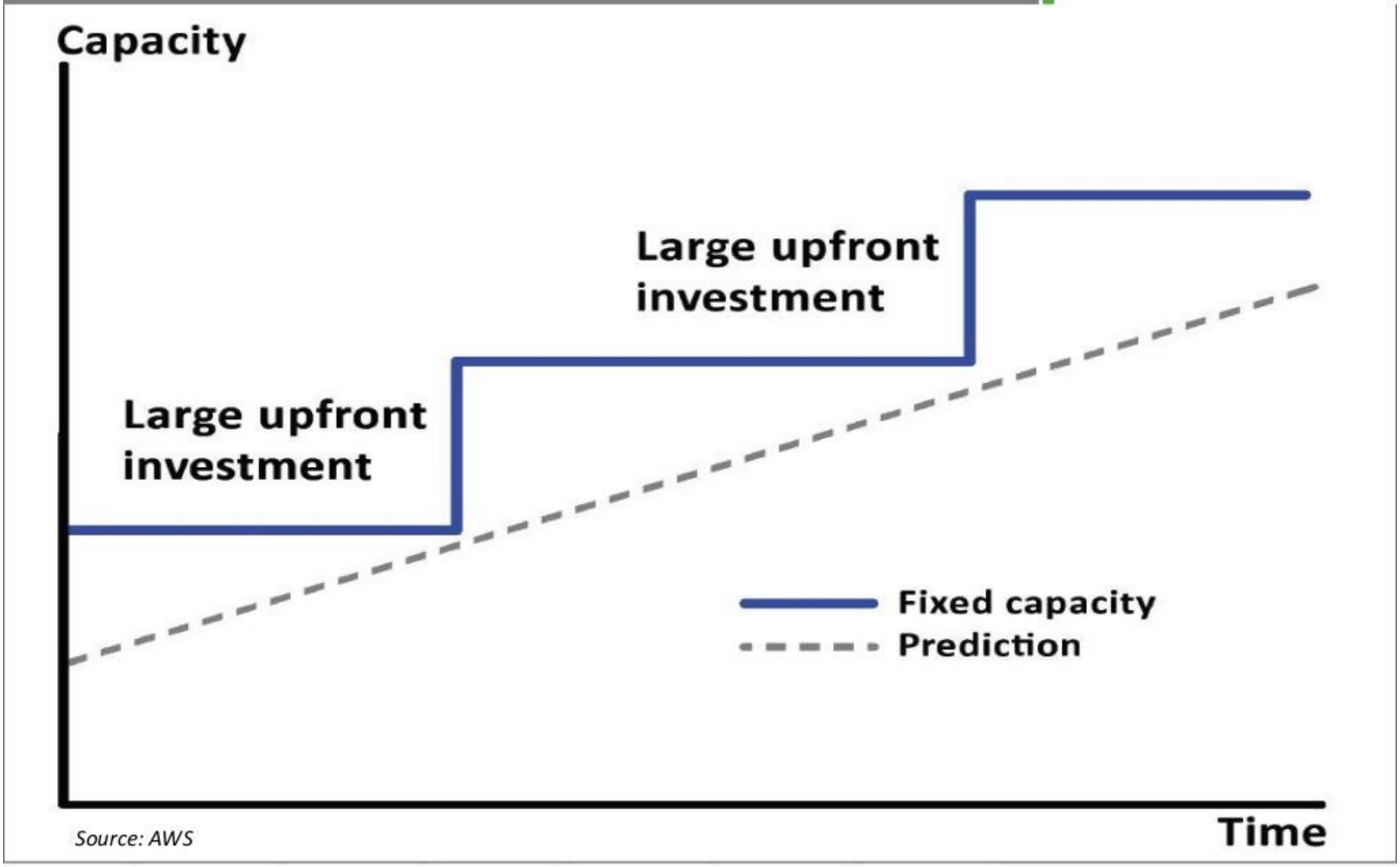
