Cloud Computing Architecture

Scaling Policies

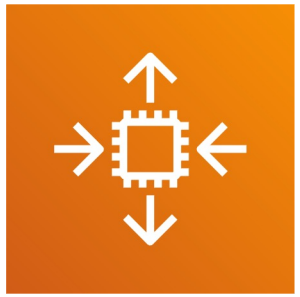
SWINBURNE UNIVERSITY OF TECHNOLOGY

Scaling Policies

This presentation:

- Auto Scaling Revisited
- How Does Auto Scaling Work?
- Auto Scaling Steps
- Auto Scaling Considerations

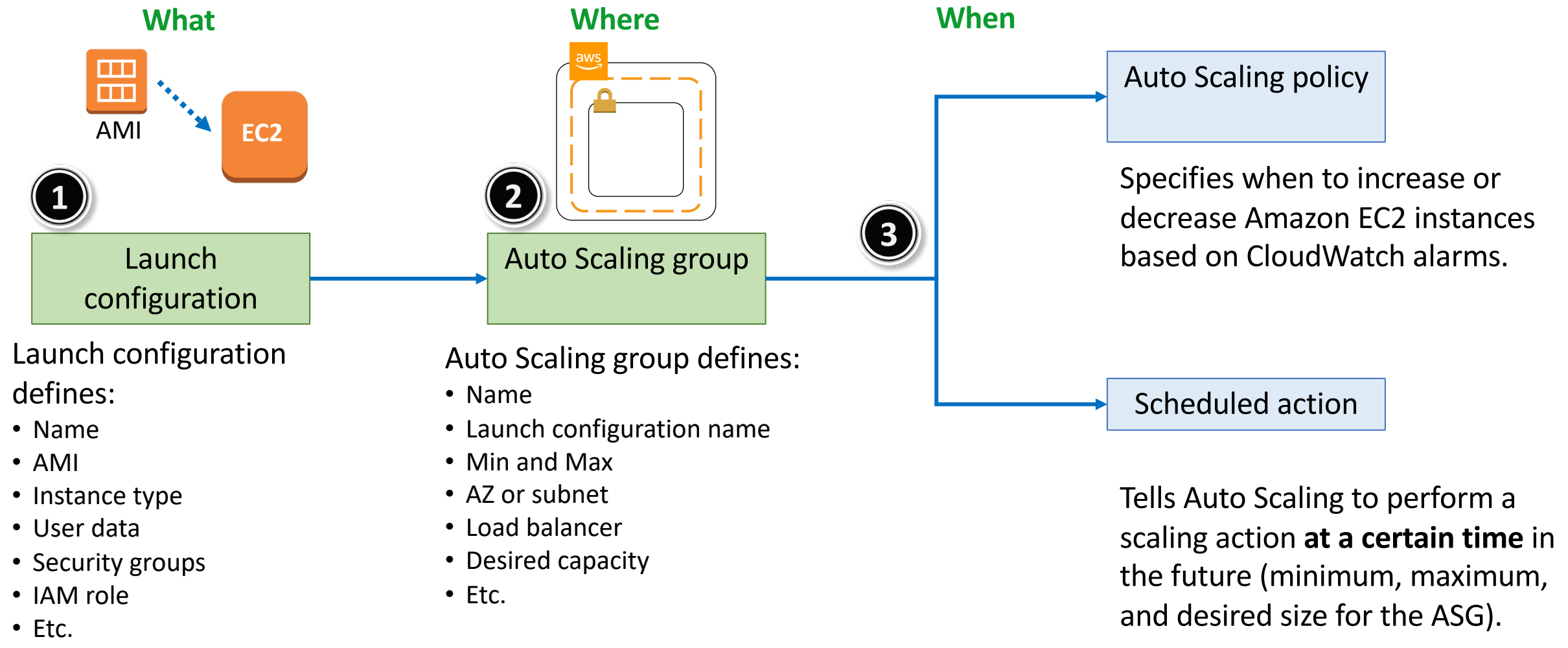
Amazon EC2 Auto Scaling (revisited)



Amazon EC2
Auto Scaling

- Launches or terminates instances based on specified conditions
- Automatically registers new instances with load balancers when specified
- Can launch across Availability Zones

How Does Auto Scaling Work? (revisited)



Launch configuration defines:

