Auto Scaling Introduction

Auto Scaling helps you in maintaining correct number of EC2 instances according to application workload. We can create collections of instances called Auto Scaling Groups.

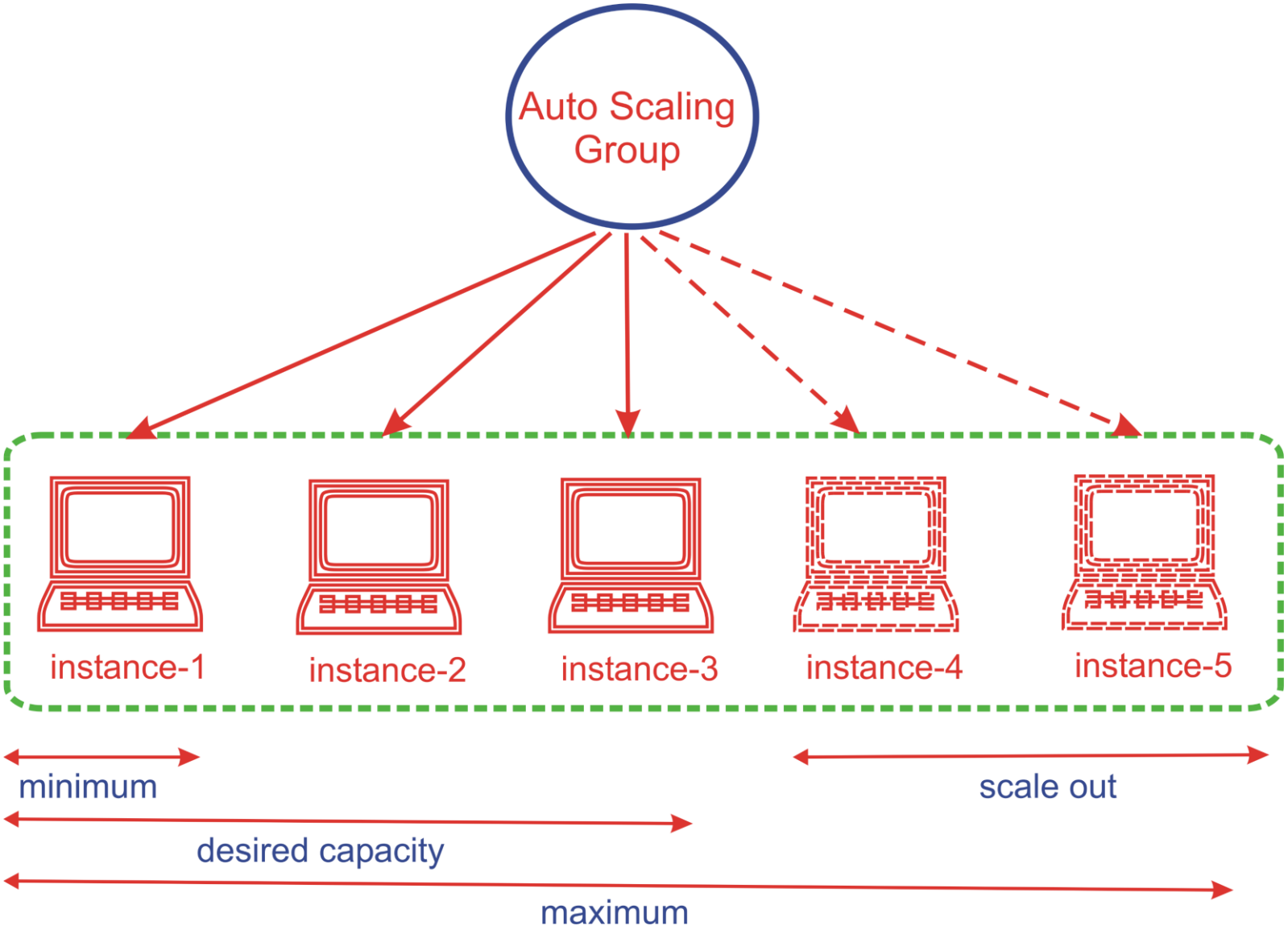
We can specify minimum number of instances in AS Group and AS will ensure that instances do not go below that.

We can specify maximum number of instances in AS Group and AS will ensure that instances do not go above that.

If we specify desired capacity then AS ensures that you always have fixed number of instances.

