# DAY- 8
# AWS
# AUTO SCALING

# AWS Architecture and Design

1. Day 1 Overview of Cloud Computing
2. Day 2 Overview of AWS
3. Day 3 Amazon EC2*
4. Day 4 Amazon EBS *
5. Day 5 Amazon CloudWatch *
6. Day 6 Amazon S3*
7. Day 7 Amazon Elastic Load Balancer *
8. **Day 8 Amazon Auto Scaling ***
9. Day 9 Amazon VPC *
10. Day 10 Amazon IAM *
11. Day 11 Amazon RDS
12. Day 12 Amazon Route 53 *
13. Day 13 Amazon DynamoDB* & Glacier
14. Day 14 Amazon Cloudfront* & Import Export & Amazon SES *
15. Day 15 Amazon ElasticBeanStalk & Amazon Cloudformation & Amazon OpsWorks
16. Day 16 AWS Economics & AWS Account Overview *
17. Day 17 AWS Architecture
18. Day 18 AWS Certification Preparation

[ * - With Hands on Demo]

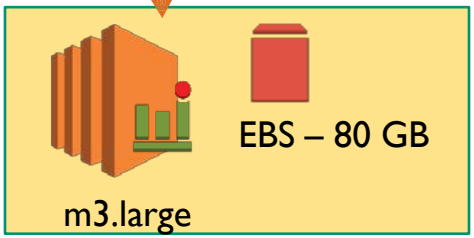
# AWS Auto Scaling

- → Scaling Types
- → What is Auto Scaling
- → Types of Auto Scaling
- → Auto Scaling Deep down

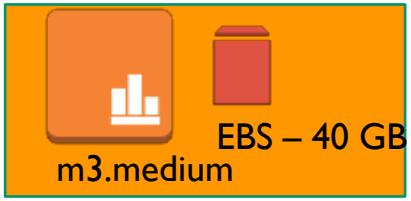
# Amazon Auto Scaling

# Scaling

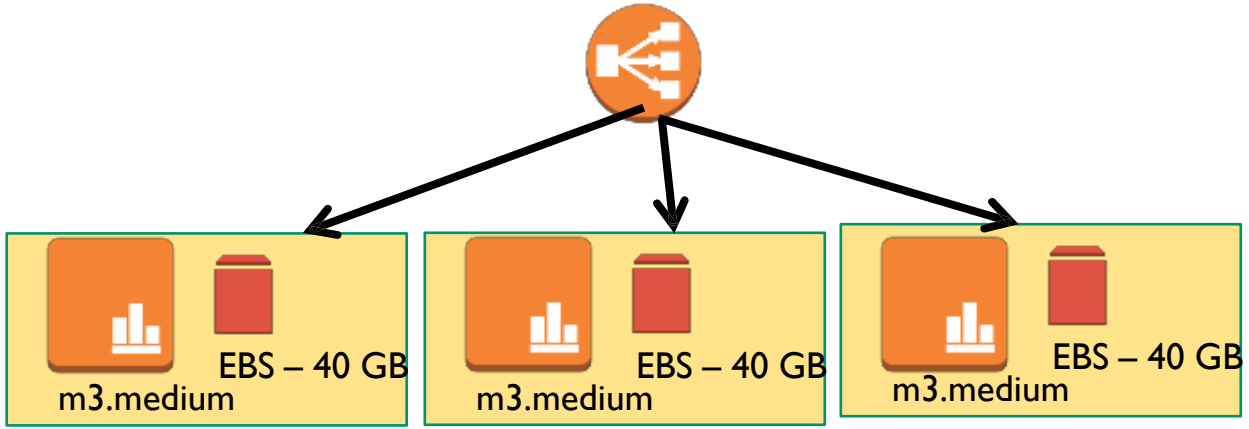





Vertical Scaling




Horizontal Scaling



# Amazon AutoScaling

- AWS Auto Scaling help achieve horizontal scalability of your application
  - ✓ Helps achieve High Availability
  - ✓ Scale Up & Down EC2 capacity
  - ✓ Maintain the desired capacity
  - ✓ Increase – decrease capacity seamlessly based on demand
  - ✓ Cost Optimization
- Ideal for hourly, daily, or weekly variability. Its not very well suited for spikes.
- Works with ELB & CloudWatch
- Free