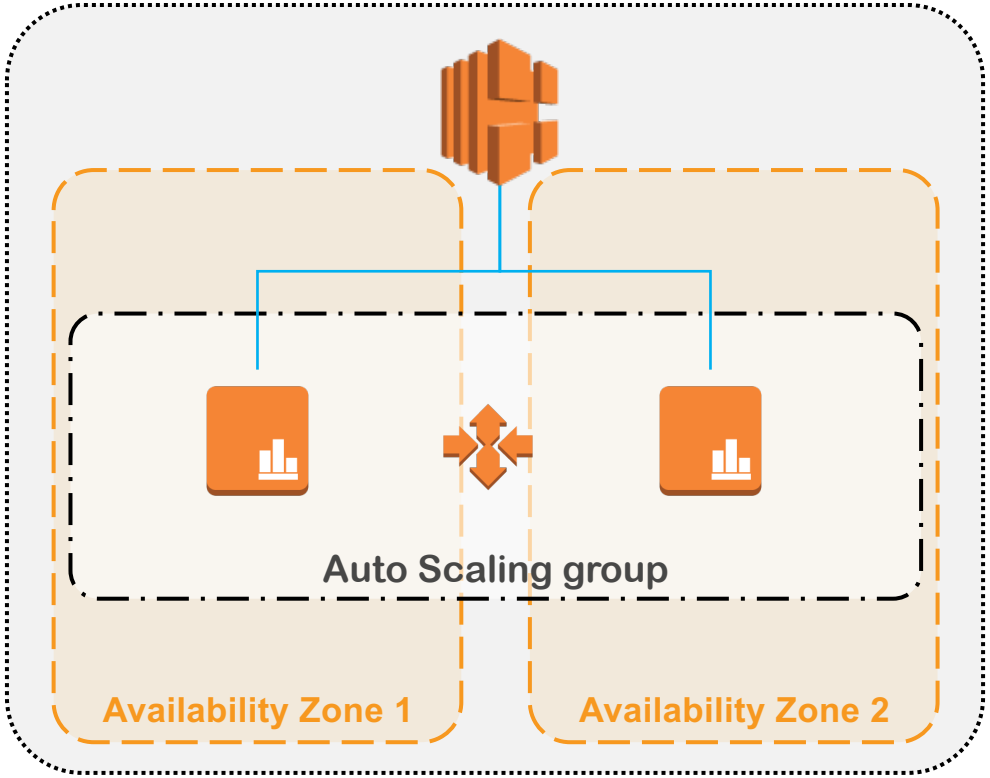
- Name
- AMI
- Instance type
- User data
- Security groups
- IAM role
- Etc.

Auto Scaling group defines:

- Name
- Launch configuration name
- Min and Max
- AZ or subnet
- Load balancer
- Desired capacity
- Etc.

Auto Scaling (revisited)

- Auto Scaling group defines:
  - Desired capacity
  - Minimum capacity
  - Maximum capacity
- What would be a good **minimum** capacity to set it to?
- What would be a good **maximum** capacity to set it to?



Minimum = two instances (# of AZs)

Desired capacity = two instances (Min.)