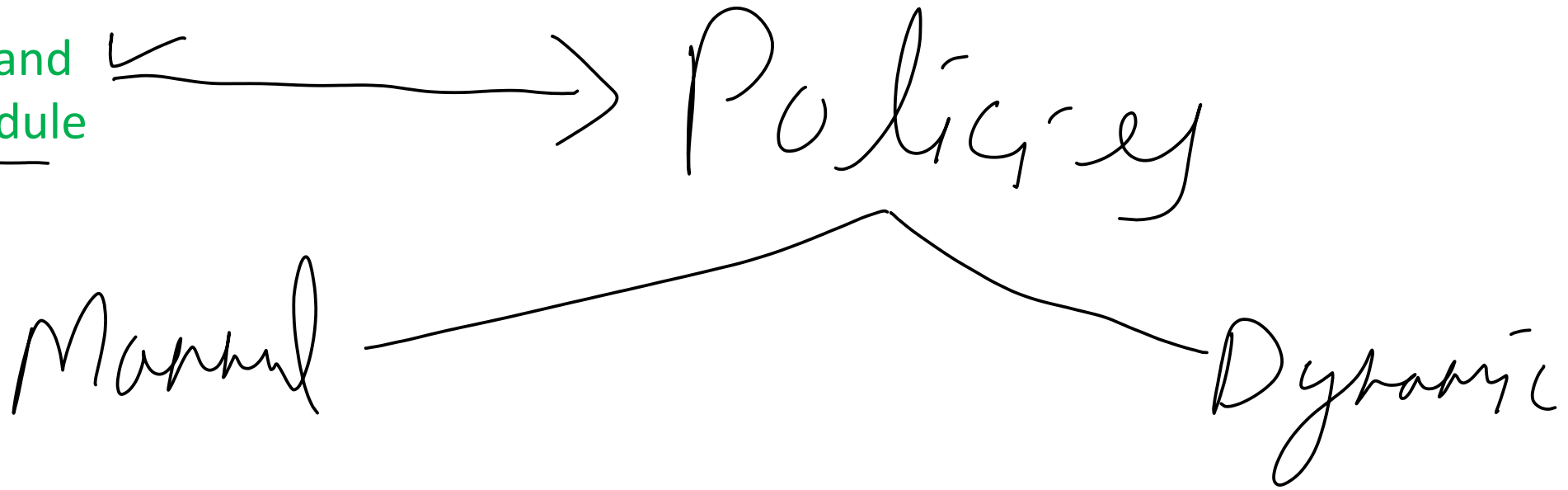
If we specify scaling polices then AS launces/terminates instances when application load increases/decreases.