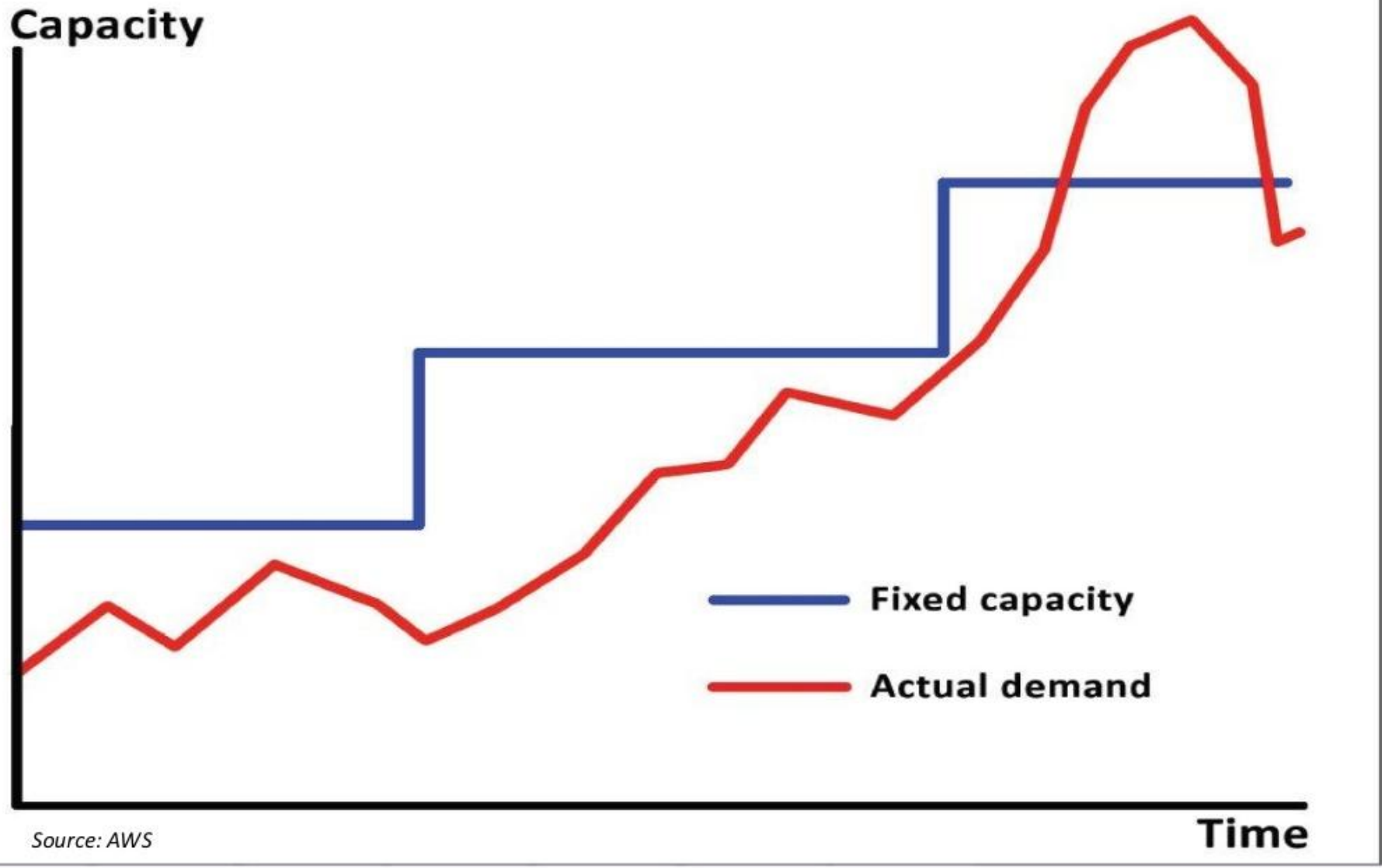
AWS auto scaling group

