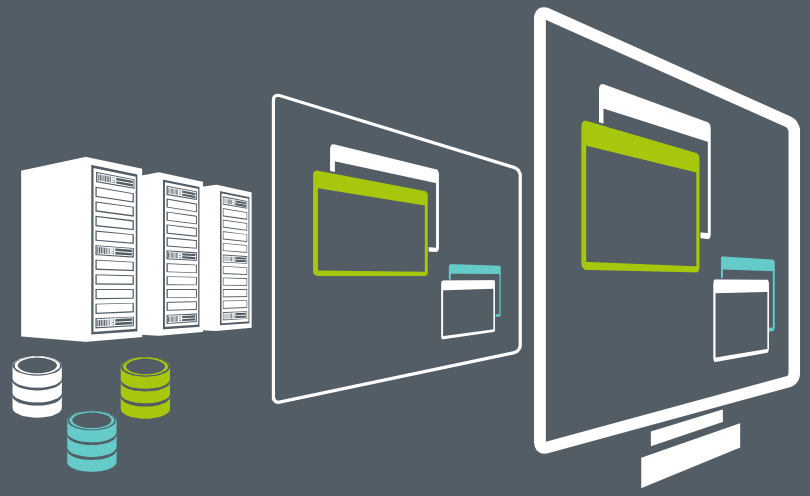
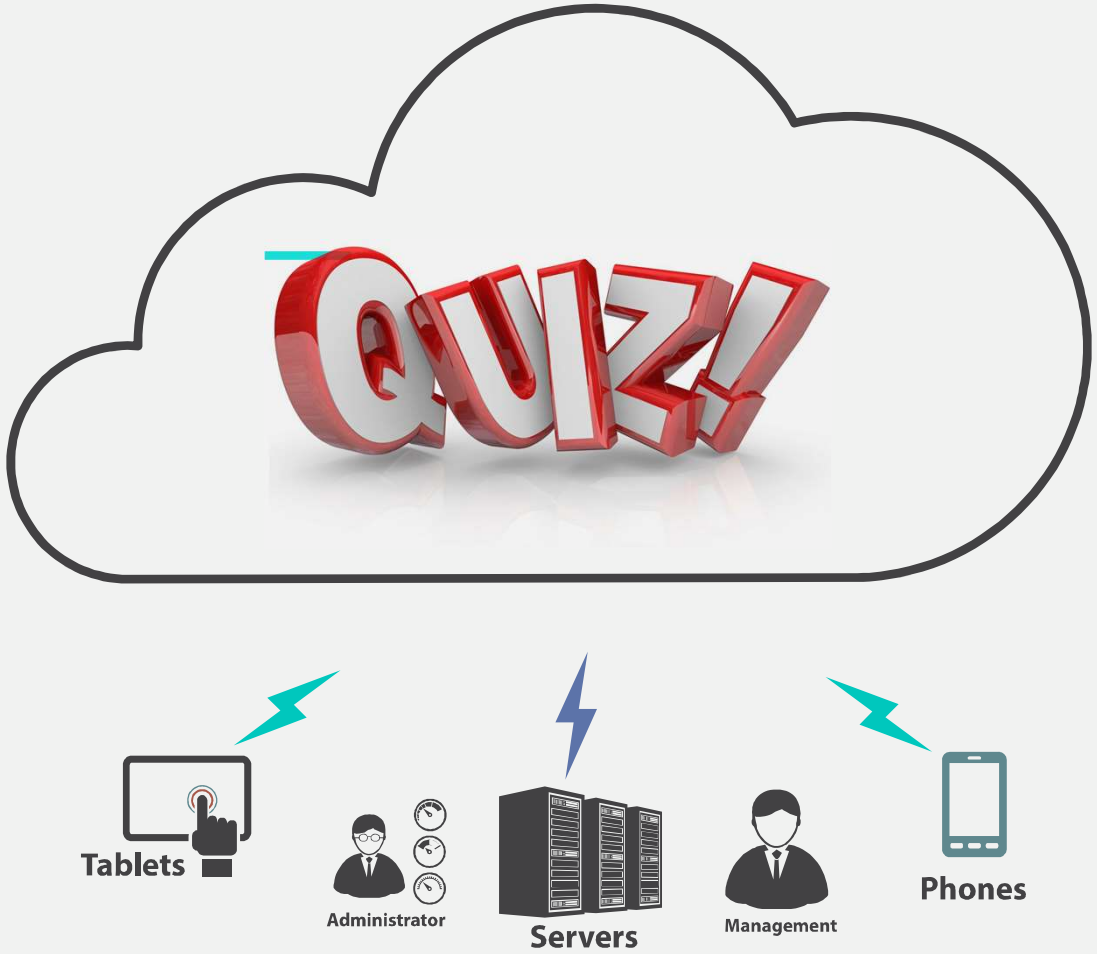
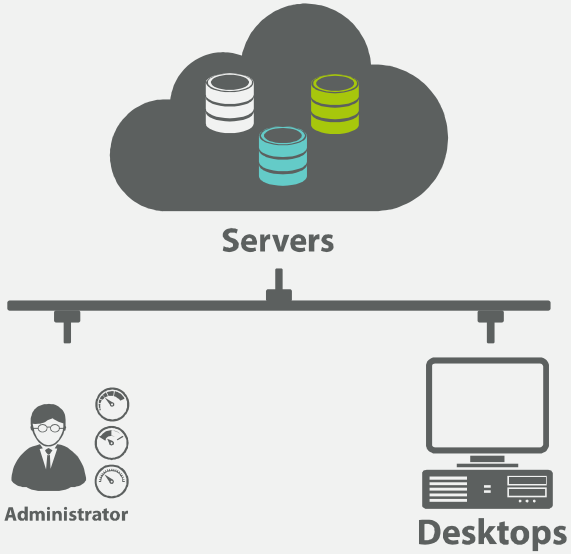