Auto Scaling Introduction



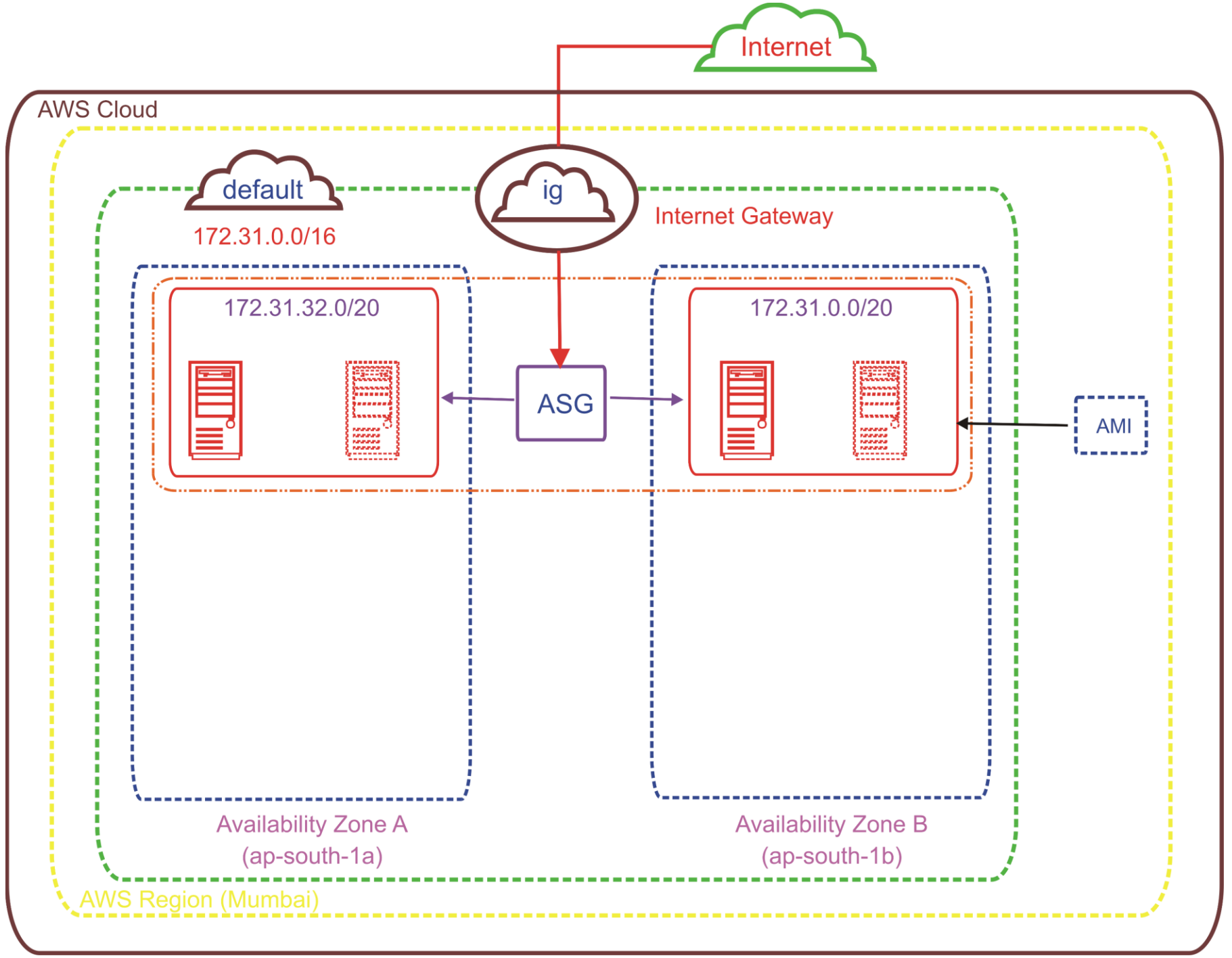
Scaling Plans

- ~~Maintain current number of instances~~
- Manual scaling
- Scale based on demand
- Scale based on schedule



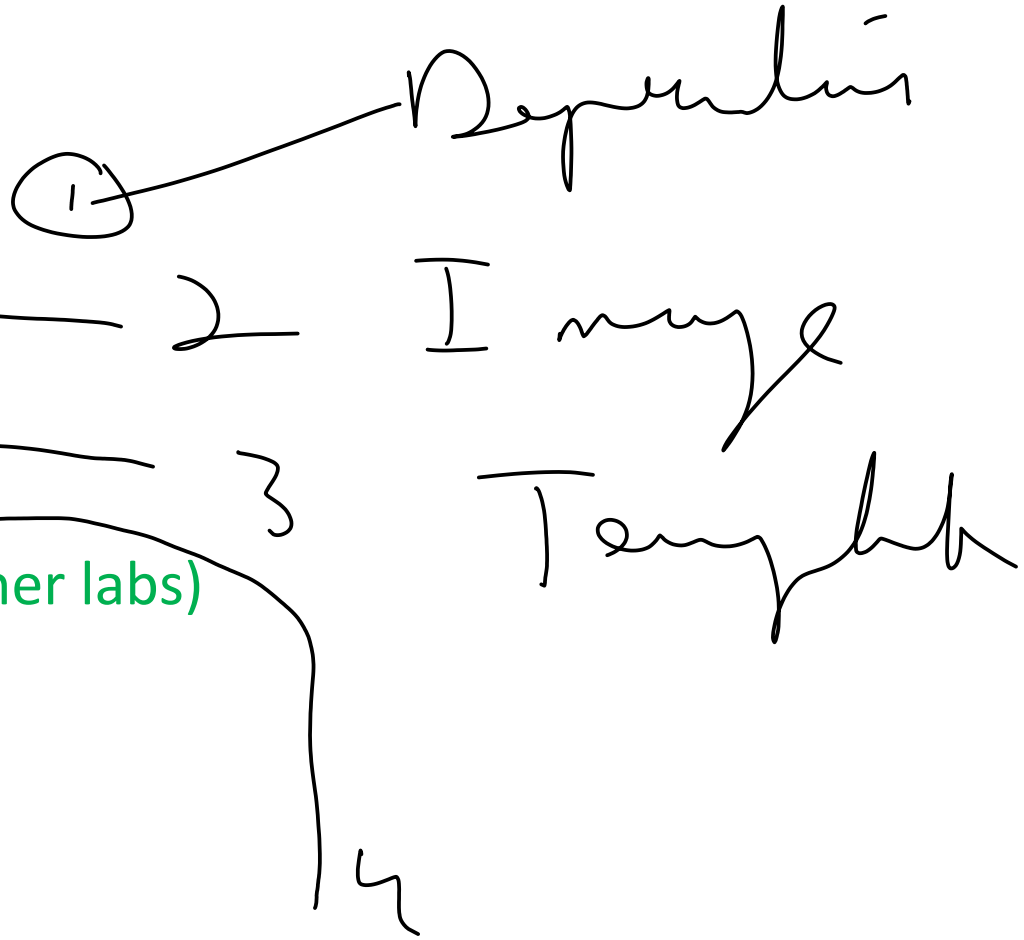
Auto Scaling Lab Setup

Default VPC



Various Steps

- Create EC2 Webserver Instance with user data
- Create Custom AMI From Instance
- Create Launch Template
- Create Auto Scaling Group
- Test Auto Scaling Group
- Delete Resources (Which are not required in other labs)