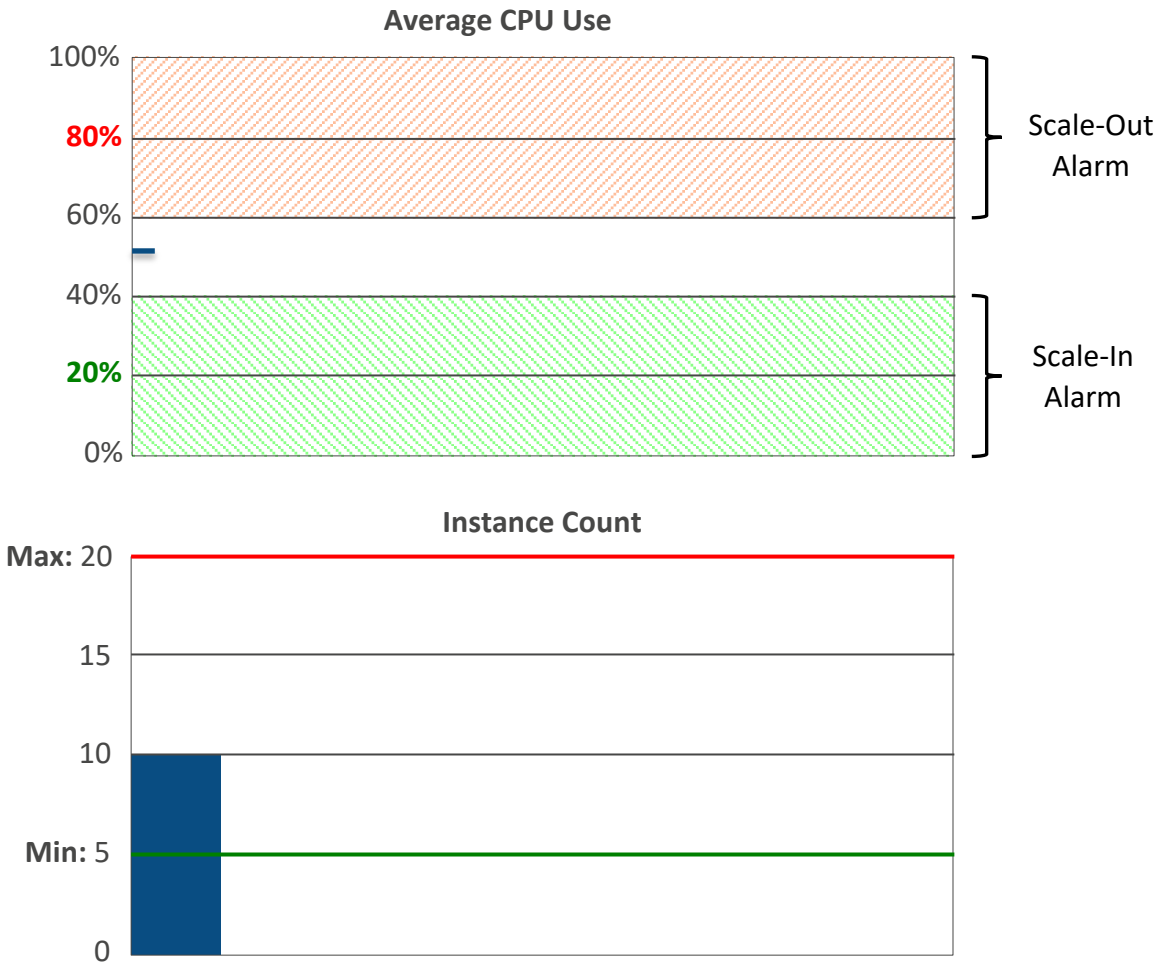
Auto Scaling Steps

Step Adjustments:

- **Add 2** instances when average CPU is **80-100%**
- **Add 1** instance when average CPU is **60-80%**
- **Remove 1** instance when average CPU is **20-40%**
- **Remove 2** instances when average CPU is **0-20%**

Limits:

- Minimum: 5 instances
- Maximum: 20 instances



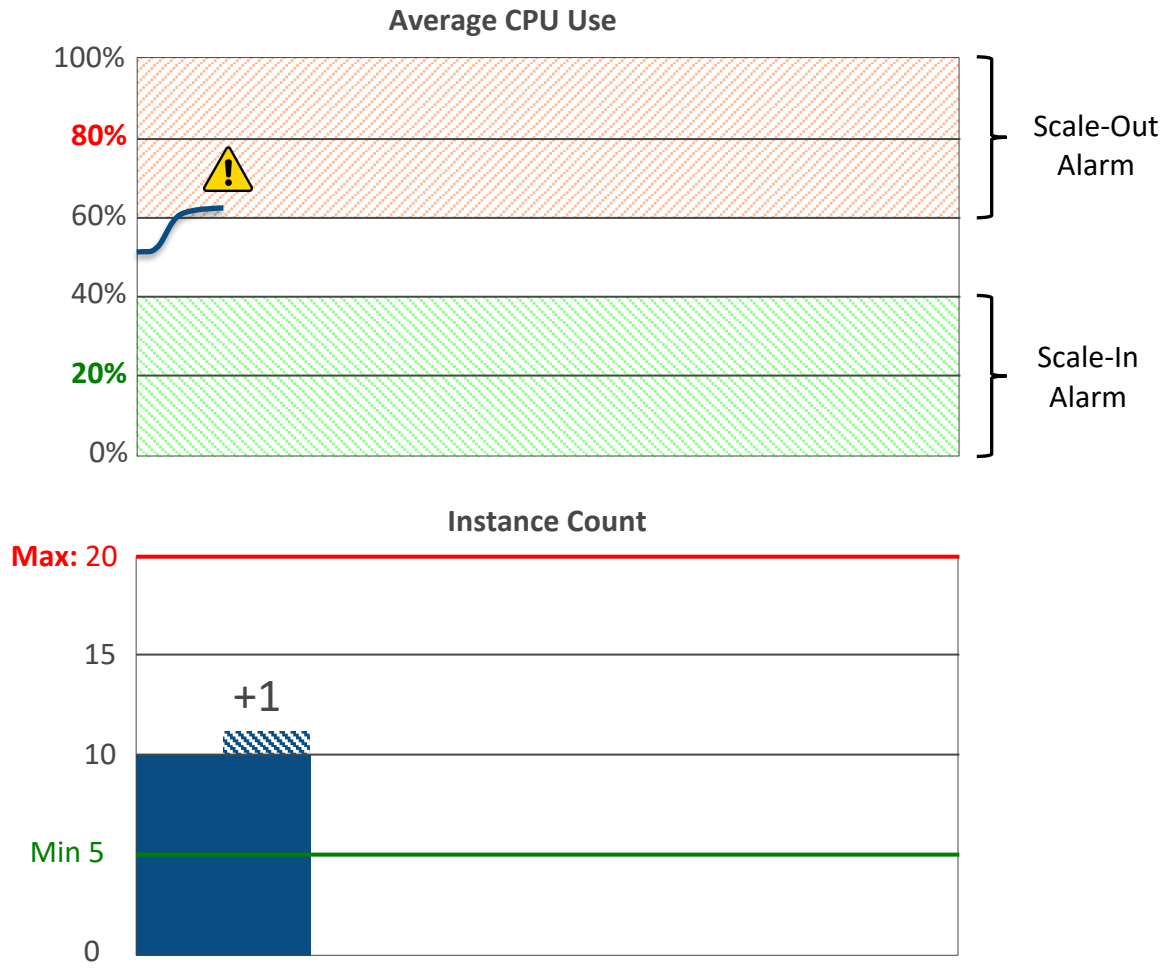
Auto Scaling Steps

As usage increases:

- CPU use goes up.