Need for Auto Scaling

Upfront Capacity Investment

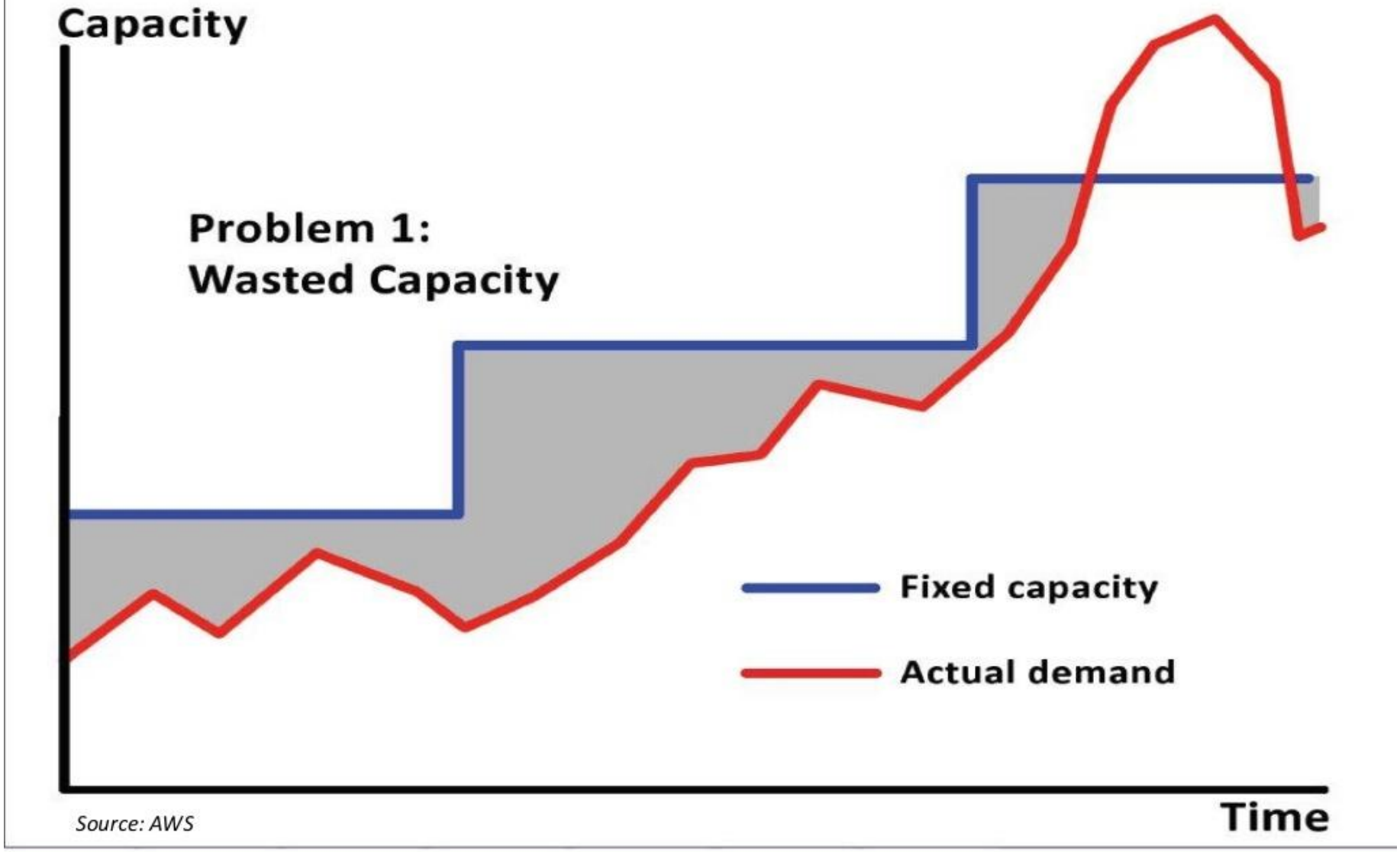


Actual Demand vs Fixed Capacity



