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.
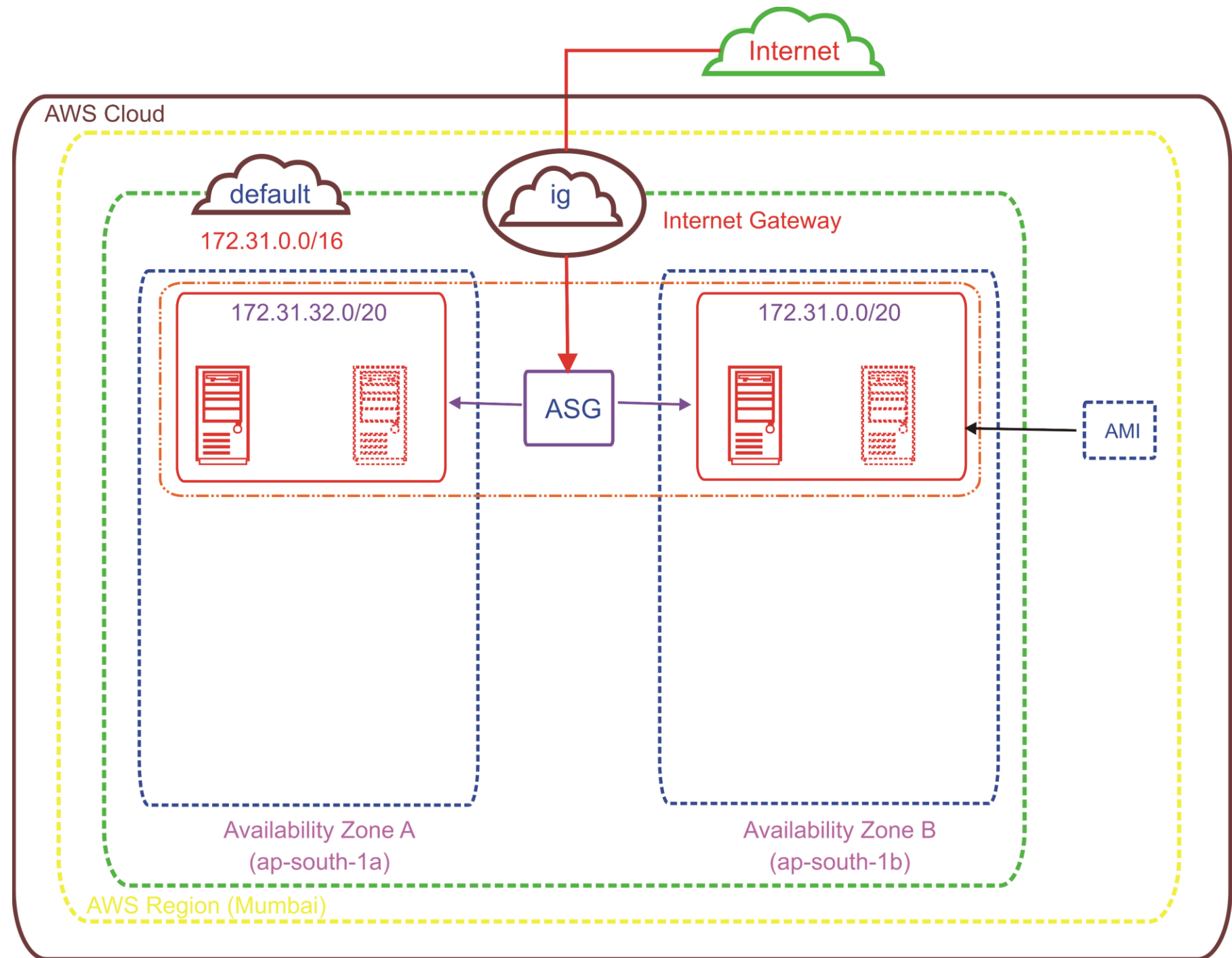
Auto Scaling Introduction
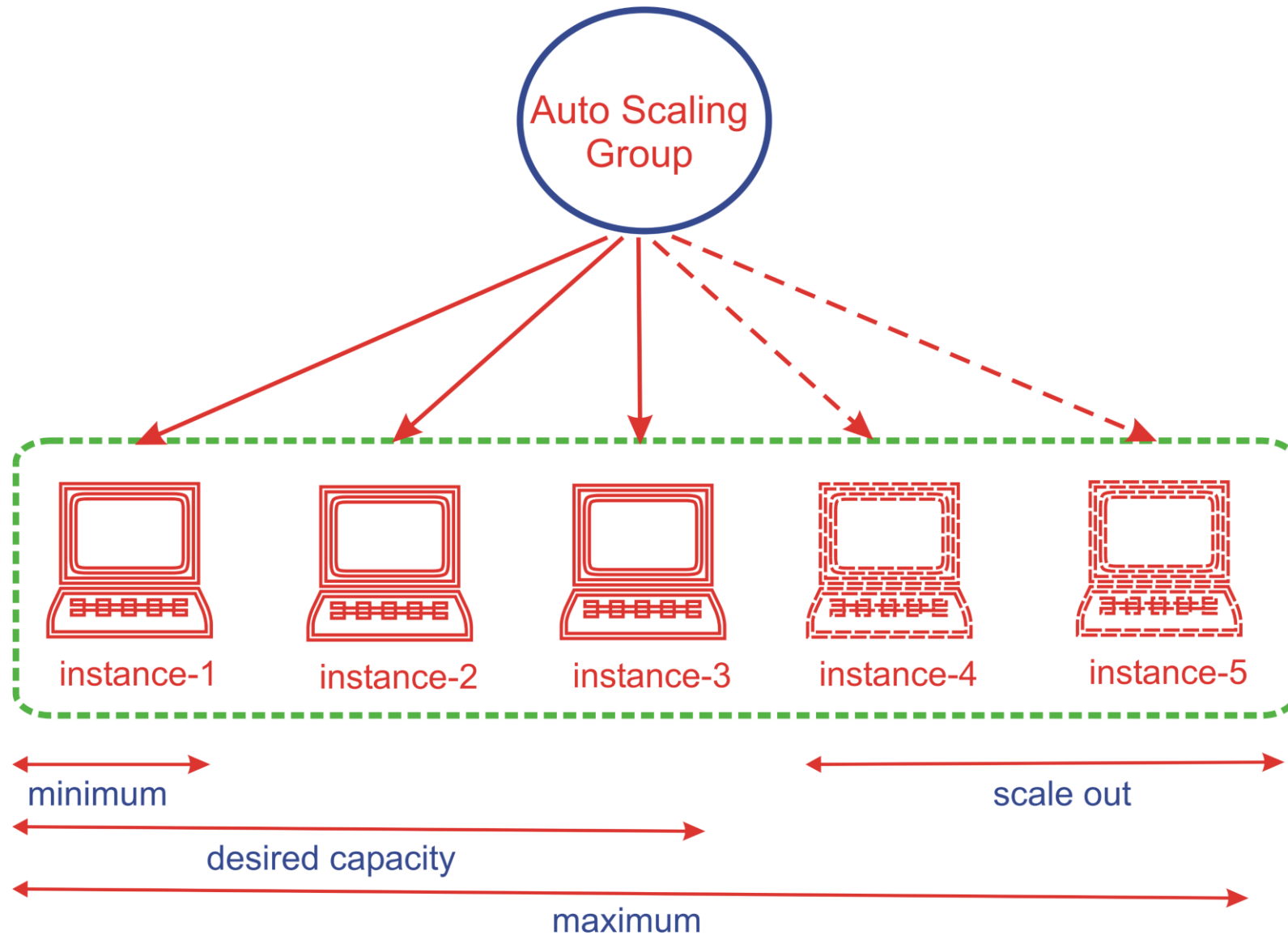
Scaling Plans

