

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.

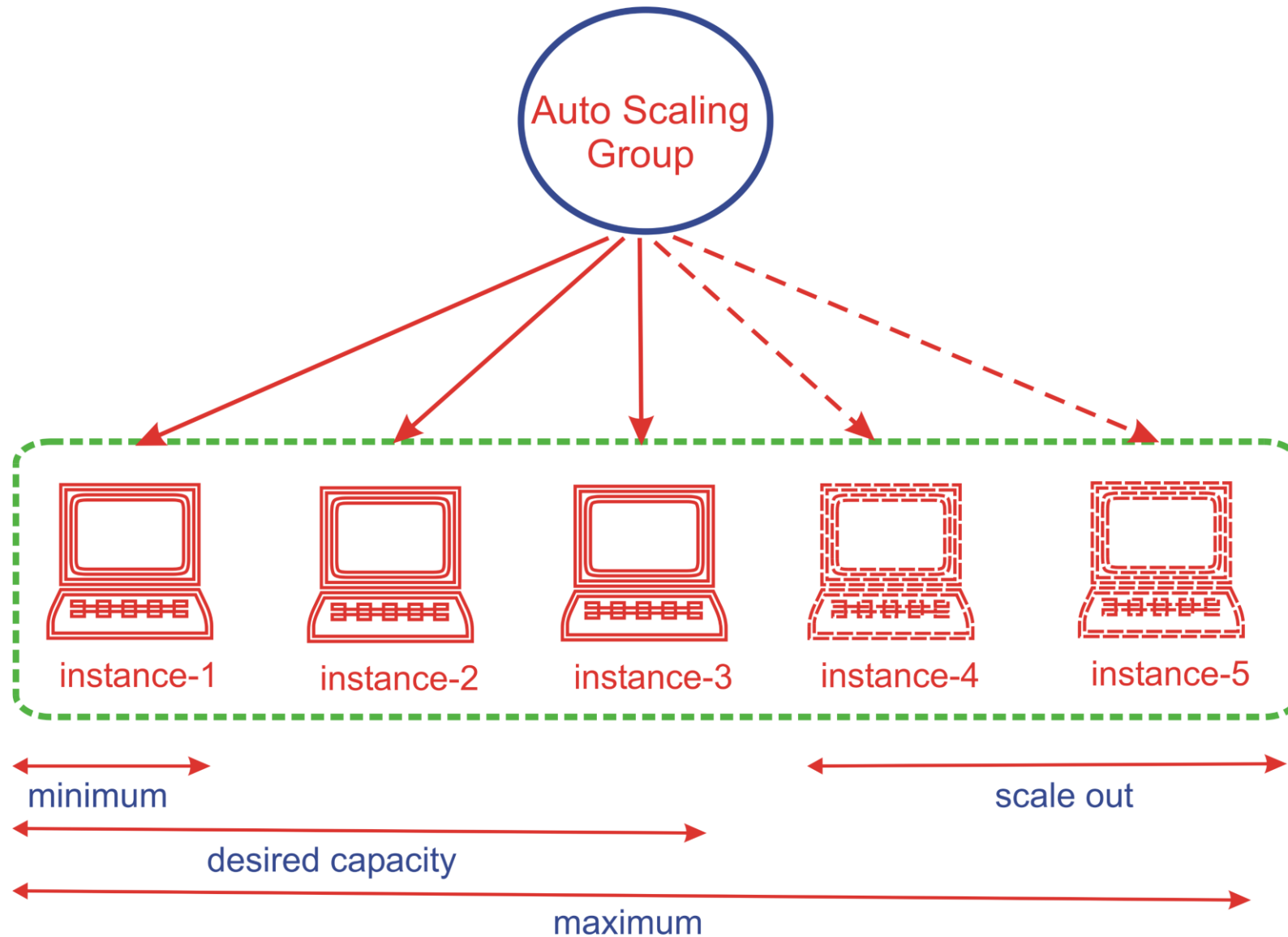
We can specify **minimum number** of instances in AS Group and AS will ensure that instances do not go below that.

