



# AWS Architecture Training

# Auto Scaling

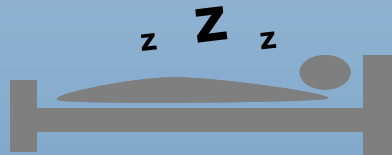
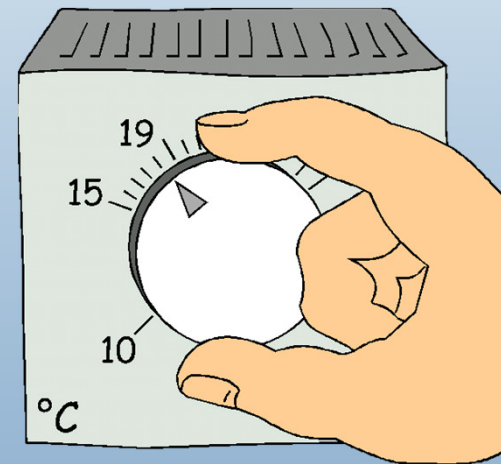
# Module 5

©2012 Amazon Web Services May not be reused or redistributed without permission



# Auto Scaling Overview

- Automatically Scale Server Farms
  - Scale Up/Down
  - (Re)Balance Across AZs
  - Add/Remove from ELB
- Set a Thermostat
  - Don't manage the furnace burners



# Types of scaling

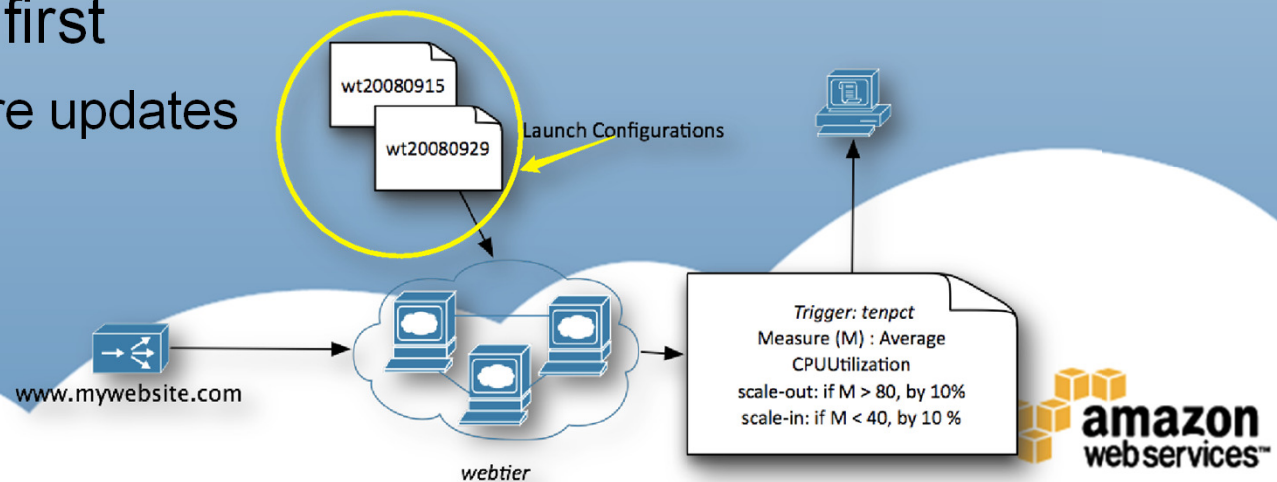
- Manual
  - Send an API call or use CLI to launch/terminate instances
  - Only need to specify capacity change (+/-)
- By Schedule
  - Scale up/down based on date and time
- By Policy
  - Scale in response to changing conditions, based on user configured real-time monitoring and alerts
- Automatic Rebalance
  - Instances are automatically launched/terminated to ensure the application is balanced across multiple AZs

# Auto Scaling Components

- Launch Configuration
- Auto Scaling Group
- Auto Scaling Policy
- CloudWatch Alarms