http://aws.amazon.com/autoscaling/

# Auto Scaling Features

Elastic

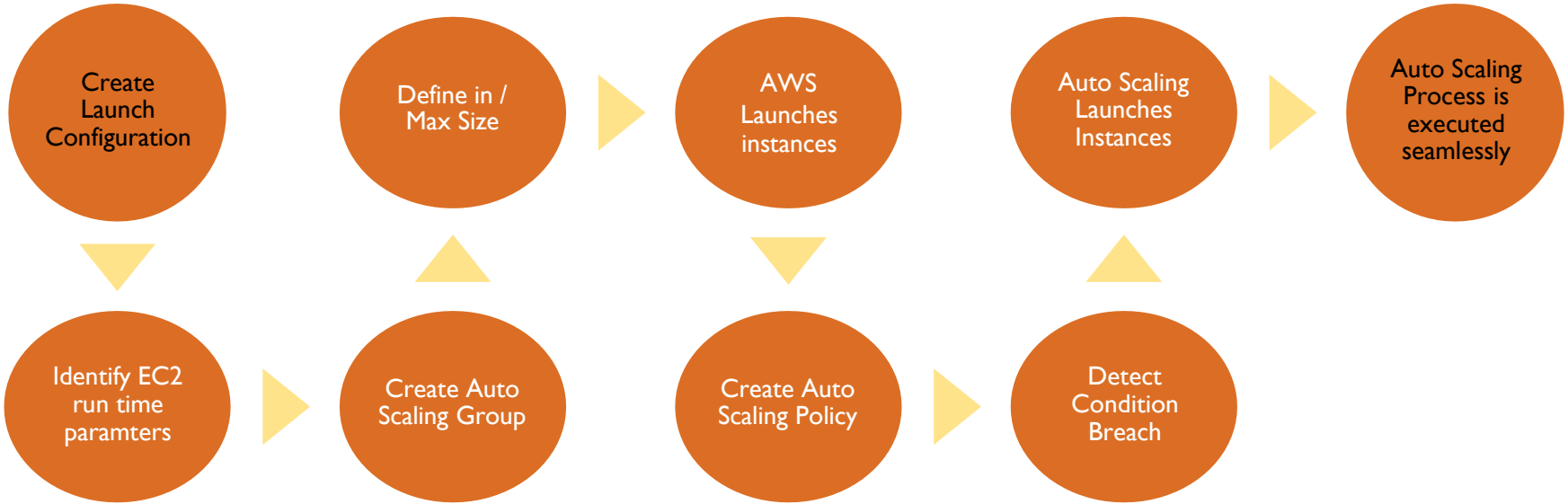
Dynamic Scaling

Cost Savings

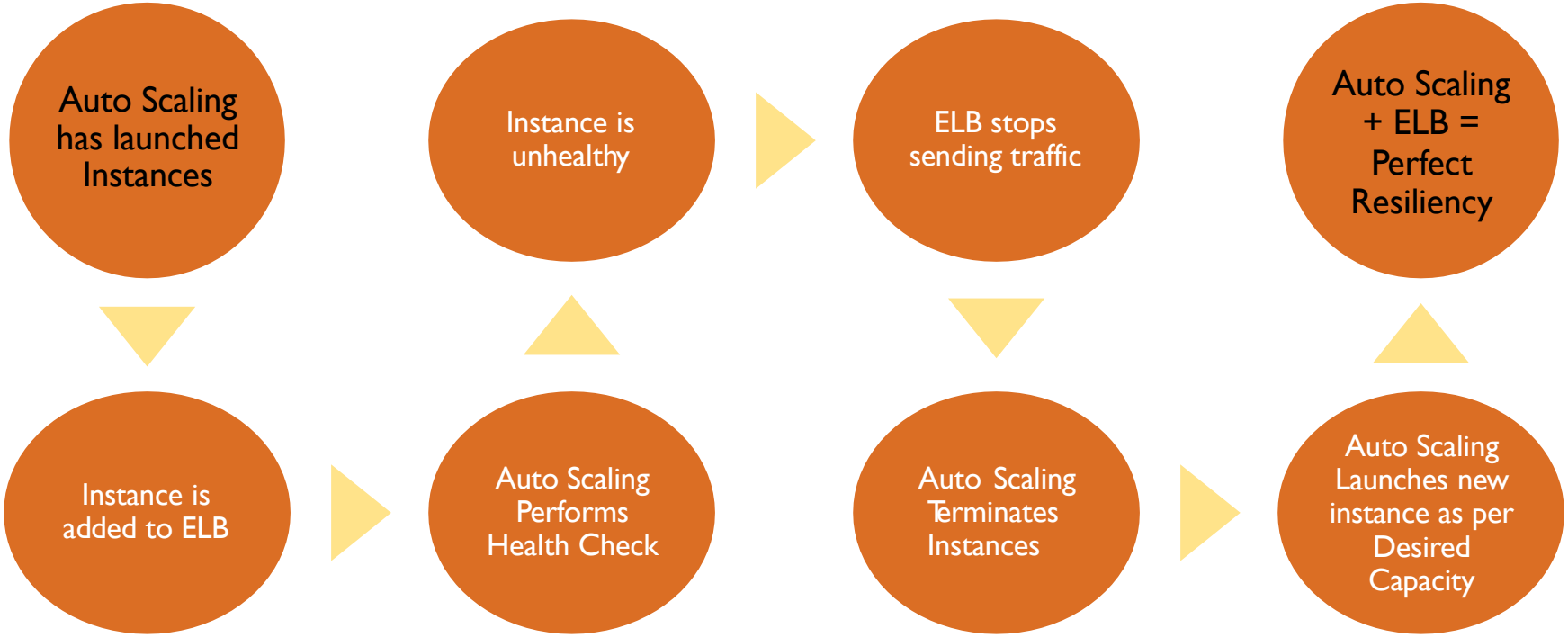
HA

With in Region

# Auto Scaling Features



# Auto Scaling + ELB : Resiliency

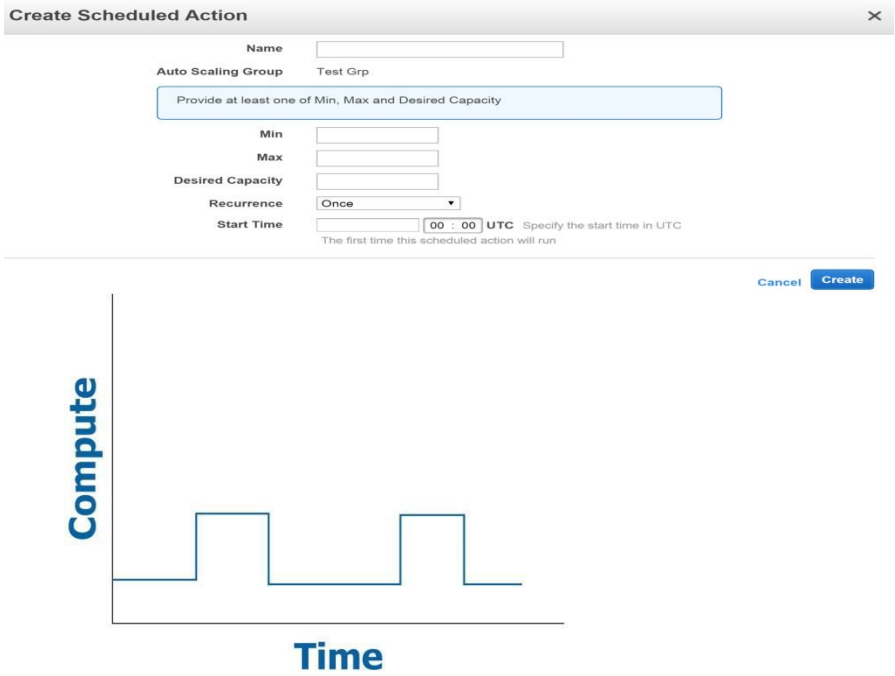