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 launches/terminates instances when application load increases/decreases.

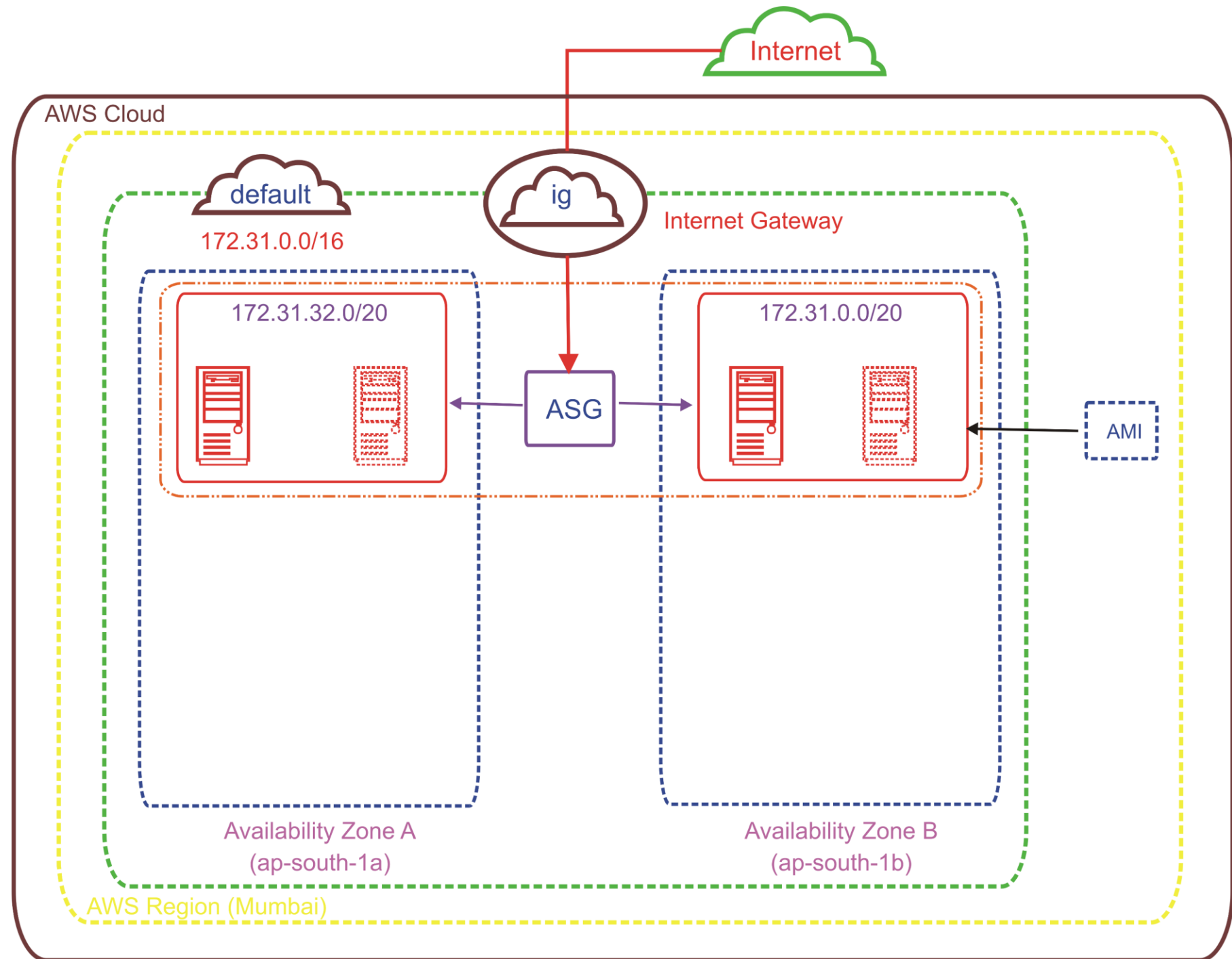
Auto Scaling Introduction



Scaling Plans

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

Auto Scaling Lab Setup



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)