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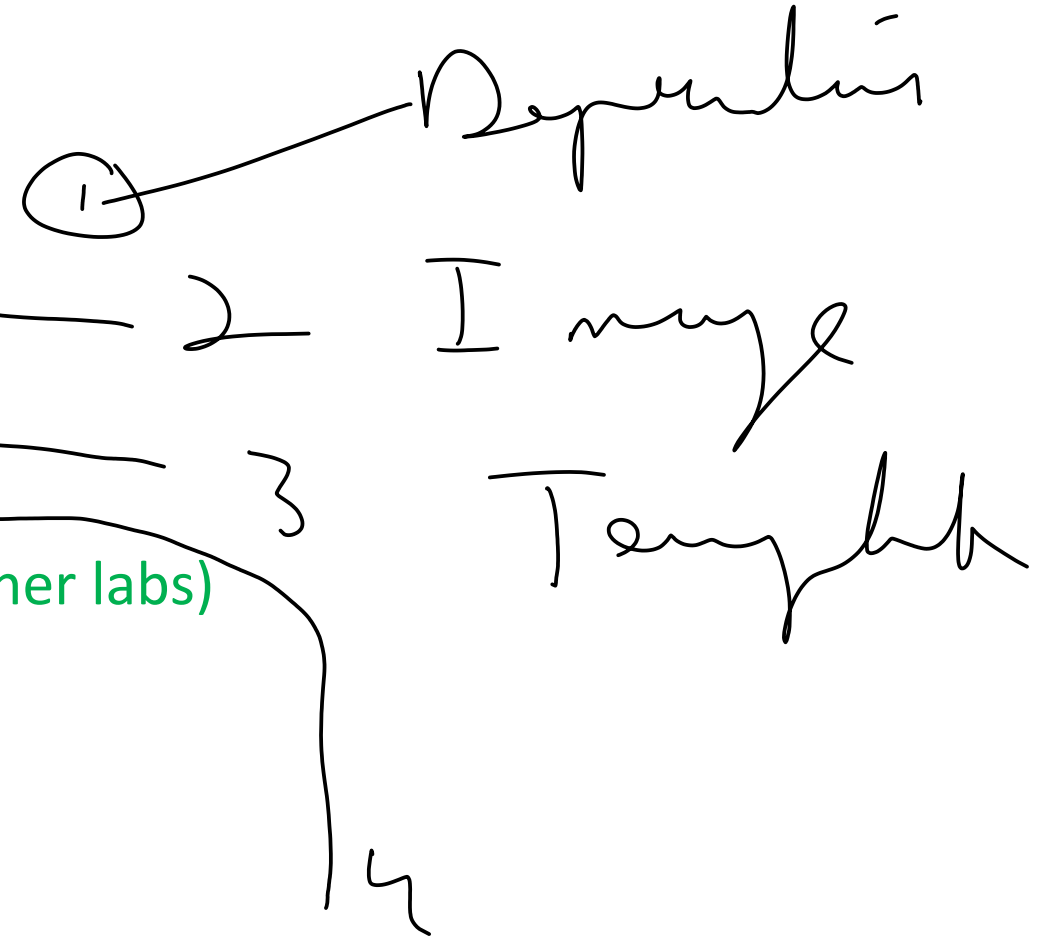- Scale based on schedule