When CPU use is 60-80%:

- Scale-out alarm is triggered.
- Add 1 step policy is applied.
- New instance is launched but not added to the aggregated group metrics until after **warm up** period expires.



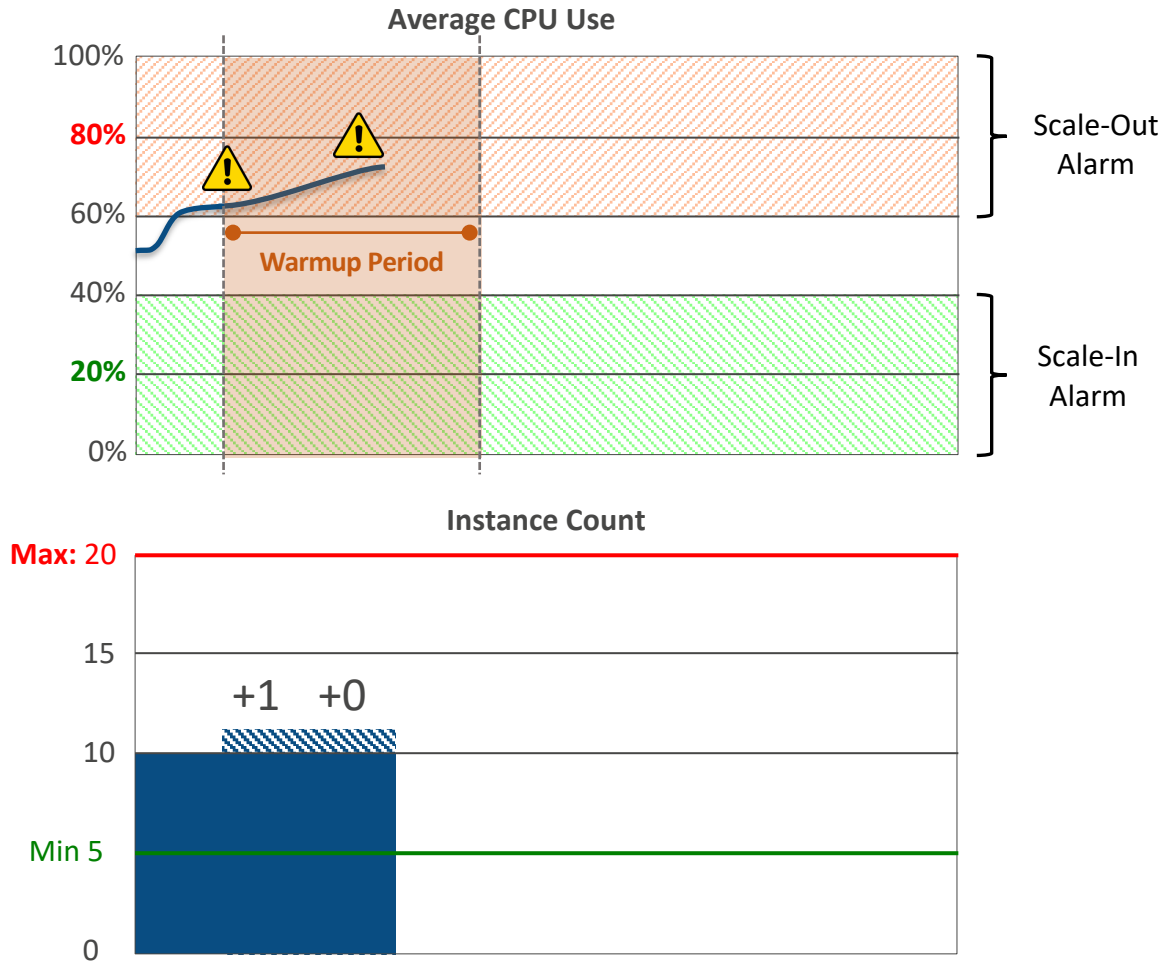
Auto Scaling Steps

As usage increases:

- CPU use goes up.