Week - 8 - a

AutoScaling :- (Theory)

Scaling → Business

↓

Scalability → Automate

↓

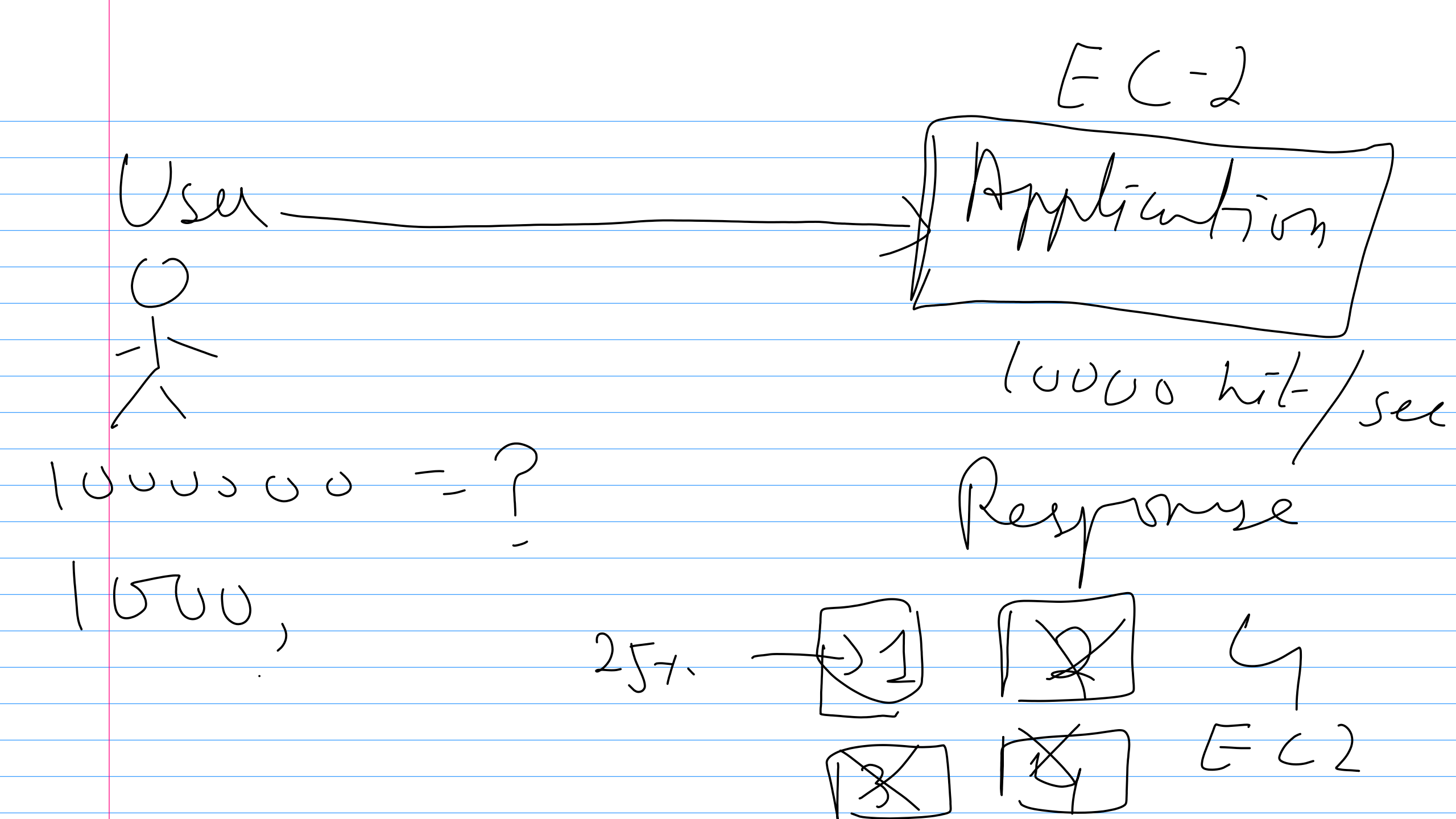
Cloud Migrate

Scalability :-

Adding up
Remove } Resource → as per requirement

Scale up