클라우드 아키텍처 구조

탄력적인 Resource 운영을 위한 Service

MEGAZONE CLOUD

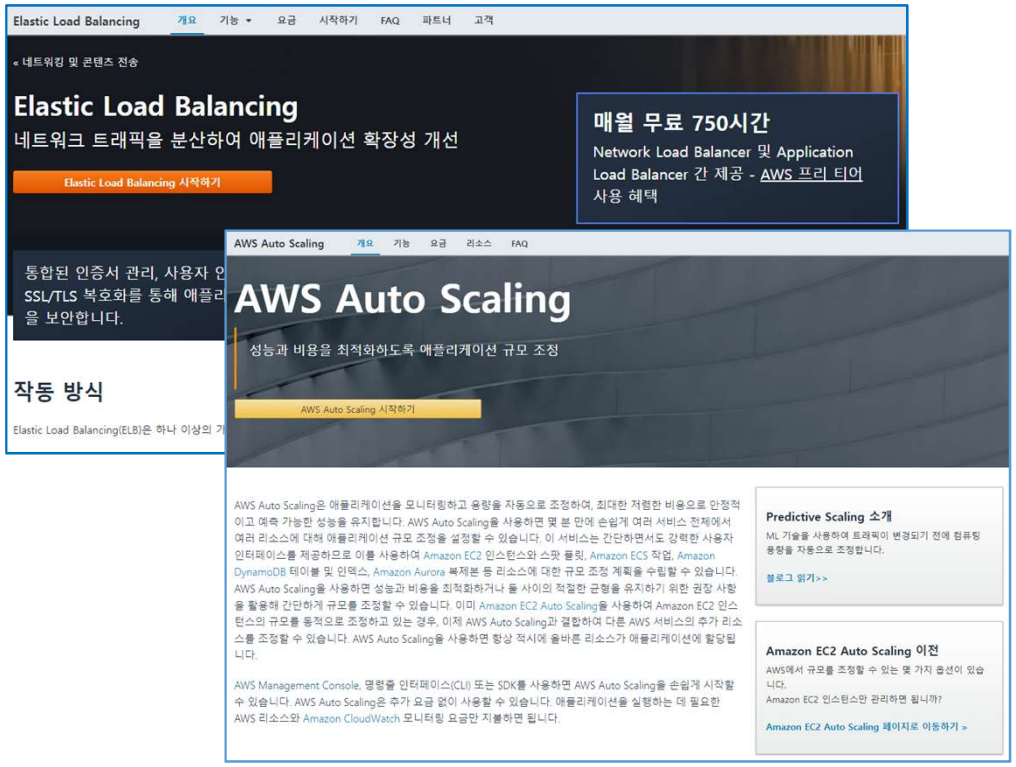

QUIZ!
Tablets
Administrator
Servers
Management
Phones



Index

수업 목표

개요

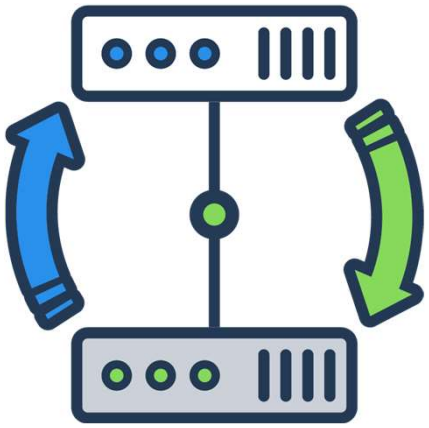


- 가용성과 확장성
- AWS Elastic Load Balancer
- AWS EC2 Auto Scaling

AWS ELB

AWS Elastic Load Balancing

Availability(HA)



https://www.pikpng.com/transpng/iJbmJRo/

- Is a quality of computing infrastructure.
- Allows it to continue functioning, even when some of its components fail.