While waiting for new instance:

- CPU use remains high.
- Another alarm period is triggered.
- Since current capacity is still 10 during the warmup period, and desired capacity is already 11, **no additional instances are launched.**



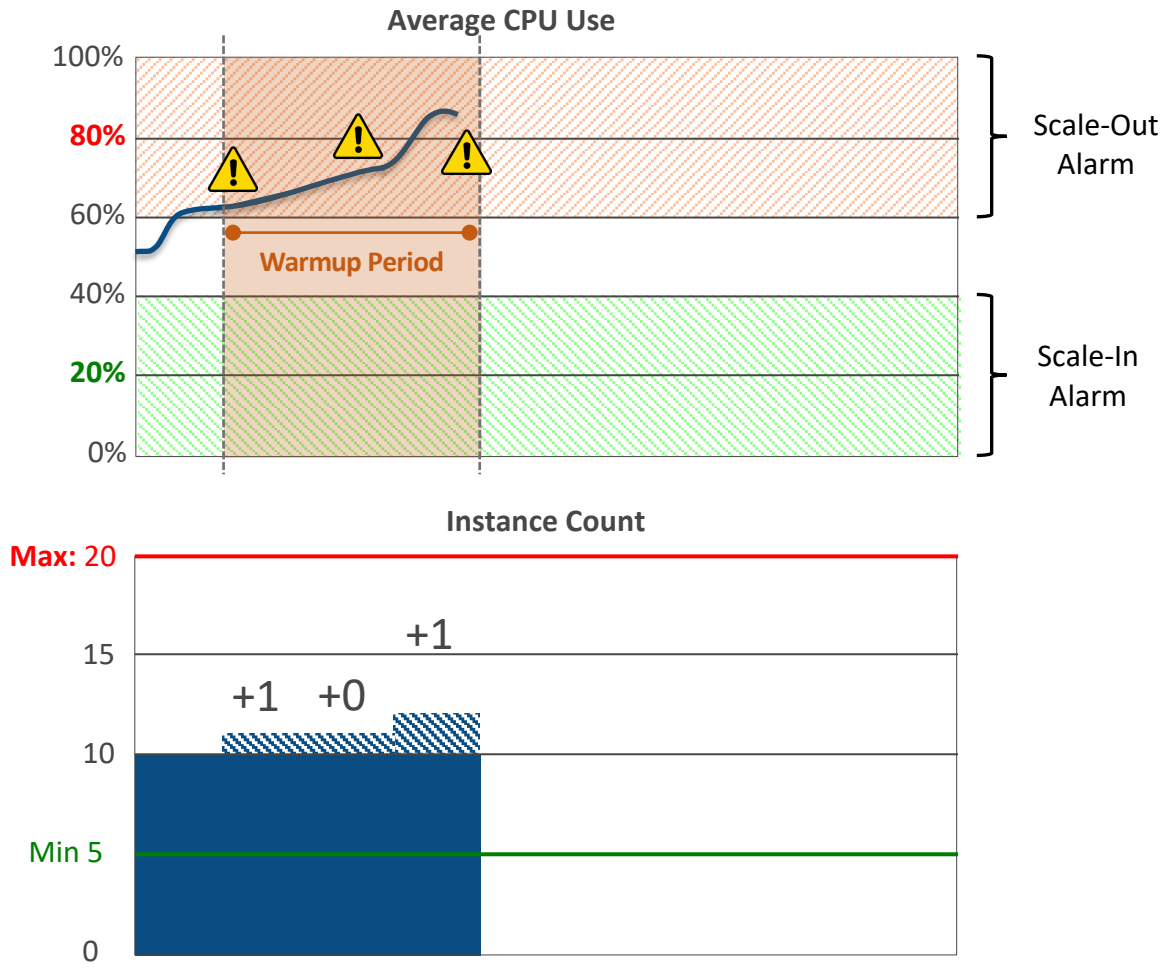
Auto Scaling Steps

As usage increases further:

- CPU use goes up.

When CPU use is 80-100%:

- Scale-out alarm is triggered.
- Add 2 step policy is applied.
- Since the alarm occurred during a warm up period, two instances are launched less the one instance added during the first alarm.
- Again new instances are not added to aggregated group metrics.