# Applying Amazon Auto Scaling for 3 Different Load Scenarios

# Auto Scaling Policy : Scaling Based on Time

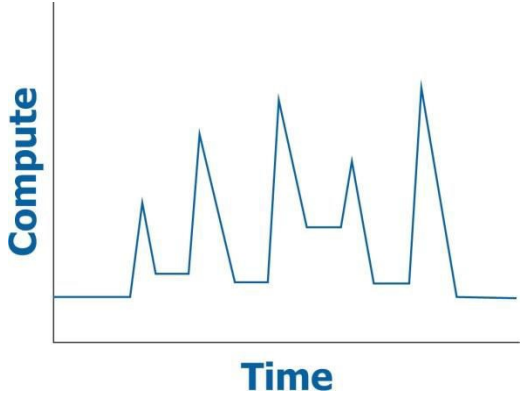
**I. Predictable Bursts**

- ✓ When there is a fixed Pattern (e.g. 9 AM to 5 PM highest load)
- ✓ Configure Time-Based auto Scaling Plan.
- ✓ You can configure as one time or recurring



# Auto Scaling Policy: Scaling Based on Condition

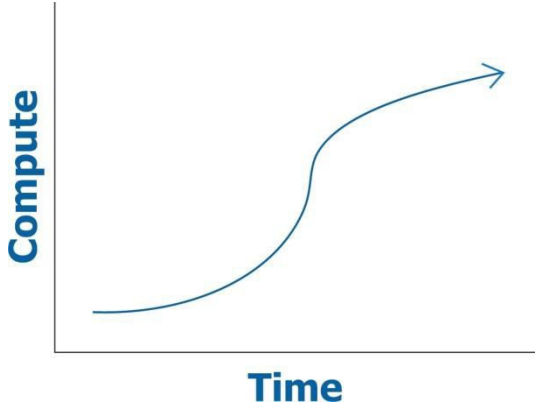
**2. Dynamic Scaling**




- ✓ Configure On demand Auto Scaling Policy based on conditions
- ✓ Create policies for both Scaling out and Scaling In scenarios.
- ✓ Auto Scaling will respond to changing conditions dynamically based on conditions defined
- ✓ Supports Simple & Step wise Scaling

