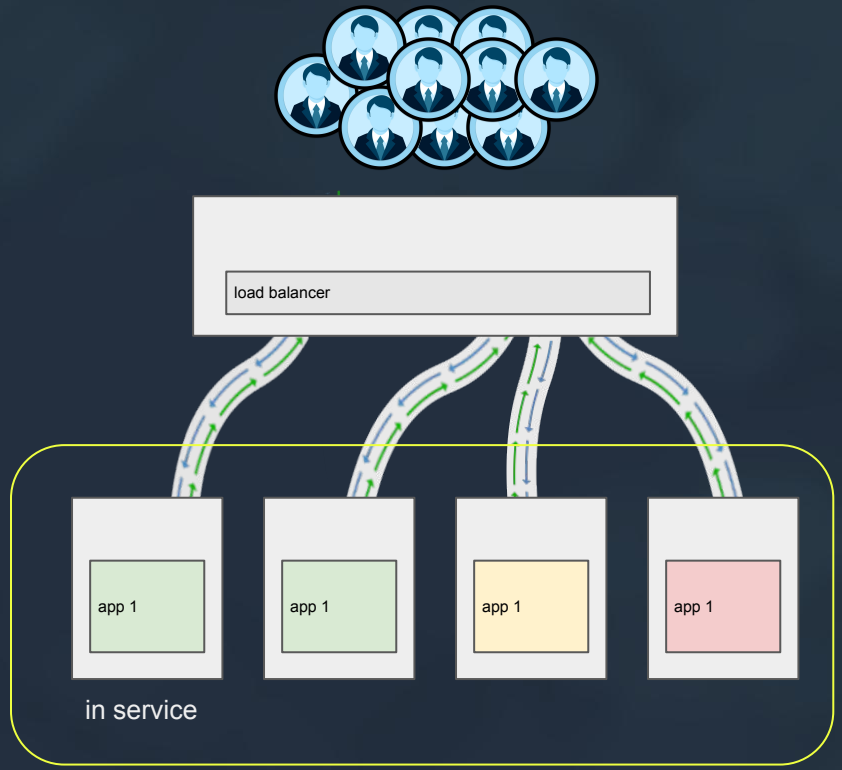
auto
scaling
group



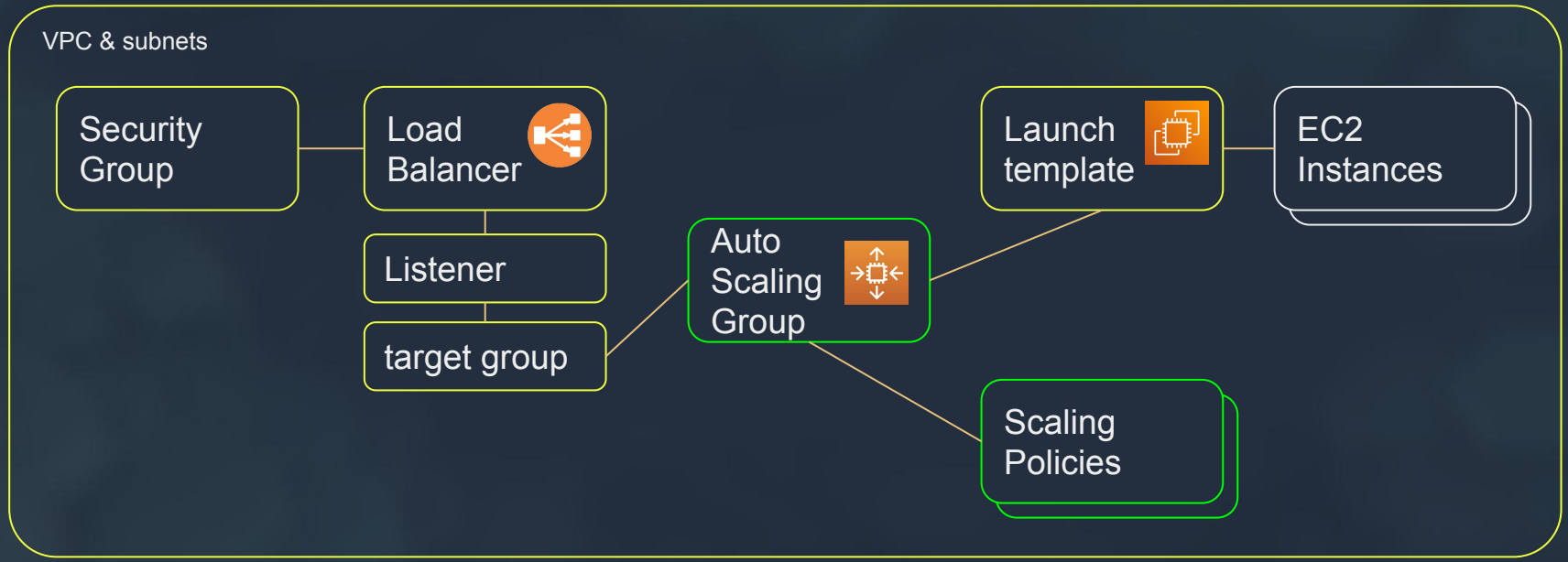


auto
scaling
group

auto
scaling
group



auto scaling group & related resources



Scaling Policies

Target Tracking Scaling

Depends on cloud watch metric value to scale up and down the EC2s

increase instances one by one

cloud watch alarms will created on your behalf

Step Scaling

Depends on cloud watch metric value to scale up or down the EC2s

increase instances by specific count according to metric values

could have more than one step

you need to create the alarm yourself

Simple Scaling

Depends on cloud watch metric value to scale up or down the EC2s

increase instances by specific count whatever the metric value

just one rule

you need to create the alarm yourself

Predictive Scaling

Scale daily or weekly according to traffic flow

using machine learning based on cloud watch historical data

auto scaling group creation (console)

- create launch template
- create auto scaling group with load balancer
  - select launch template
  - select same vpc that configured in the template
  - attach to a new load balancer
    - network load balancer
    - internet-facing
    - configure tcp port and target group
  - no VPC Lattice
  - enable load balancer health check
  - configure group size (desired, minimum and maximum capacity)
  - enable target tracking scaling policy
    - average cpu utilization
    - configure target value
  - no notification
  - no tags

auto scaling group creation (cli)

Ingredients:

- VPC
- Subnets
- Security Group
- Launch Template
- Load Balancer
  - Listener
  - Target Group
- Auto Scaling Group
- Scaling Policies

Directions:

- Create VPC, Or retrieve the Id of existing one
- Create at least 2 Subnets, Or retrieve their Ids
- Create Security Group, Or retrieve Id.
- Create Launch Template, Or use existing one's name.
- Create Load Balancer or retrieve ARN.
  - Create Listener or just use existing one
  - Create Target Group or retrieve ARN.
- Create Auto Scaling Group
- Attach Scaling Policies to the scaling group