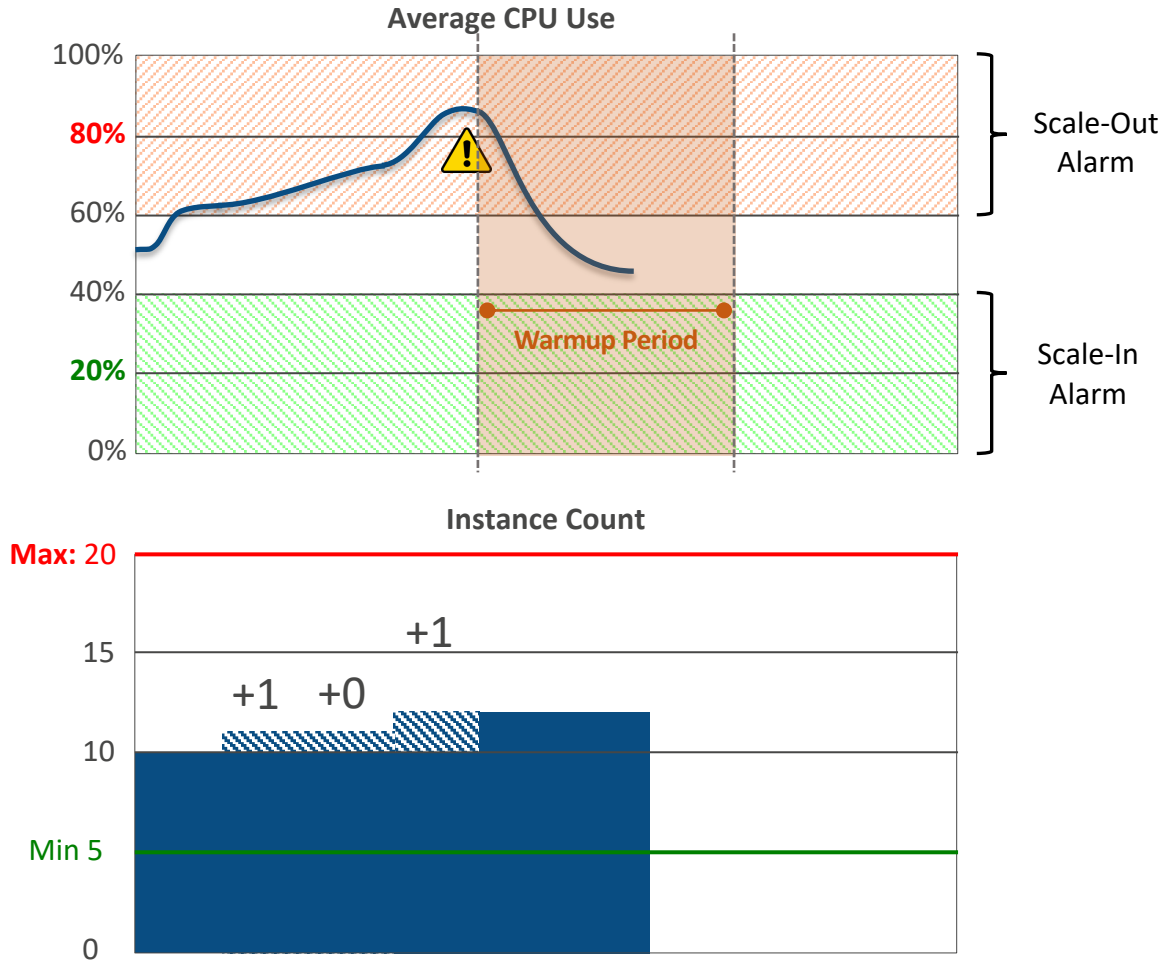
Auto Scaling Steps

As capacity matches usage:

- CPU use stabilizes.

When CPU use is 40-60%:

- No alarms are triggered.
- After warmup period expires, new instances are added to the aggregated group metrics.