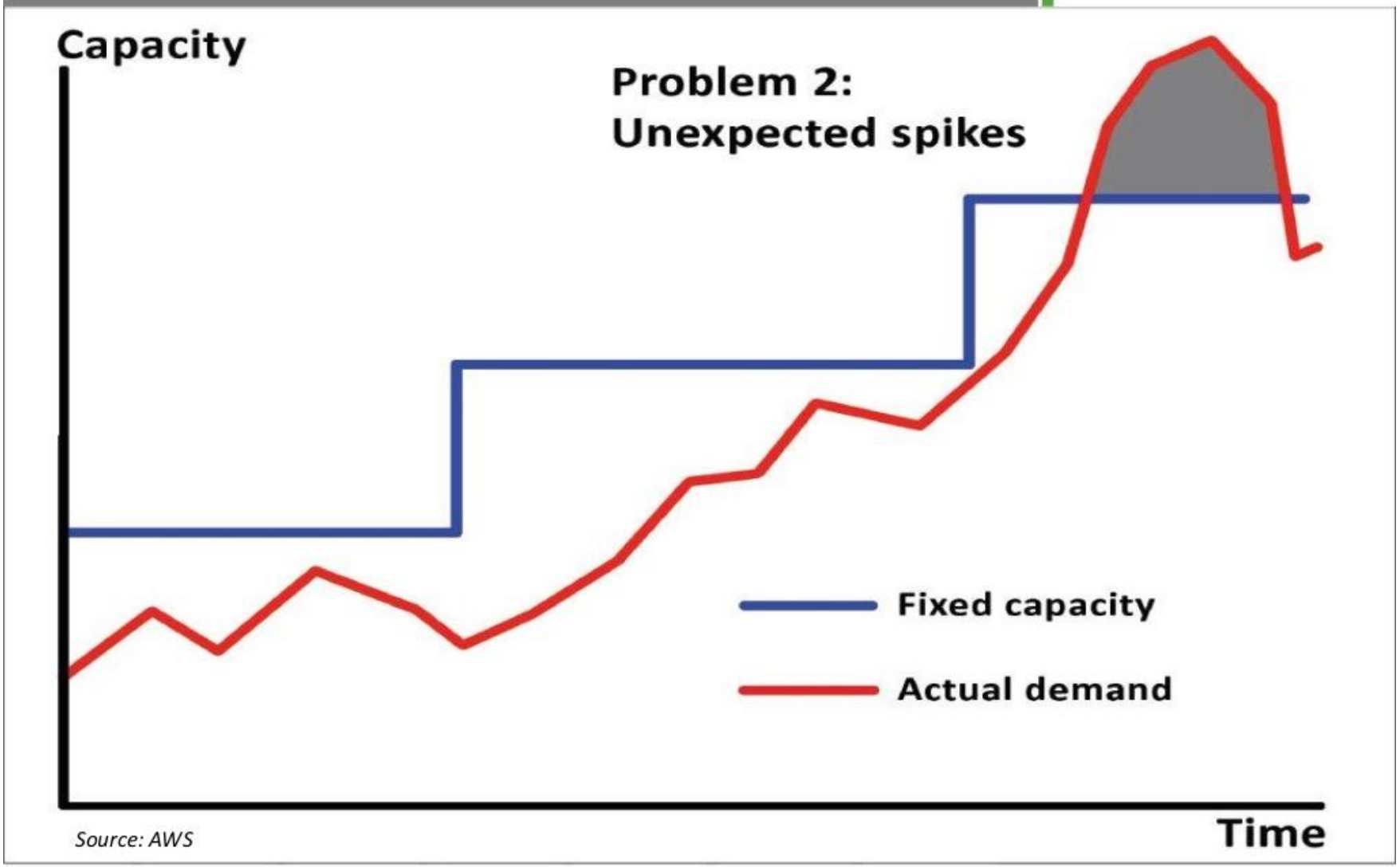
Source: AWS

Problem 1: Wasted Capacity



Source: AWS