Policies

Manual

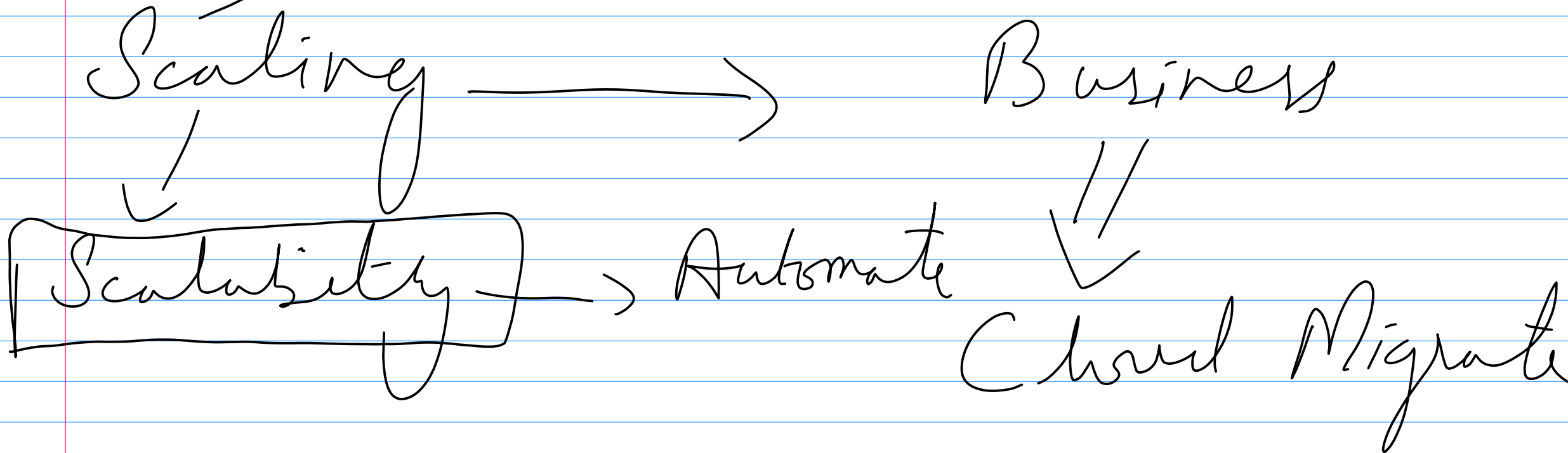Dynamic

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)