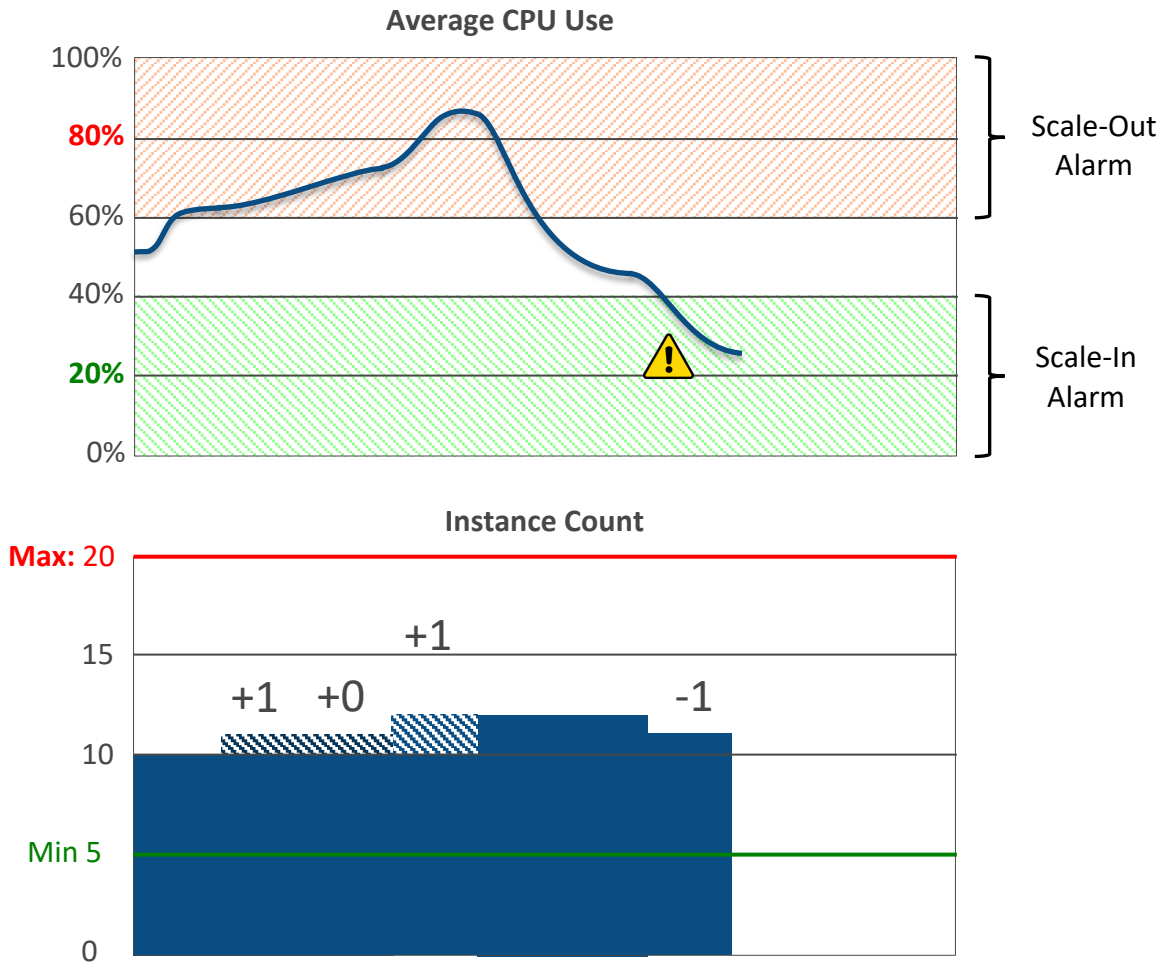
Auto Scaling Steps

As usage decreases:

- CPU use goes down.

When CPU use is 20-40%:

- Scale-in alarm is triggered.
- Remove 1 step policy is applied.
- An instance is removed from the Auto Scaling group and from the aggregated group metrics.



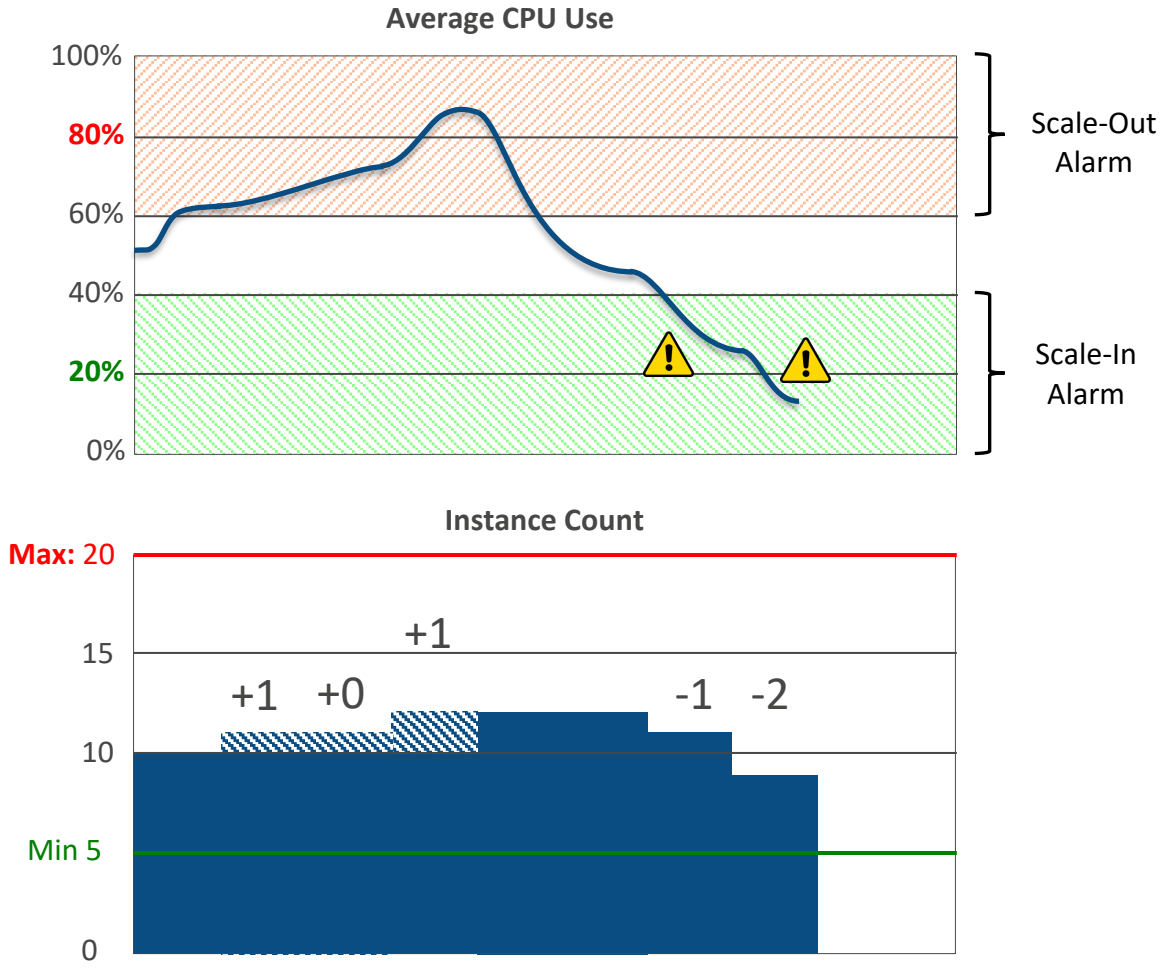
Auto Scaling Steps

As usage decreases:

- CPU use goes down further.