- This is important for mission-critical systems that cannot tolerate interruption in service, and any downtime can cause damage or result in financial loss.

AWS Elastic Load Balancing

Availability(HA)

TABLE 1: Translating the Metrics

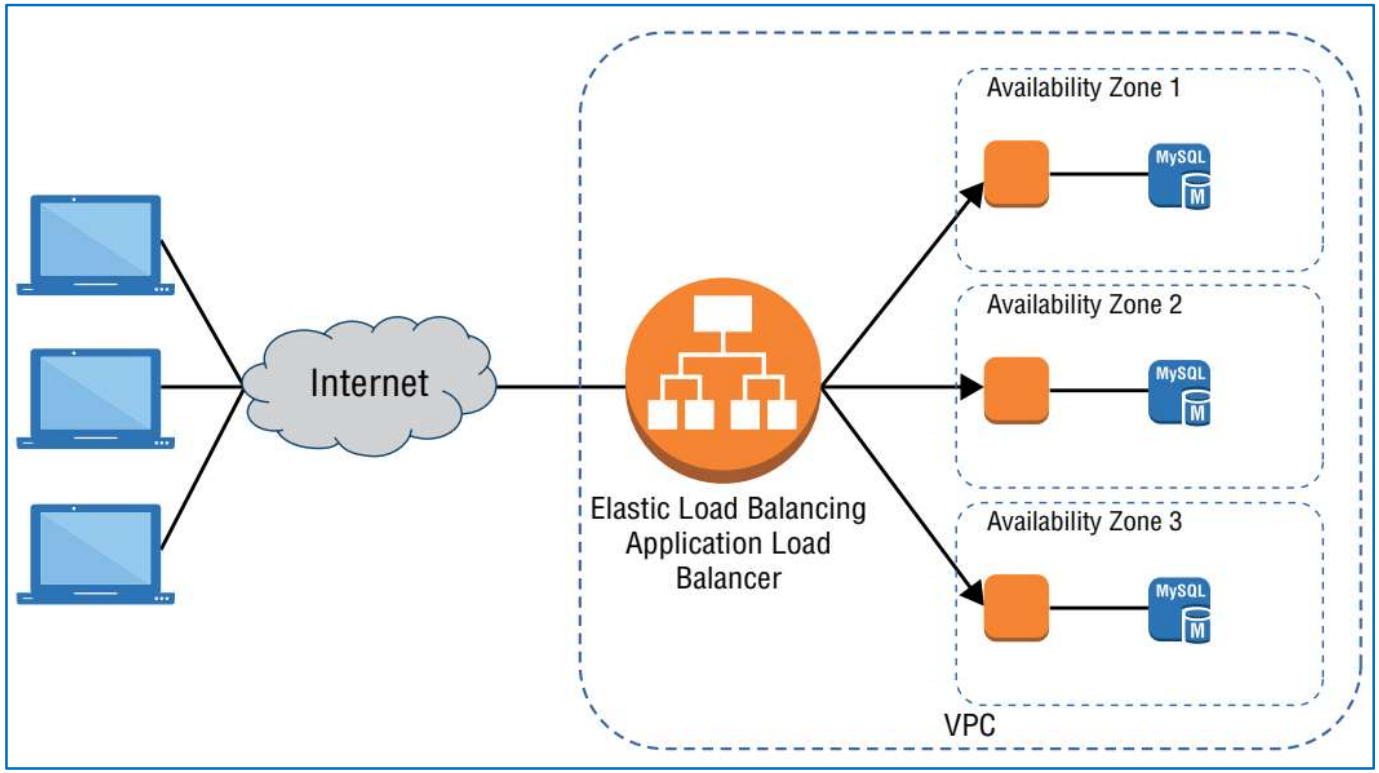
Availability	Downtime Per Year (3651/4 × 24)
99.9999%	32 seconds
99.999%	5 minutes, 15 seconds
99.99%	52 minutes, 36 seconds
99.95%	4 Hours, 23 minutes
99.9%	8 Hours, 46 minutes
99.5%	1 day, 19 hours, 48 minutes
99%	3 days, 15 hours, 40 minutes