1 → Depending

2 → Image

3 → Template

4

Week - 8 - a

AutoScaling :- (Theory)

Scaling ⟶ Business

Scaling → [Scalability] → Automate

Cloud Migrate

Scalability :-

Adding up
Remove up } Resource ⟶ as per requirement

Scale up
Scale Out

Scale Down
Scale In

User

EC-2

Application
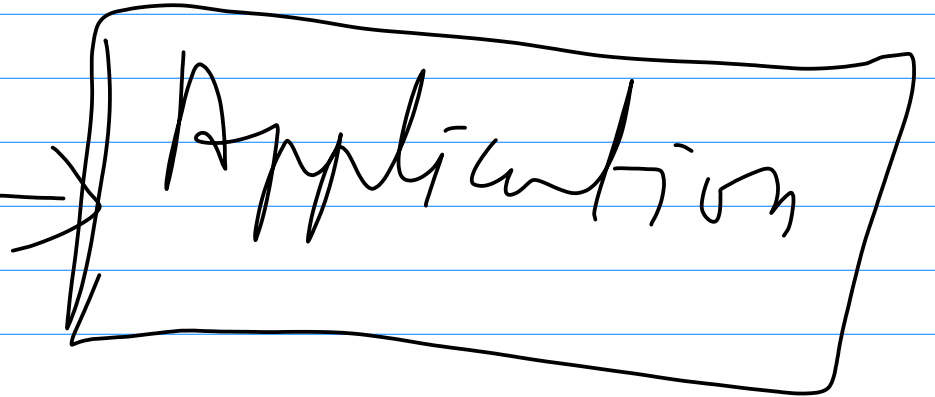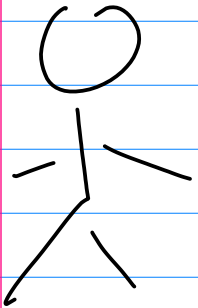
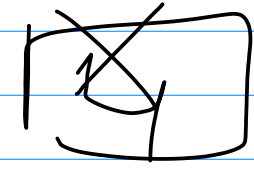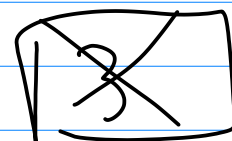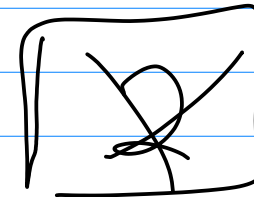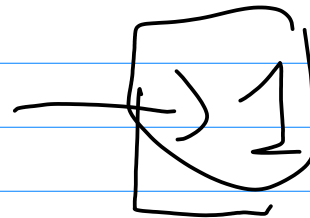10000 hit/sec

1000000 = ?

1000,

Response

25% —→ 1

4

EC2

Scalability
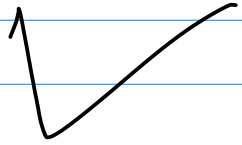
Horizontal
of EC2 in cluster
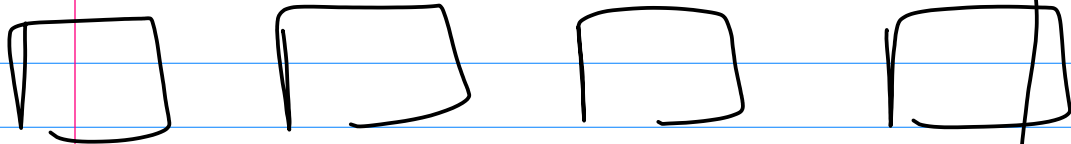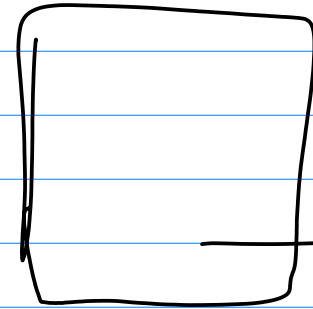
Vertical
Power

AutoScaling Components

1. Launch Configuration
   AMI ———————> Custom

2. Autoscaling Group

3. Scaling Policy

AMI $\longrightarrow$

(1) EC2 Instance (create)