Scale out

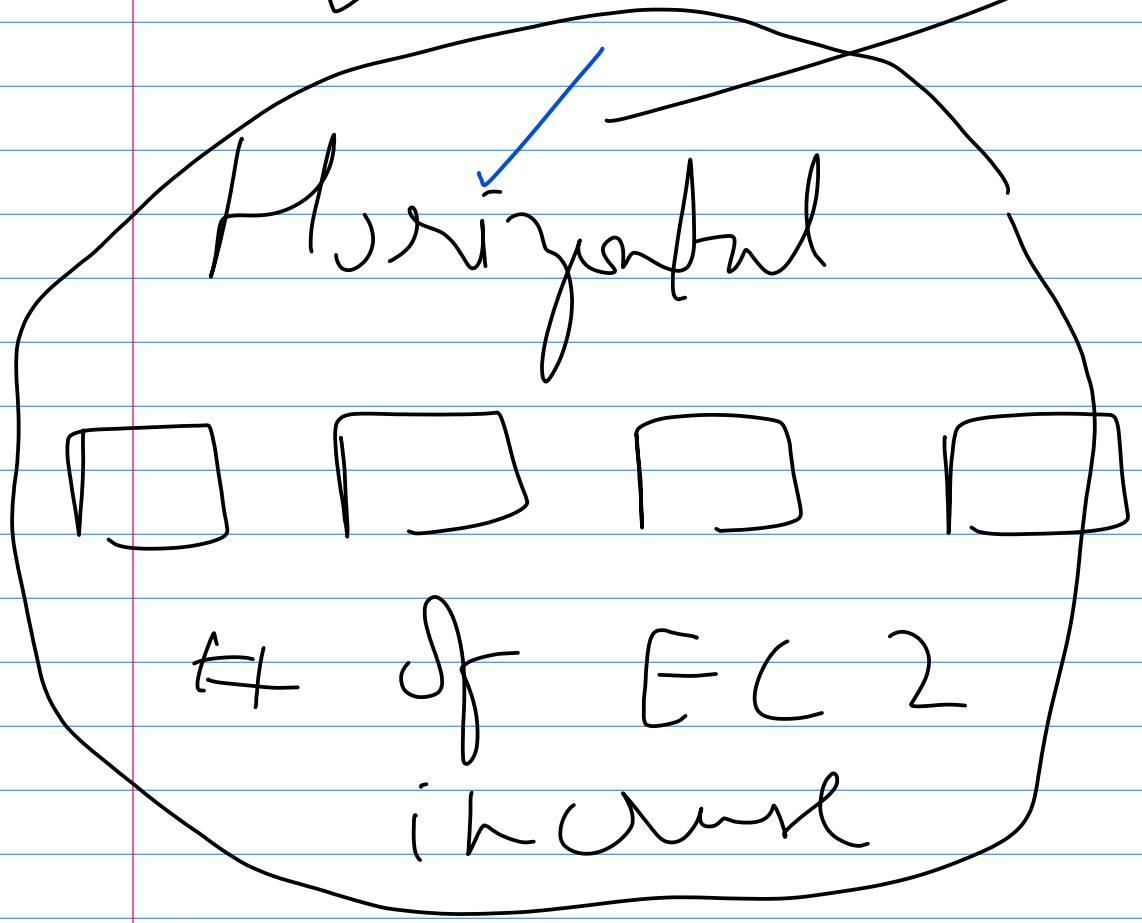
Scale Down
Scale in


EC-2
User
Application
10000 hit/sec
10000000 =?
1000,
Response
25%
1
2
3
4
EC2

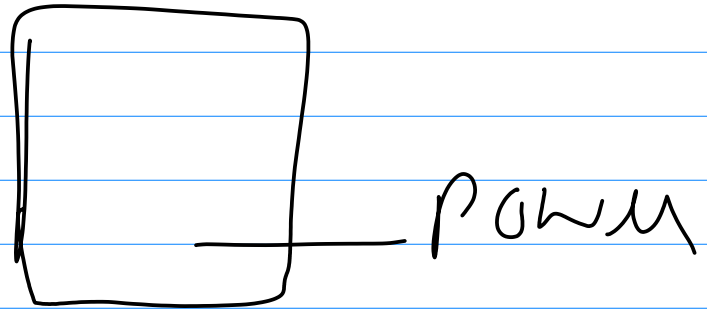
Scalability

Horizontal