https://www.nojitter.com/slas-burden-enterprise

- Guarantees a certain percentage of uptime.
- 99.9% uptime will be down only 0.1% of the time, 0.365 days or 8.76 hours per year.
- The number of “nines” is commonly used to indicate the degree of high availability.
- For example, “five nines” indicates a system that is up 99.999% of the time.

AWS Elastic Load Balancing

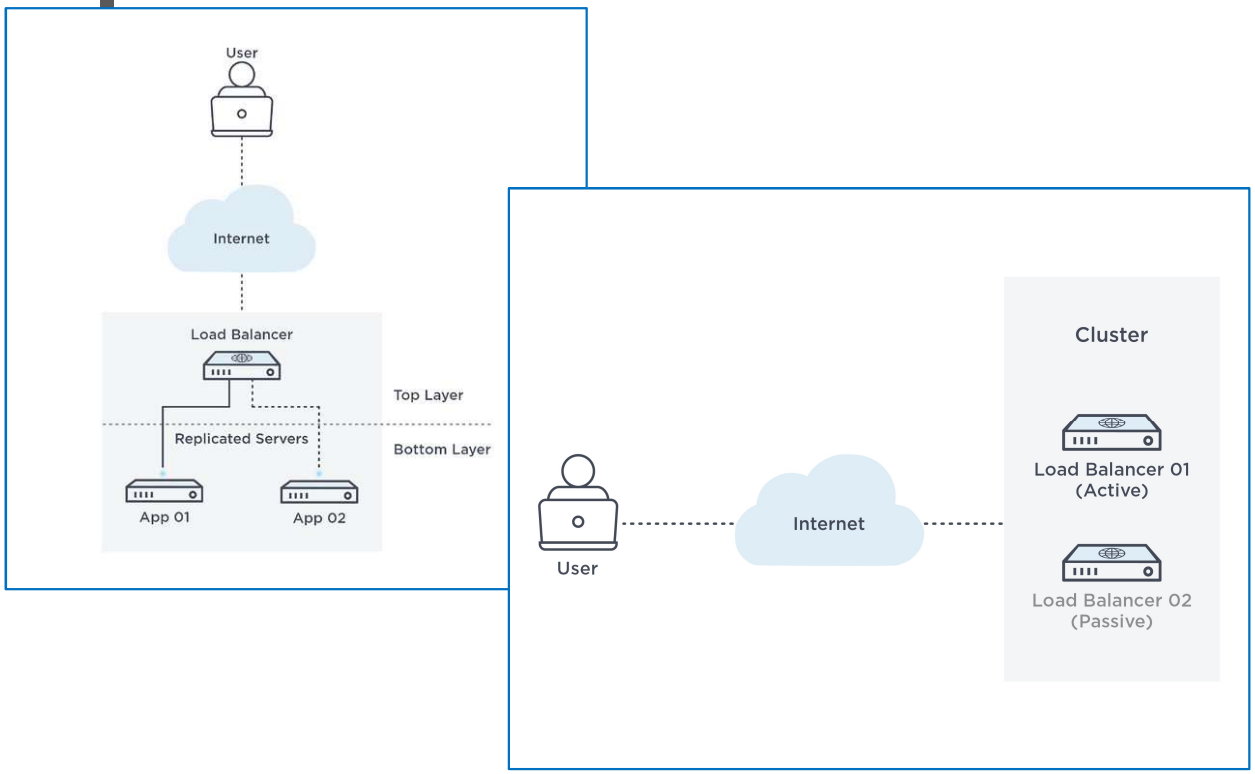
Availability(HA)