Apache, PHP, MYSQL

Application   Deploy

Image

Template $\longrightarrow$ Auto Scaly Group

Template $\downarrow$

Image

① Storage

②

3) Scaling Policy
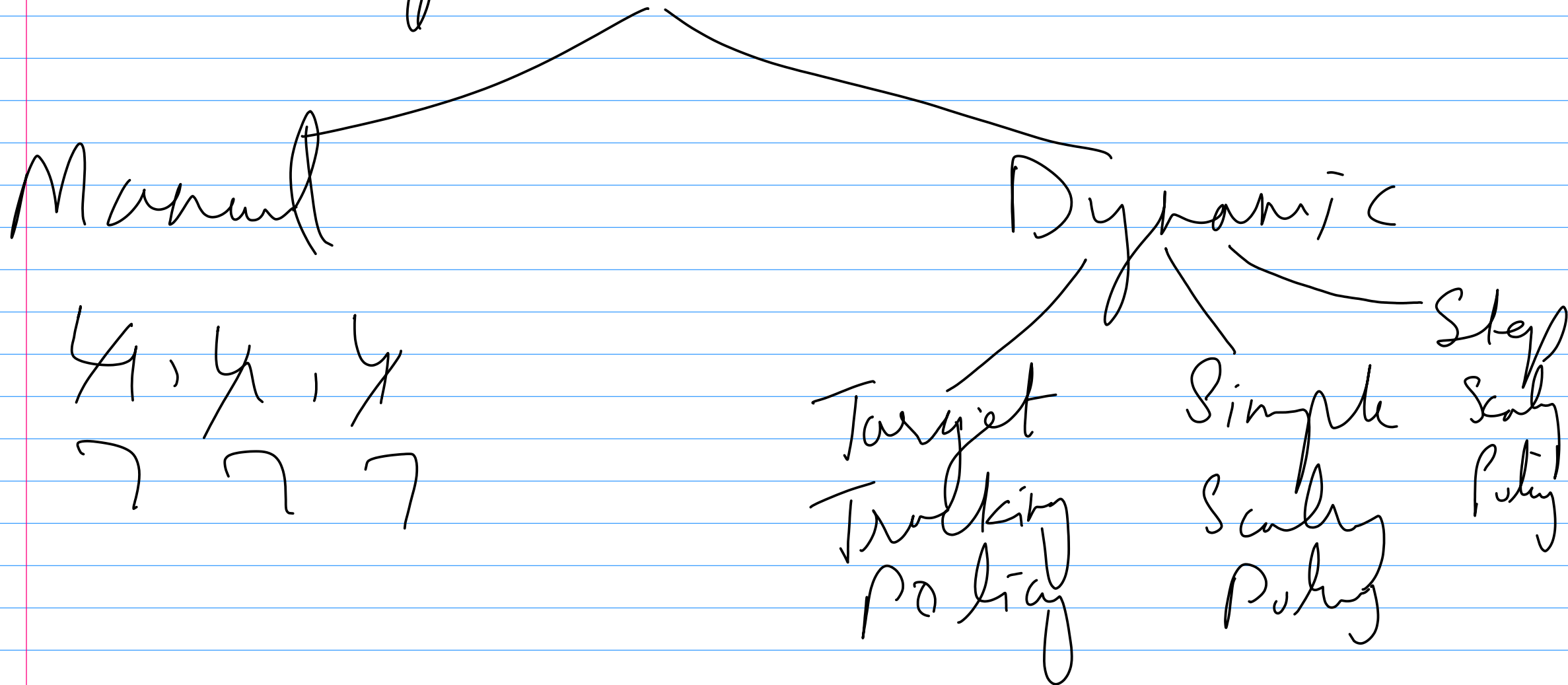
Manual

Dynamic

4, 4, 4
7 7 7
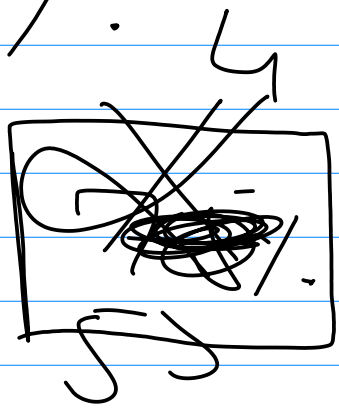
Target
Tracking
Policy
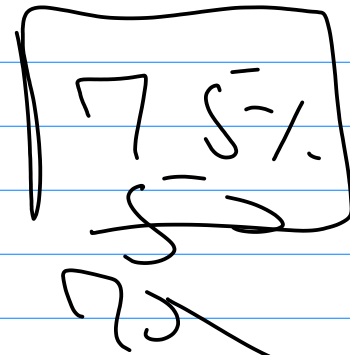