Problem 2 : Lost Customers



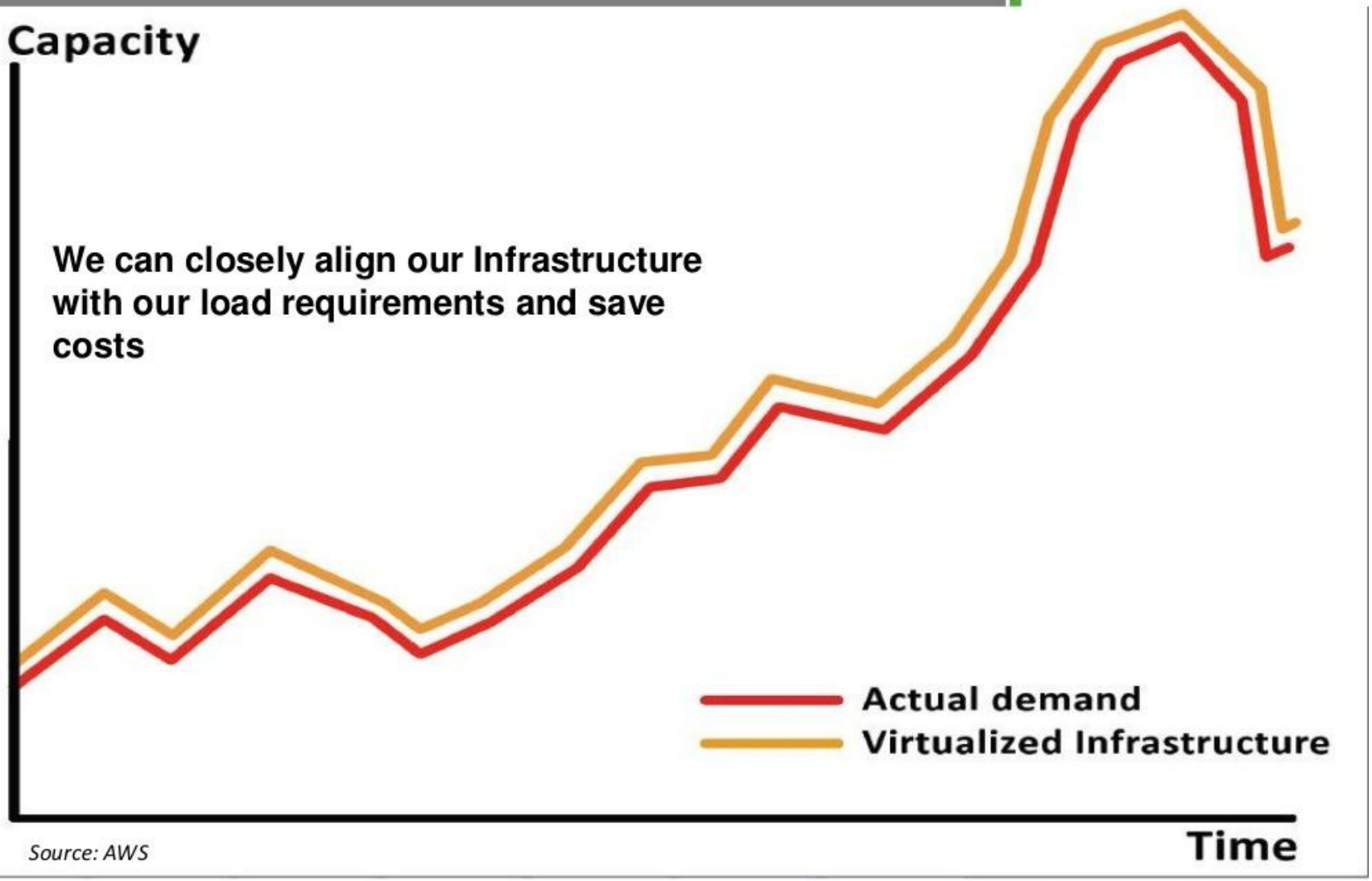
Need for Auto Scaling

What is AWS Auto Scaling ?