Vertical



+ of EC2 instance



power

AutoScaling Components

1. Launch Configuration
   AMI ⟶ custom
2. Autoscaling Group
3. Scaling ✓ Policy ✓

AMI ⟶

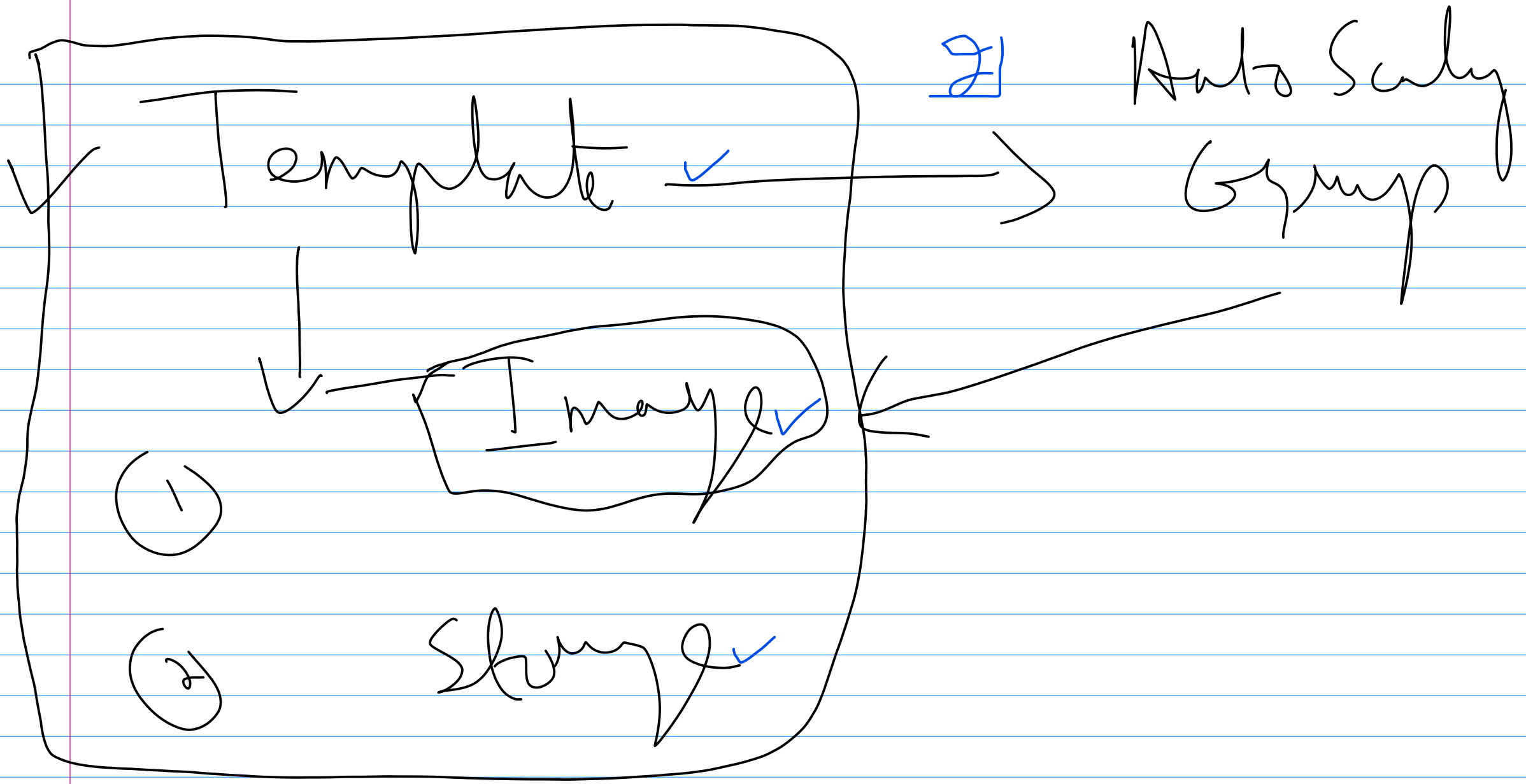
1) EC2 Instance (create)