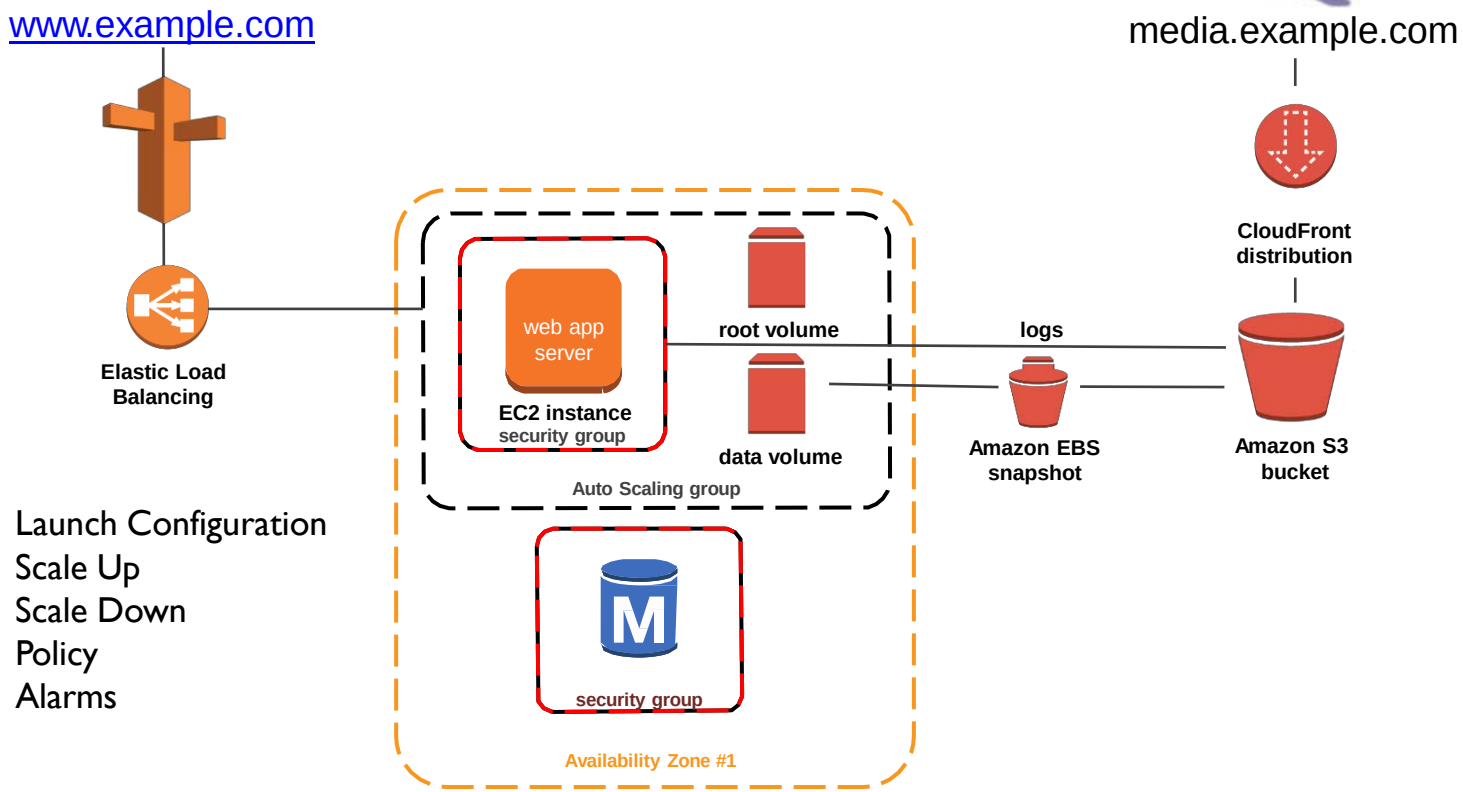
# Auto Scaling Policy: Manual Scaling

### 3. Manual Scaling / Constantly Growing




- ✓ When you do not want to follow certain pattern (time) or dynamic but want to scale manually
- ✓ Periodically monitor the load requirements and manually configure the Auto Scaling parameters.
- ✓ Modify the minimum fixed capacity gradually.

# Amazon AutoScaling Architecture




Launch Configuration
Scale Up
Scale Down
Policy
Alarms

# Auto Scaling Vocabulary

- ✓ **Launch Configuration :** First step for each Auto Scaling group create. Takes EC2 instance launch configurations which is used while scaling out new instance.
- ✓ **An Auto Scaling Group :** It is a representation of multiple Amazon EC2 instances that share similar characteristics and are treated as a logical grouping for the purposes of instance scaling and management.
- ✓ **A policy** is a set of instructions for Auto Scaling that tells the service how to respond to CloudWatch alarm messages. An Amazon CloudWatch alarm is an object that watches over a single metric. An alarm can change state depending on the value of the metric
- ✓ **Triggers:** A trigger is a concept that combines two AWS features: a CloudWatch alarm and an Auto Scaling policy that describes what should happen when the alarm threshold is crossed. You can set a trigger to activate on any metric published to Amazon CloudWatch, such as CPU Utilization.
- ✓ **Health Check :** It is a call to check on the health status of each instance in an Auto Scaling group. If an instance reports degraded performance, Auto Scaling terminates the instance and launches another one to take its place.