"AWS 공인 솔루션스 아키텍트 스터디 가이드-어소시에이트 3/e", 벤 파이퍼, 데이비드 클린턴 공저, 동준상 역, 에이콘 출판사, 2022, p406

AWS Elastic Load Balancing

Availability(HA)



- The basic elements of HA
  - Redundancy
  - Monitoring
  - Failover
- Technical components enabling HA
  - Data backup and recovery
  - Load balancing
  - Clustering

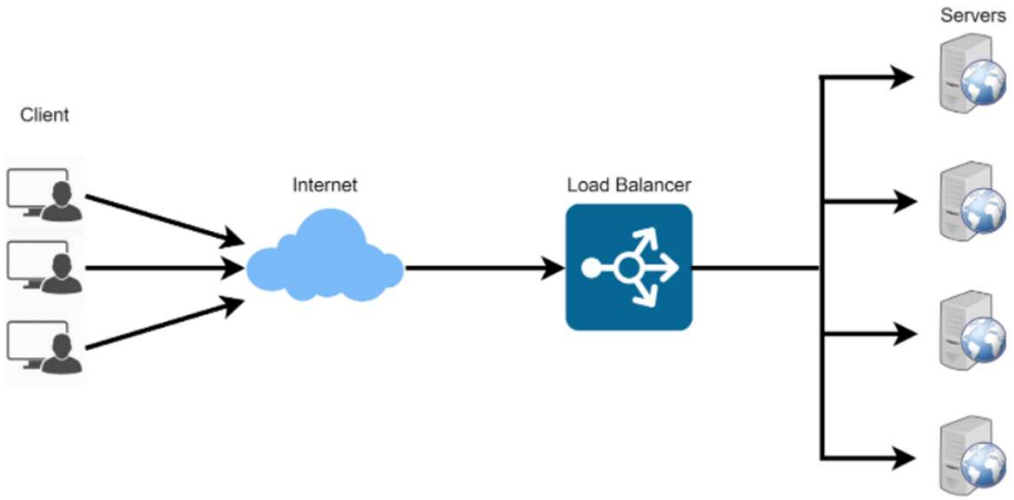
AWS Elastic Load Balancing

What's Load Balancing

- Automatically distributes incoming traffic across multiple targets.
- Multiple targets are EC2 instances, containers, and IP addresses, in one or more Availability Zones.
- Monitors the health of its registered targets.
- Routes traffic only to the healthy targets.
- Scales load balancer as incoming traffic changes over time.
- Can automatically scale to the vast majority of workloads.

