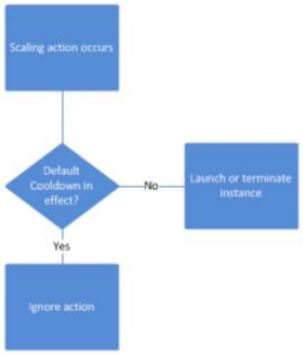
Auto Scaling Groups – Scaling Policies

- Target Tracking Scaling
  - Most simple and easy to set-up
  - Example: I want the average ASG CPU to stay at around 40%
- Simple / Step Scaling
  - When a CloudWatch alarm is triggered (example CPU > 70%), then add 2 units
  - When a CloudWatch alarm is triggered (example CPU < 30%), then remove 1
- Scheduled Actions
  - Anticipate a scaling based on known usage patterns
  - Example: increase the min capacity to 10 at 5 pm on Fridays

Auto Scaling Groups - Scaling Cooldowns

- The cooldown period helps to ensure that your Auto Scaling group doesn't launch or terminate additional instances before the previous scaling activity takes effect.
- In addition to default cooldown for Auto Scaling group, we can create cooldowns that apply to a specific simple scaling policy
- A scaling-specific cooldown period overrides the default cooldown period.
- One common use for scaling-specific cooldowns is with a scale-in policy—a policy that terminates instances based on a specific criteria or metric. Because this policy terminates instances, Amazon EC2 Auto Scaling needs less time to determine whether to terminate additional instances.
- If the default cooldown period of 300 seconds is too long—you can reduce costs by applying a scaling-specific cooldown period of 180 seconds to the scale-in policy.
- If your application is scaling up and down multiple times each hour, modify the Auto Scaling Groups cool-down timers and the CloudWatch Alarm Period that triggers the scale in



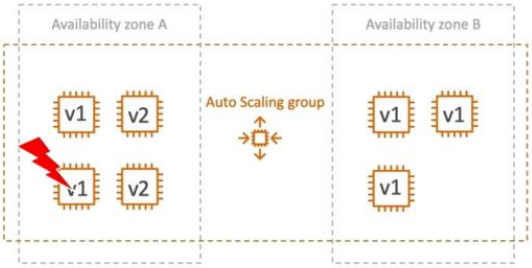
https://docs.aws.amazon.com/autoscaling/ec2/userguide/Cooldown.html

ASG for Solutions Architects

- ASG Default Termination Policy (simplified version):

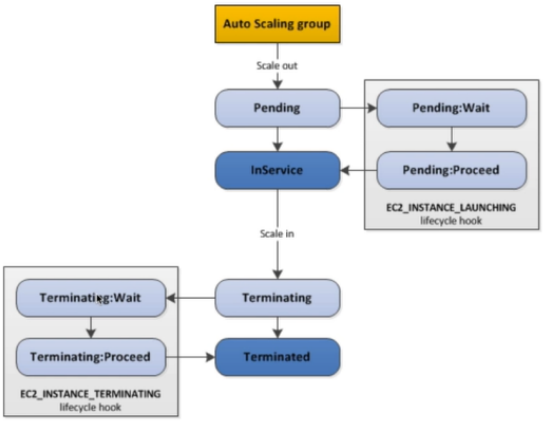
1. Find the AZ which has the most number of instances
2. If there are multiple instances in the AZ to choose from, delete the one with the oldest launch configuration

- ASG tries the balance the number of instances across AZ by default



ASG for Solutions Architects Lifecycle Hooks

- By default as soon as an instance is launched in an ASG it's in service.
- You have the ability to perform extra steps before the instance goes in service (Pending state)
- You have the ability to perform some actions before the instance is terminated (Terminating state)



https://docs.aws.amazon.com/autoscaling/ec2/userguide/lifecycle-hooks.html

ASG for Solutions Architect
Launch Template vs Launch Configuration

- Both:
  - ID of the Amazon Machine Image (AMI), the instance type, a key pair, security groups, and the other parameters that you use to launch EC2 instances (tags, EC2 user-data...)

- Launch Configuration (legacy):
  - Must be re-created every time

- Launch Template (newer):
  - Can have multiple versions
  - Create parameters subsets (partial configuration for re-use and inheritance)
  - Provision using both On-Demand and Spot instances (or a mix)
  - Can use T2 unlimited burst feature
  - Recommended by AWS going forward