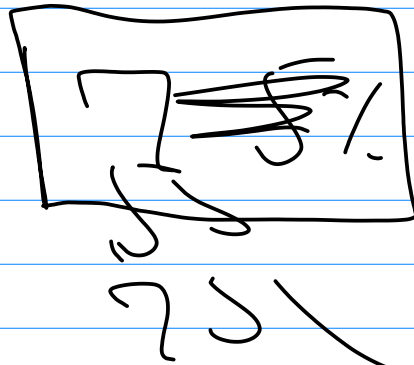
Simple
Scaling
Policy
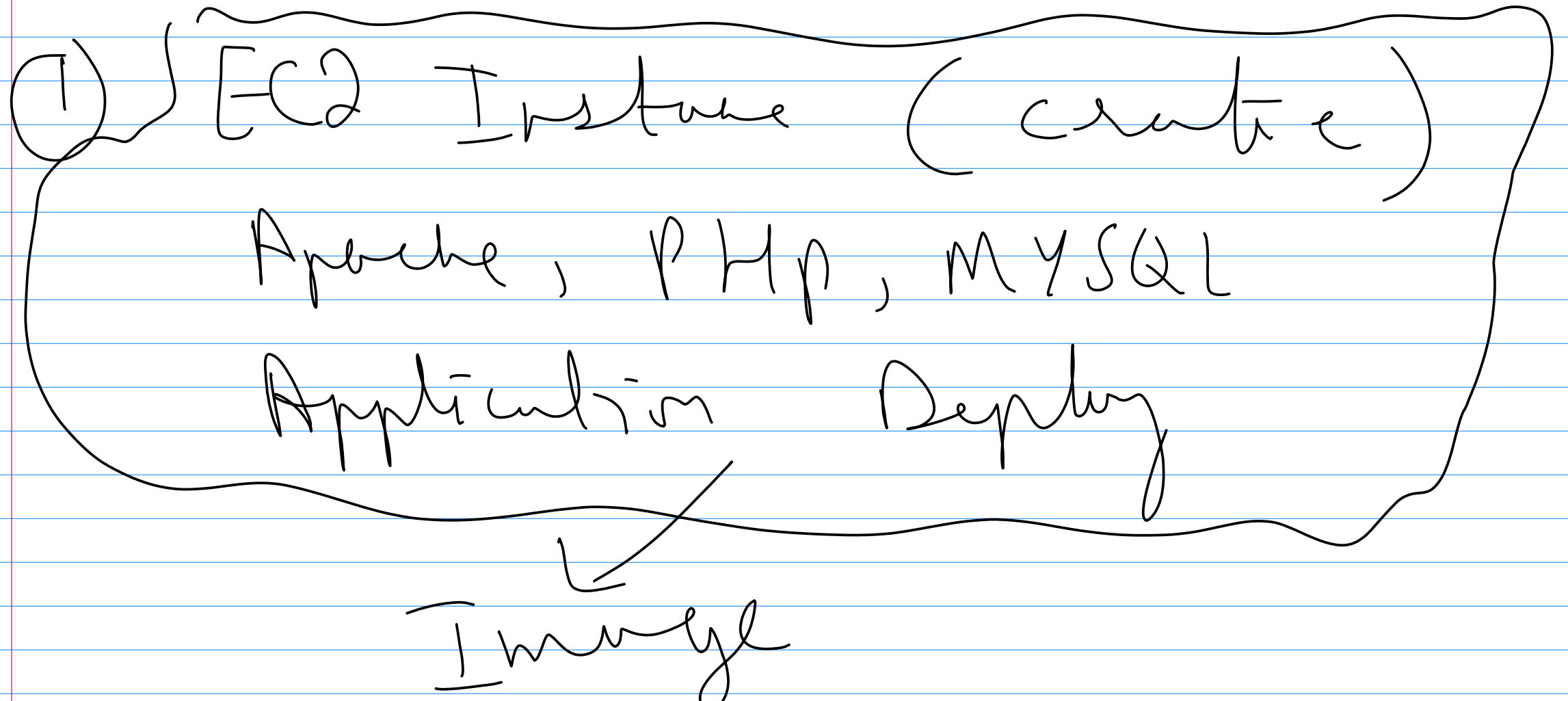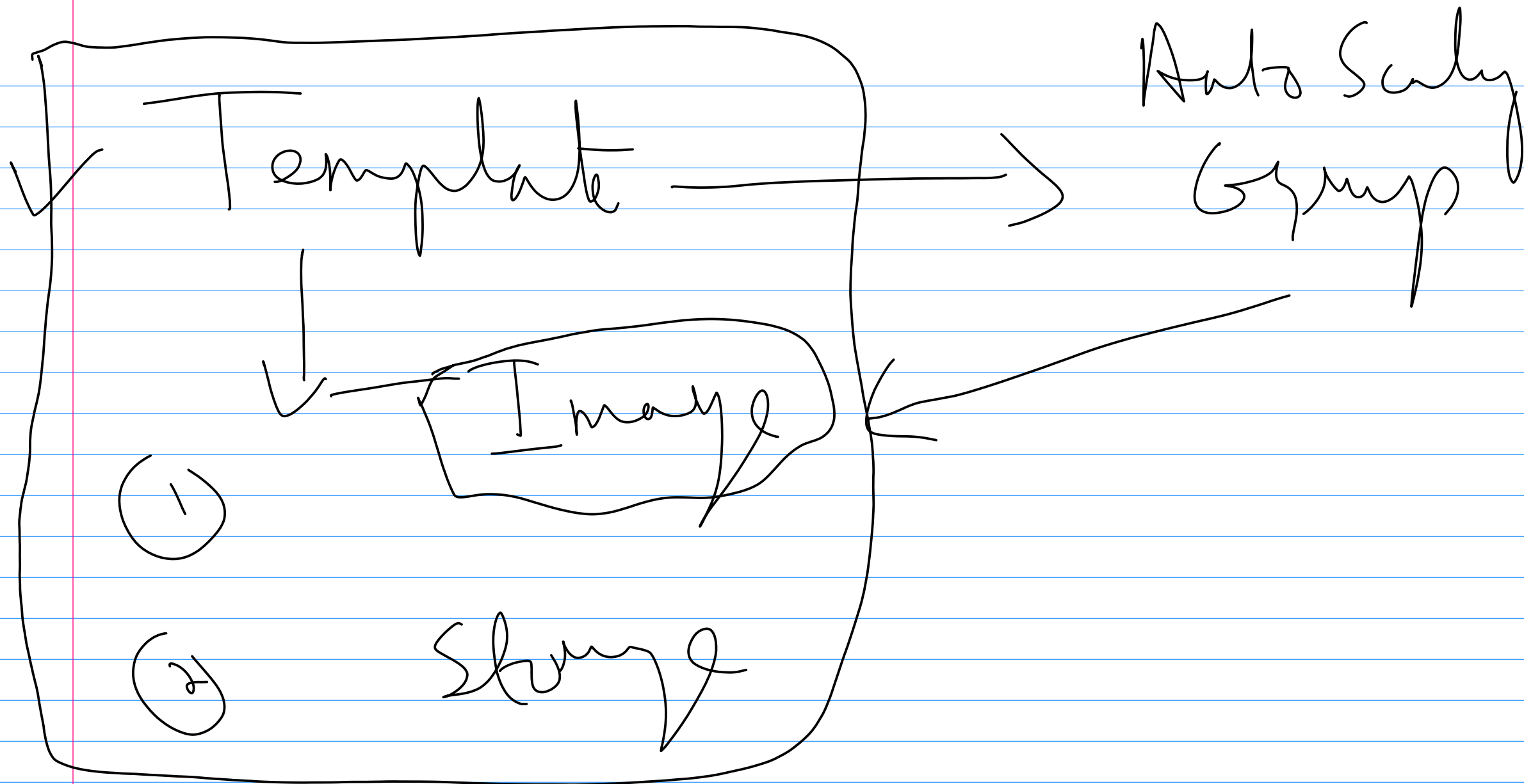
Step
Scaling
Policy