AWS Auto Scaling allows us to scale our Amazon EC2 capacity out or down automatically according to the load patterns . Example :

- We can **expand** the number of Amazon EC2 instances from 1 to 100+ automatically during load peaks
- We can **reduce** the number of Amazon EC2 instances from 100+ to 1 automatically during load valleys

AWS Auto Scaling



Source: AWS

What Amazon Auto Scaling can do ?

- Handle all the 3 load scenarios (Candidates)
- Scale out Amazon EC2 instances seamlessly and automatically when demand increases
- Scale down unwanted Amazon EC2 instances automatically and save money when demand subsides
- Decide the scaling based on AWS CloudWatch metrics
- Auto Scale your Web servers(Amazon EC2) in combination with AWS Elastic Load Balancing

How much does it Cost ?

Cost for using AWS Auto Scaling service = **0 $**

Value= **PRICELESS**

Note : AWS Auto scaling needs Amazon CloudWatch monitoring service to function . Amazon CloudWatch is billed on usage basis.

Some AWS Auto scaling Concepts

Auto Scaling group : Logical grouping of multiple Amazon EC2 instances for easy scaling and Management

Health Check: Calls to check on the health status of each Amazon EC2 instance in an Auto Scaling group

Launch Configuration: Captures the parameters necessary to create new EC2 instances in Auto Scaling mode

Some AWS Auto scaling Concepts

Triggers: A CloudWatch alarm and an Auto Scaling policy that describes the actions when the alarm threshold is crossed . Two Triggers – Scaling out and Scaling down needs to be created

Policy : Set of instructions for Auto Scaling that tells the service how to respond to AWS CloudWatch alarm messages

Risks involved in AWS Auto Scaling

Risk 1: AWS Auto Scaling takes between 30 – 180 seconds sometimes to launch a new instance(s) . This intermediate time may cause impaired performance for our customers

Risk 2: AWS Auto Scaling cannot differentiate between valid (vs) malicious traffic , it can scale out servers even for malicious traffic

1)How will you change the instance type for instances which are running in your application tier and are using Auto Scaling. Where will you change it from the following areas?

Auto Scaling policy configuration
Auto Scaling group
Auto Scaling tags configuration
Auto Scaling launch configuration

Answer D.

Explanation: Auto scaling tags configuration, is used to attach metadata to your instances, to change the instance type you have to use auto scaling launch configuration

2) **You have a content management system running on an Amazon EC2 instance that is approaching 100% CPU utilization. Which option will reduce load on the Amazon EC2 instance?**

Create a load balancer, and register the Amazon EC2 instance with it
Create a CloudFront distribution, and configure the Amazon EC2 instance as the origin
Create an Auto Scaling group from the instance using the CreateAutoScalingGroup action
Create a launch configuration from the instance using the CreateLaunchConfigurationAction

Answer A.

Explanation: Creating alone an autoscaling group will not solve the issue, until you attach a load balancer to it. Once you attach a load balancer to an autoscaling group, it will efficiently distribute the load among all the instances. Option B – CloudFront is a CDN, it is a data transfer tool therefore will not help reduce load on the EC2 instance. Similarly the other option – Launch configuration is a template for configuration which has no connection with reducing loads.


Auto Scaling will keep trying to launch the instance for 72 hours
Auto Scaling will suspend the scaling process
Auto Scaling will start an instance in a separate region
The Auto Scaling group will be terminated automatically

Answer B.

Explanation: Auto Scaling allows you to suspend and then resume one or more of the Auto Scaling processes in your Auto Scaling group. This can be very useful when you want to investigate a configuration problem or other issue with your web application, and then make changes to your application, without triggering the Auto Scaling process.

4) What happens to my ec2 instances after deletion of autoscaling group associated with it?

5) How do I attach new intsances to auto scaling group?