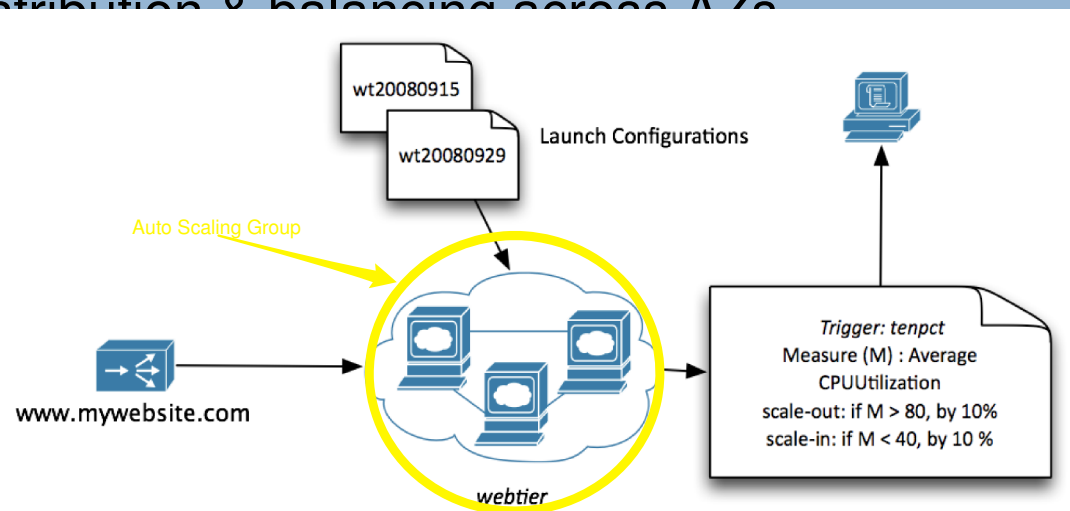
# Launch Configuration

- Describes what Auto Scaling will create when adding instances
  - AMI
  - Instance Type
  - Security Group
  - Instance Key Pair
- Only one active launch configuration at a time
- Auto Scaling will terminate instances with old launch configurations first
  - Rolling software updates



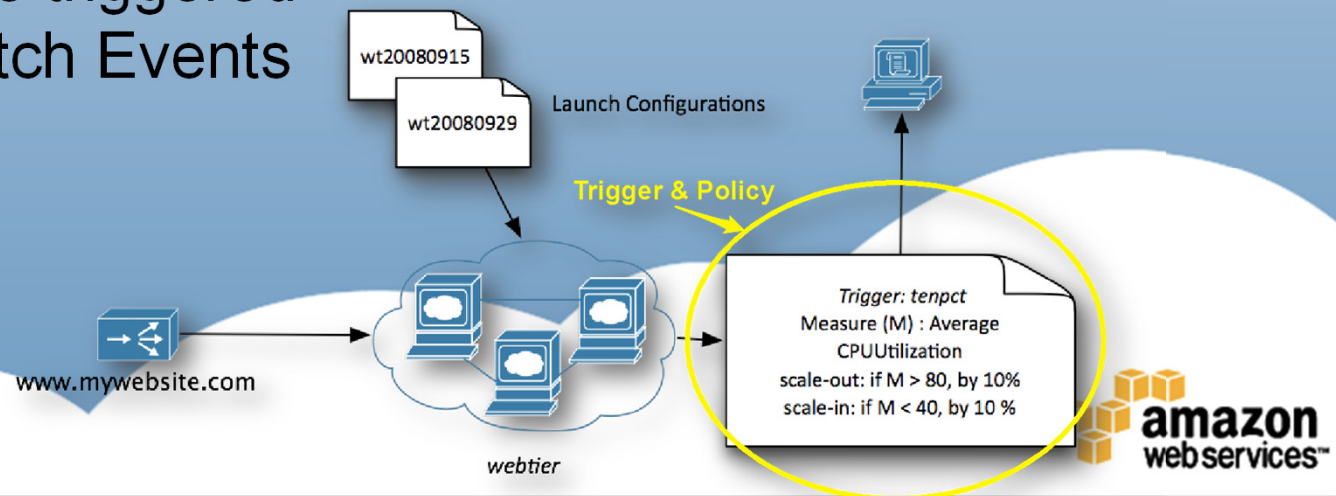
# Auto Scaling Group

- Auto Scaling managed grouping of EC2 instances
- Automatic health check to maintain pool size
- Automatically scale the number of instances by policy
  - Min, Max, Desired
- Automatic Integration with ELB
- Automatic Integration with AZs
  - Automatic distribution & balancing across AZs



# Auto Scaling Policy




- Parameters for performing an Auto Scaling action
  - Scale Up/Down
  - By how much
    - ChangeInCapacity (+/- #)
    - ExactCapacity (#)
    - ChangeInPercent (+/- %)
  - Cool Down (seconds)
- Policy can be triggered by CloudWatch Events



# CloudWatch Alarms

- Monitors a CloudWatch metric
  - Threshold (> 50% CPU)
  - Period (for 3 minutes)

- Alarm States

Icon	State	Description
	OK	Within defined threshold
	ALARM	Outside defined threshold
	INSUFFICIENT_DATA	Metric does not have enough data to determine state

- Available Actions
  - Trigger Auto Scaling Policy (scale up/down event)
  - Send SNS notification



# Scaling Activity

- Instance Launch
  - Instances are launched (not started) from a “gold” image
  - Bootstrapping is important!
  - Automatically added to ELB (if configured)
- Instance Termination
  - Instances are terminated (not stopped)
  - Longest running instance from the oldest launch config first
  - Automatically removed from ELB (if configured)
- Cooldown
  - The period after an Auto Scaling activity during which no other scaling activity can take place
  - Gives the system time to perform and adjust to before executing a new scaling activity
  - 300 seconds (5 minutes) by default

# Auto Scaling Considerations

- Suspend/Resume Processes
  - Allows you to manually suspend and resume Auto Scaling activities
- EC2 Bills Hourly
  - Less obvious when manually launching instances
  - Can be costly with Auto Scaling “thrashing”
- AWS Recommendation
  - Scale up quickly, scale down slowly
  - No extra cost as long as it's within an hour

The Auto Scaling lab examples do not follow this advice for illustrative purposes and should not be considered best practice or recommended settings!



©2012 Amazon Web Services May not be reused or redistributed without permission