When CPU use is 0-20%:

- Scale in alarm is triggered.
- Remove 2 step policy is applied.
- Two instances are removed from the Auto Scaling group and from the aggregated group metrics.



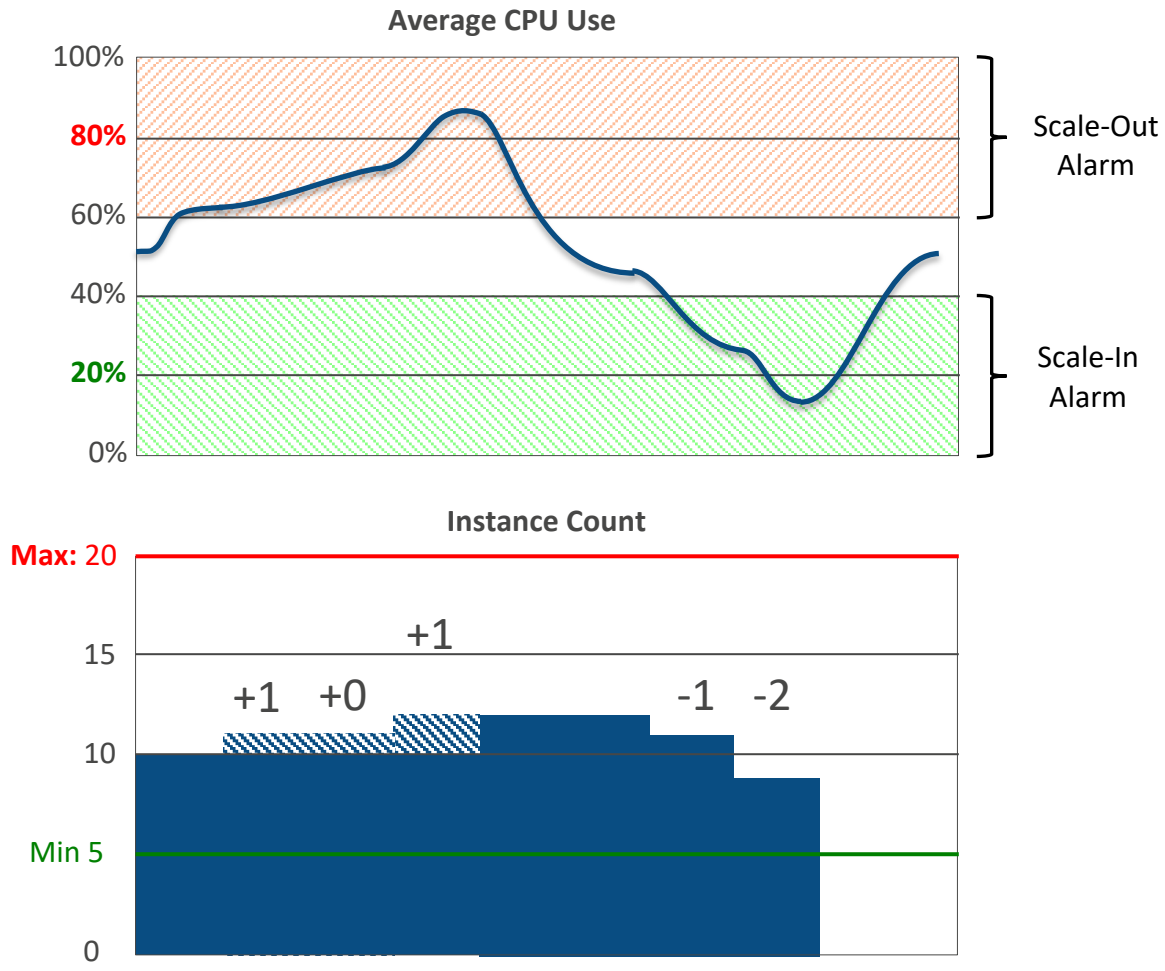
Auto Scaling Steps

As capacity matches usage:

- CPU use stabilizes.

When 40% < CPU Use < 60%

- No step adjustment is triggered.
- No step policies are applied.
- No instances are added or removed from service.



Auto Scaling Considerations

- Avoid Auto Scaling thrashing.
  - Be more cautious about scaling in; avoid aggressive instance termination.
  - Scale out early, scale in slowly.
- Set the min and max capacity parameter values carefully.
- Use lifecycle hooks.
  - Perform custom actions as Auto Scaling launches or terminates instances.
- Stateful applications will require additional automatic configuration of instances launched into Auto Scaling groups.

Remember: Instances can take several minutes after launch to be fully usable.

Learn more.

Lecture References

Recommend Viewing

Swinburne Lecture – High Level Overview

AWS Academy – Deeper dive

ACA Module 9