AWS Elastic Load Balancing

Load Balancing Algorithms



https://medium.com/geekculture/load-balancing-da0bde7882f1

- Round Robin
- Hash
- Least Connection
- Response Time

AWS Elastic Load Balancing

ELB's Types

High availability percentages of SLAs

PERCENTAGE	YEARLY DOWNTIME*
99.9	8hr 45m 57s
99.99	52m 35.7s
99.999	5m 15.6s
99.9999	31.6s
99.99999	3.2s
99.999999	0.3s
99.9999999	31.6 ms

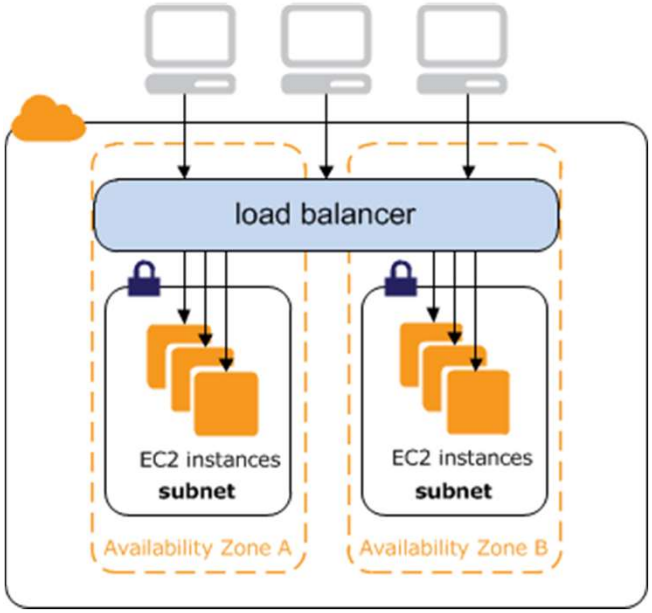
*APPROXIMATE VALUES; SOURCE: HTTPS://UPTIME.IS/

https://www.techtarget.com/searchdatacenter/definition/high-availability

- Application Load Balancer
- Network Load Balancer
- Gateway Load Balancer
- Classic Load Balancer

AWS Elastic Load Balancing

ELB Key Features



https://docs.aws.amazon.com/ko_kr/elasticloadbalancing/latest/classic/elb-internet-facing-load-balancers.html

- Security
- High availability
- High throughput
- Health checks
- Sticky sessions

Lab1. Create and Deploy Application Load Balancer

Lab2. Application Load Balancer Test on Failure

Lab3. Application Load Balancer Sticky Session Setting

AWS Auto Scaling

AWS Auto Scaling

Scalability

Amazon
Auto Scaling

https://www.offsetup.com/supported-technologies

- Involves beginning with only the resources need.
- Involves designing architecture to automatically respond to changing demand by scaling out or in.
- As a result, can pay for only the resources use.
- Don’t have to worry about a lack of computing capacity to meet customers’ needs.

AWS Auto Scaling

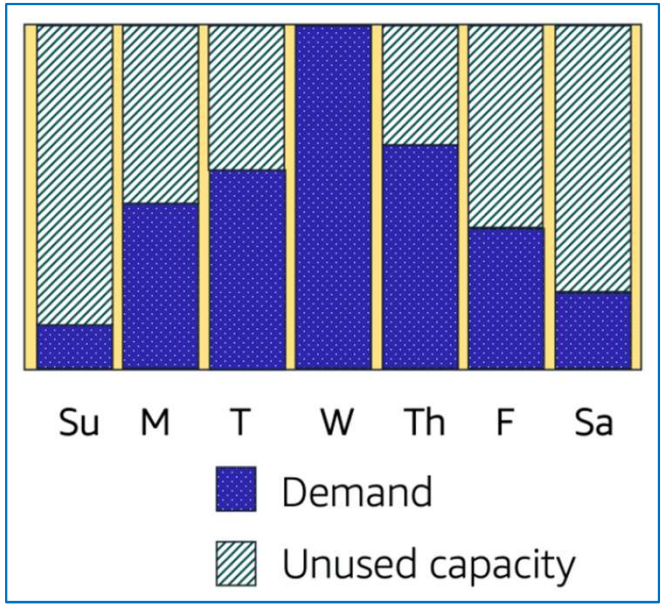
Scalability

https://digitalcloud.training/amazon-ec2-auto-scaling/

- If wanted the scaling process to happen automatically, which AWS service would use?
- The AWS service that provides this functionality for Amazon EC2 instances is ***Amazon EC2 Auto Scaling***.