Apache, PHP, MYSQL

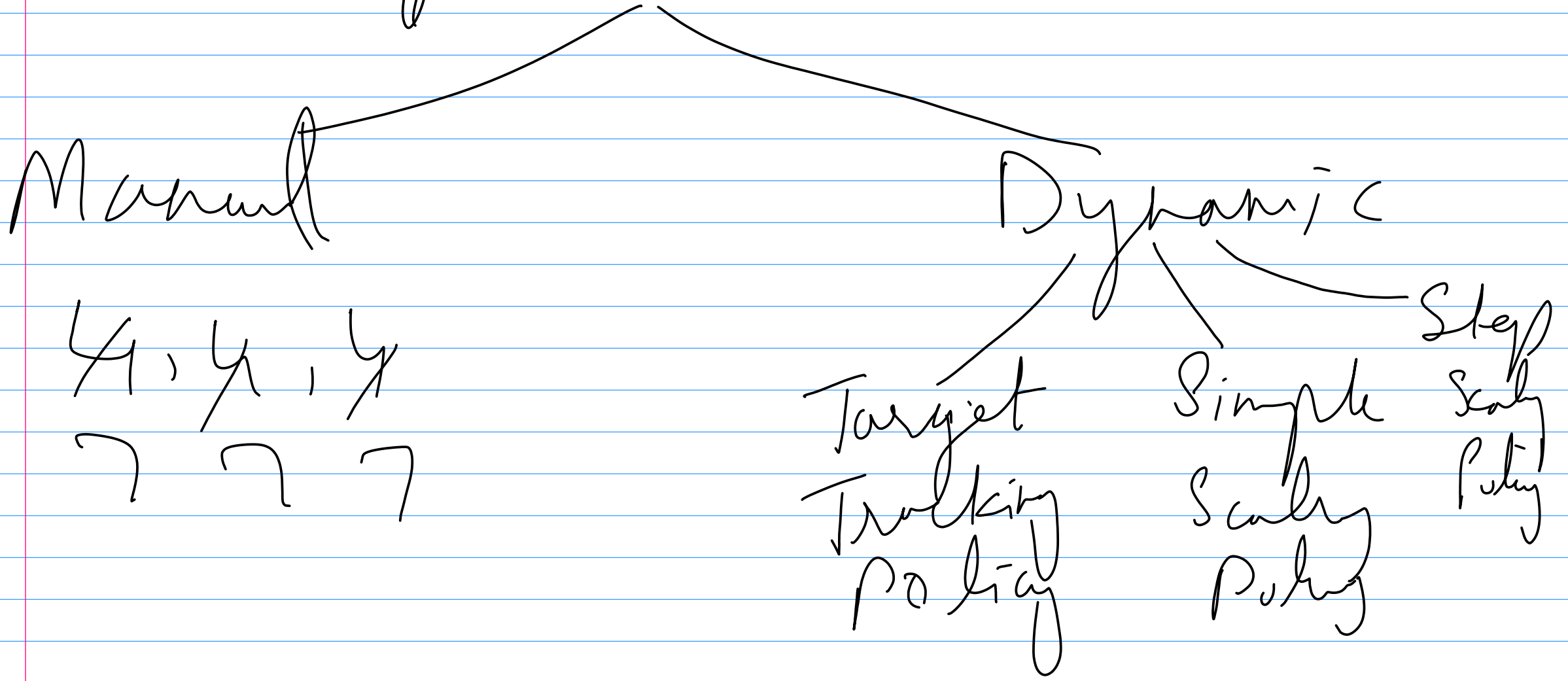
Application Deploy

↓

Image


Template
Auto Scaling Group
Image
1
2
Skunge

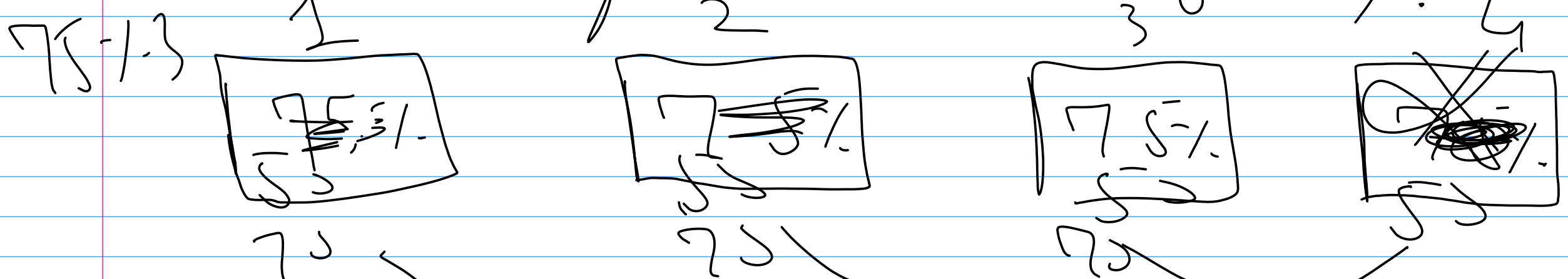
3) Scaling Policies



Dynamic Scaling

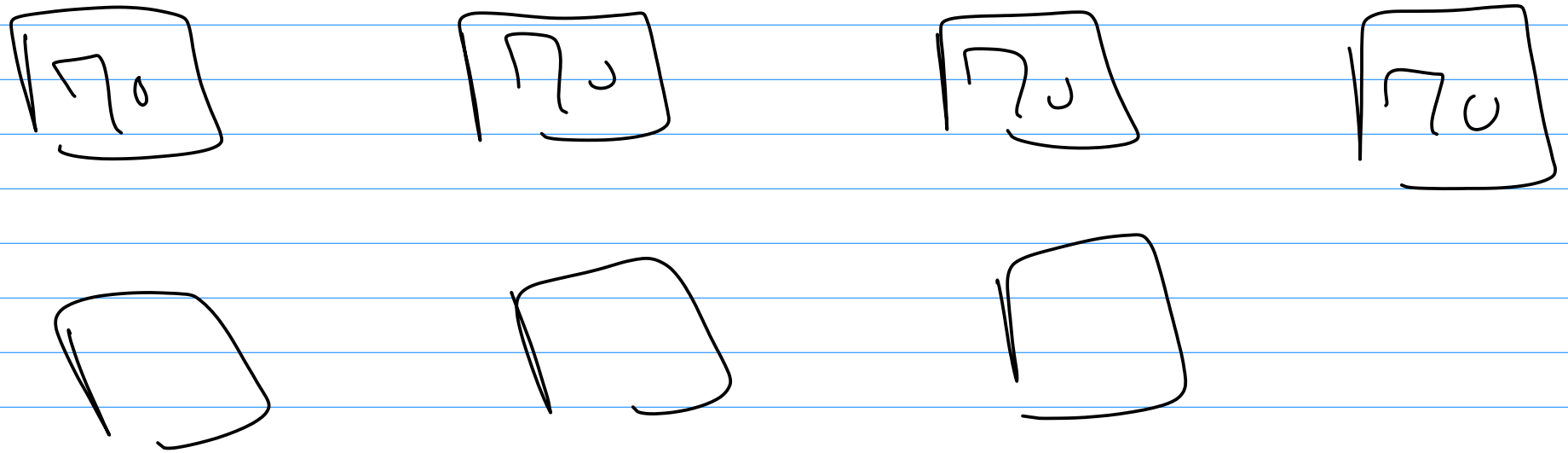
1) Target Scaling Policy

CPU Utilization → 80%



