Amazon Auto Scaling
What is Amazon’s AutoScaling?

• A group of instances that runs a particular application which can:
  – Add or remove instances in response to changes in demand.

• Amazon lets you use AutoScaling to:
  – Keep a set number of instances running and automatically launch instances to replace failing instances (as determined by a health check).
  – Scale instances based on a schedule
  – Dynamically scale instances based on a condition
Instances and Launch Configurations

Launch Configuration + AMI which auto-configures services on startup
Define an AutoScaling Group

Instance
Instance

Auto Scaling Group

Min, Max and Desired Number of Instances
How does it work?
What is required?

• Group of EC2 Instances (AutoScaling Group)
  – Group Parameters
    • A minimum, maximum and desired number of instances.
  – A Launch Configuration for the instances.
    • AMI, Security Groups, Keypairs etc.
  – Scaling Policies
    • Actions to perform on the auto scaling group (add or reduce the number of instances).
    • These are triggerable actions, either manually or through CloudWatch.
  – Triggers
    • CloudWatch Alarm to trigger a certain policy.
How can I access AutoScaling?

- AutoScaling CLI Tools.
- AutoScaling API
  - Included in the Amazon AWS SDK.

aws.amazon.com/documentation/documentation/autoscaling/