AWS Auto Scaling

What's Auto Scaling

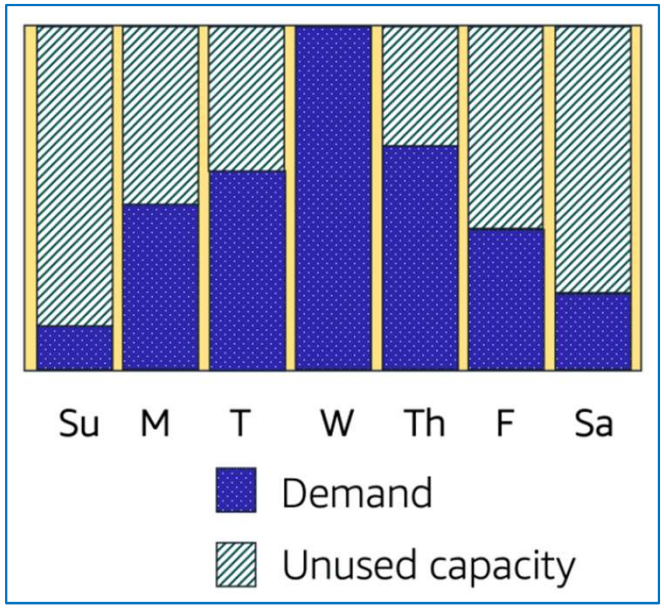


AWS Cloud Practioner Essentials

- Enables to automatically add or remove Amazon EC2 instances in response to changing application demand.
- By automatically scaling instances in and out as needed, be able to maintain a greater sense of application availability.

AWS Auto Scaling

What's Auto Scaling

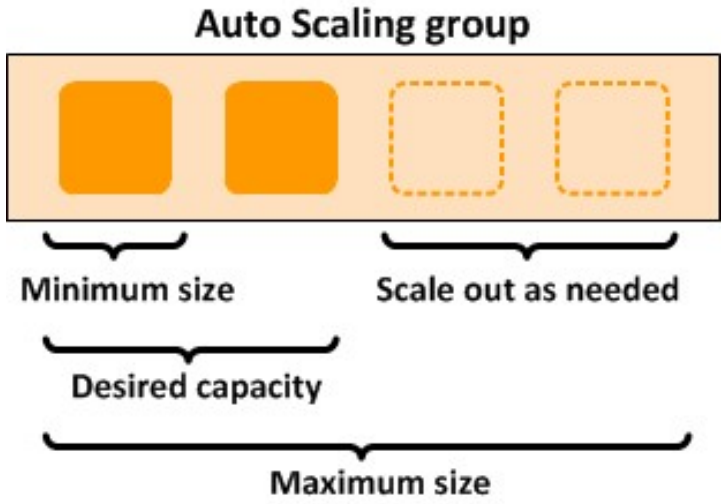


AWS Cloud Practioner Essentials

- Within Amazon EC2 Auto Scaling, can use two approaches: ***dynamic scaling*** and ***predictive scaling***.
- ***Dynamic scaling*** responds to changing demand.
- ***Predictive scaling*** automatically schedules the right number of Amazon EC2 instances based on predicted demand.

AWS Auto Scaling

What's Auto Scaling

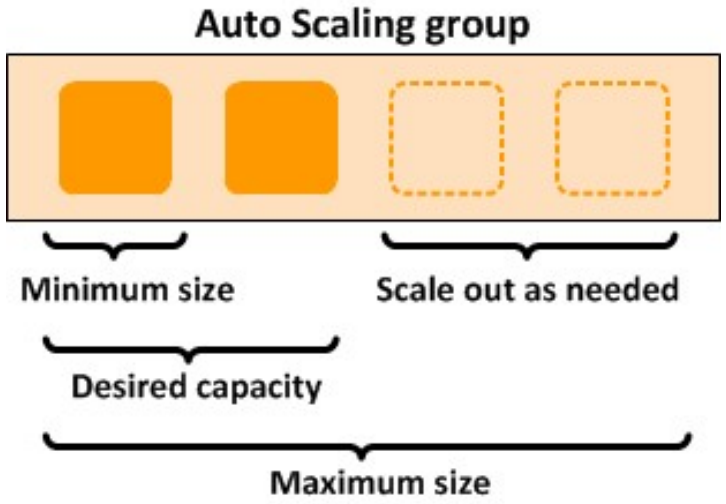


https://docs.aws.amazon.com/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html

