Auto Scaling

- Setup processing instances dynamically
    - Respoding to workload
    - Provisioning new when demand is spiky , Terminating old when demand is low brigs up savings

Auto scaling schemes

- Maintain current instance level
    - Steady state workload
    - Automatically performs health check
    - Specify minimum number of instances
- Manual
    - Specify Max / Min and desired
    - Suitable for Infrequent event - launching a new game -reg website
- Scheduled
    - Monthly, quarterly load
    - Scheduling decisions made as a function of date/time
    - Suitable for Recurring workload
- Dynamic
    - Define params that control scheduling decision
    - N/w bandwidth or cpu usage monitored by CW

Autoscaling Components

- Launch configuration
    - components - name, ami , security group, keypair , (bid price)
    - 100 autoscale configs per account by default
    - but only 20 instances ec2 of one configuration
- Autoscaling group
    - It has configuraration options
        - AG name, Min, max ,desired #, az (can be multiple), load balancer
    - On-demand (default) or spot instance
    - Spacify spot instance price in config options
    - Rest works same.
    - Create new configuration with new bid price and attach to your scaling price
- Scaling policy
    - Associate cloud watch alarm
    - When triggered it executes policy and then launches or terminates instances
    - Scale out quickly but scale in slowly
    - Scale in cant be done quickly
        - Workload burst can cause quick scale out
        - When burst is NOT sustained scale in should not happen quickly
    - Cooldown

- After scale poicy action is taken, this is a "pause" that is added before next scaling operation can begin.