Dynamic Scaling
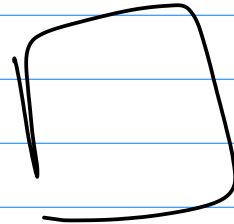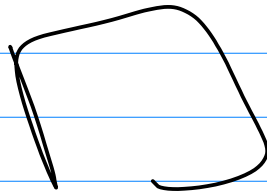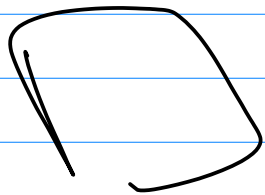
1) Target Scaling Policy

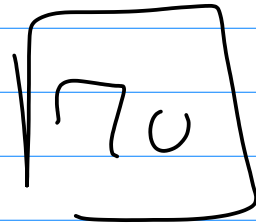CPU Utilization ⟶ 80%

75-1.3   1   2   3   80%   4

75-1.3

2) Simple Scaling

CPU Utilization 70%.

```
[70]   [70]   [70]   [70]

 ▱      ▢      ▢
```
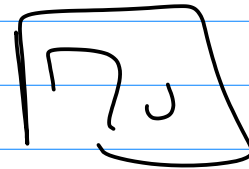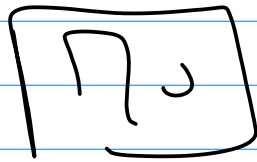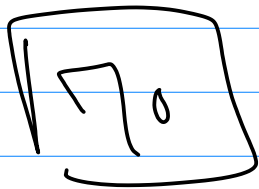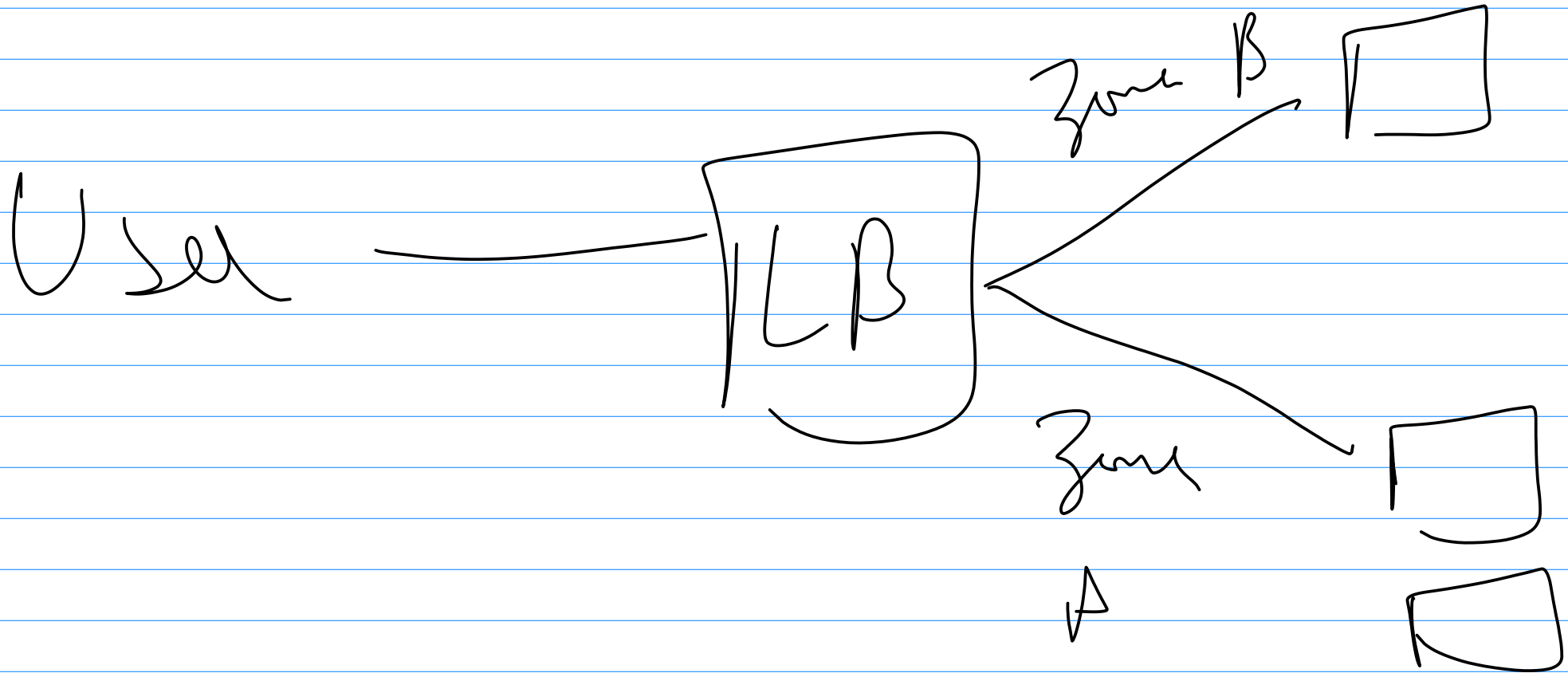
3) Step Scaling Policy

Step 1 — 10 — 20 $\frac{CPU}{\%}$

1 Instance

Step - 2 — 30 — 40

2 Instance

5 Instance   60   80

Load Balancing

User ———————— [LB]

Zone B → □

Zone A → □ □

Warm Up ~~Time~~ Period

Cool Down Period

Warm 300 Sec ⟶ 5 Minutes
Up
Lunch —— ☐