2) Simple Scaling

CPU Utilization 70%



3) Step Scaling Policy

Step 1 — 10 — 20 CPU %

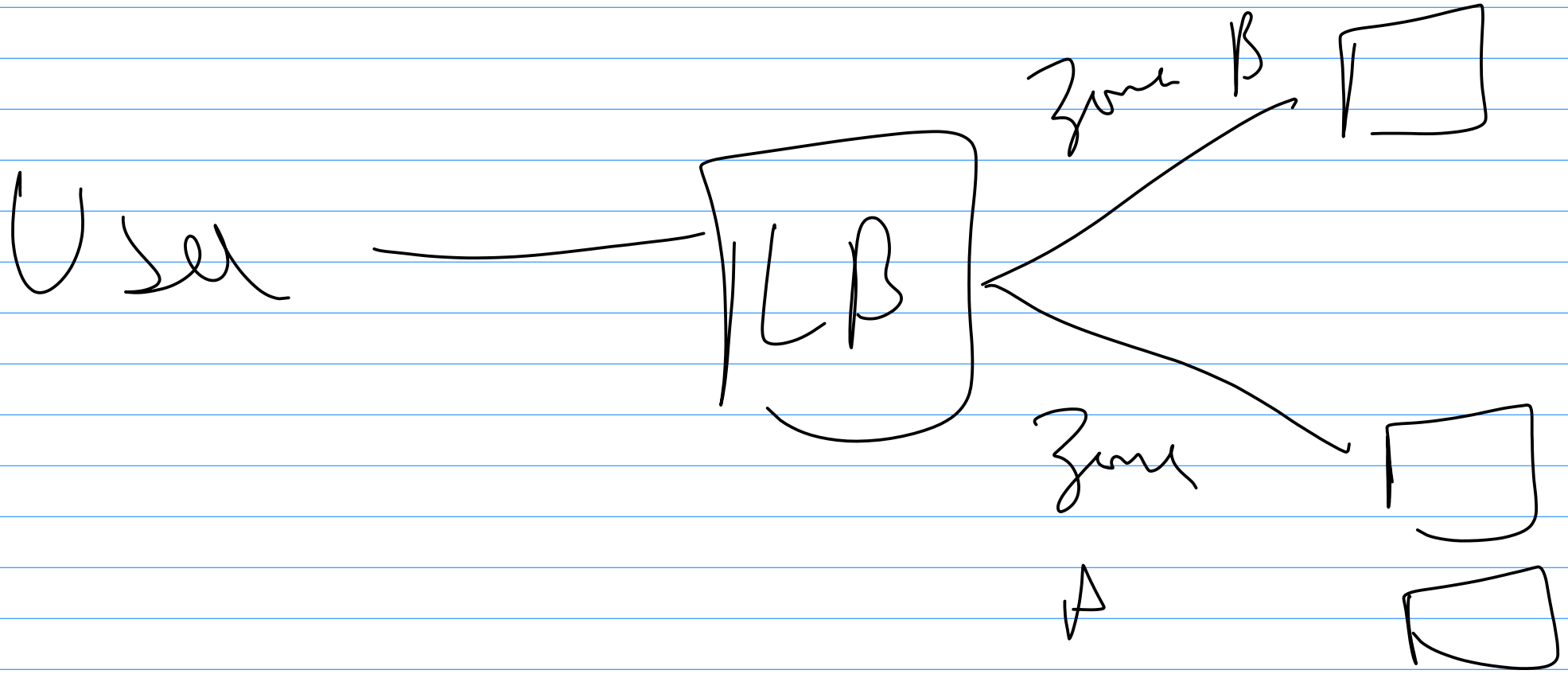
1 Instance

Step - 2 —— 30 - 40

2 Instance

5 Instance 60 — 80

Load Balancer



Warm Up ~~Time~~ Period

Cool Down Period

Warm Up 300 Sec ⟶ 5 Mins

Lunch —