# Use Case For Auto Scaling

- ✓ **Scale Web tiers** (Apache, Nginx etc)
- ✓ **Scale Application tiers** (Tomcat, Jboss, etc)
- ✓ **Manage Load Balancing Tiers** (HAProxy, Nginx, ELB etc)
- ✓ **Idle for Stateless Tiers**

# Types of Scaling

- **Manual Scaling**
  Send an API call or use the Auto Scaling command line interface (CLI) to launch or terminate an Amazon EC2 instance. You need only specify the change in capacity you want.

- **Scaling by Schedule**
  Scaling by schedule means that scaling actions are performed automatically as a function of time and date.

- **Scaling by Policy**
  A more advanced way to scale your resources, scaling by policy, lets you define parameters that inform the Auto Scaling process. **For example**, you can create a policy that calls for enlarging your fleet whenever the average CPU utilization rate stays above ninety percent for fifteen minutes.

**Note:** You should have two policies, one for scaling up and one for scaling down, for each event that you want to monitor.

# Scaling– In-depth

- ✓ Scaling frequently with Spikes is not recommended
- ✓ One account can create maximum 100 launch configs
- ✓ Suspend or resume the Auto Scaling activity as per need
- ✓ Maximum 125 scheduled scaling actions are allowed per Auto Scaling group so roughly it allows 4 actions/day scaling schedules in a month
- ✓ The user can define the instance termination policy. It first identifies the zone with the maximum number of instance and then implement one of the below termination policy:
  - ✓ Delete the oldest instance (Default policy)
  - ✓ Delete Newest Instance first
  - ✓ Delete instance with oldest launch config
  - ✓ Delete Instance with Closest to Next Instance Hour

# Scaling– In-depth

- ✓ EBS backed AMIs boots faster than S3 backed AMI
- ✓ You can not update Launch Configuration but can modify Auto Scaling group
- ✓ The user can use Spot/Reserved Instance while using Auto Scaling
- ✓ Auto Scaling Group works across multiple AZs in a single region but can not span across regions
- ✓ AS can work with or without ELB.

# Scaling – Health Check

- ✓ Auto Scaling checks the health of the EC2 instances by doing status checks periodically.
- ✓ If Auto Scaling is configured with ELB, the user can also configure the health check to be performed for ELB.
- ✓ If the status returned from the EC2 health check is anything other than “OK” and for ELB it is any status is anything other than “In Service”, Auto Scaling will mark that instance or ELB as unhealthy.
- ✓ If an ELB is unhealthy, all its instances will also be marked as unhealthy.

# Scaling – Health Check

- ✓ Auto Scaling sends metrics to CloudWatch for monitoring. These are the same metrics that are available for any EC2 instance, even if it is not in an Auto Scaling group.

- ✓ Enable Detailed monitoring for EC2 instances launched by Auto Scaling.

- ✓ The Auto Scaling also has group metrics such as:
  - ✓ GroupMinSize (Minimum number of Instances),
  - ✓ GroupMaxSize (Maximum number of Instances),
  - ✓ GroupDesiredSize (Desired Capacity of a Group),
  - ✓ GroupInServiceInstances (The Number of running instances),
  - ✓ GroupPendingInstances(The Number of pending instances),
  - ✓ GroupTotalInstances (The total number of Instances) and a few more

# Scaling– Guidelines

- ✓ Auto Scaling should not be used with Database except read only DB.
- ✓ If you have critical configuration / log / audit trail files then keep moving it to AWS S3 at regular interval.
- ✓ Optimize resources after monitoring patterns.

# Demo

In the next video we will do hands on with AWS Auto Scaling

# Thank You

Email us – support@intellipaat.com

Visit us - https://intellipaat.com