- Ensure the correct number of Amazon EC2 instances available to handle the load for application.
- Create collections of EC2 instances, called ***Auto Scaling groups***.
- The minimum number of instances in each Auto Scaling group → Never goes below this size.
- The maximum number of instances in each Auto Scaling group → Never goes above this size.

AWS Auto Scaling

What's Auto Scaling

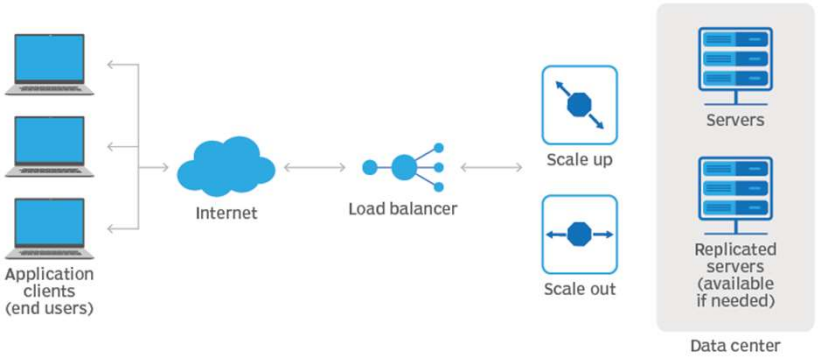


https://docs.aws.amazon.com/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html

- Specify the desired capacity, either when create the group or at any time thereafter → Amazon EC2 Auto Scaling ensures that group has this many instances.
- Specify scaling policies →Amazon EC2 Auto Scaling can launch or terminate instances as demand on application increases or decreases.

AWS Auto Scaling

AWS Auto Scaling Component



https://www.techtarget.com/searchcloudcomputing/tip/When-to-use-Amazon-EC2-Auto-Scaling-vs-AWS-Auto-Scaling

- Amazon Auto Scaling Group
- Configuration Templates
- Scaling Options

Lab4. Understanding Amazon EC2 Auto Scaling

Quiz

#1

Amazon EC2 예약 인스턴스의 약정 기간 옵션은 무엇인가? (2개 선택)

① 1년
② 2년
③ 3년
④ 4년
⑤ 5년

Quiz

#2

총 6개월 동안 실행되며 중단을 견딜 수 있는 Workload가 있다. 가장 비용 효율적인 Amazon EC2 구매 옵션은 무엇인가?

① 예약 인스턴스
② 스팟 인스턴스
③ 전용 인스턴스
④ 온디맨드 인스턴스

Quiz

#3

다음 중 가용 영역을 가장 잘 설명한 것은 무엇인가?

① AWS 리소스가 포함된 지리적 영역

② 리전 내의 단일 데이터 센터 또는 데이터 센터 그룹

③ AWS 서비스가 서비스별 작업을 수행하는 데 사용하는 데이터 센터

④ 온프레미스 데이터 센터에서 하이브리드 방식으로 AWS 인프라를 실행하는 데 사용할 수 있는 서비스

Quiz

#4

다음 중 AWS 글로벌 인프라에 대한 올바른 설명은 무엇인가?

① 리전은 단일 가용 영역으로 구성된다.
② 가용 영역은 두 개 이상의 리전으로 구성된다.
③ 리전은 두 개 이상의 가용 영역으로 구성된다.
④ 가용 영역은 단일 리전으로 구성된다.

Quiz

#5

리전을 선택할 때 고려해야 할 요소는 무엇인가? (2개 선택)

① 데이터 거버넌스 및 법적 요구 사항 준수
② 고객과의 근접성
③ 연중무휴 기술 지원 이용 가능
④ 다른 사용자에게 사용자 지정 권한을 할당하는 기능
⑤ AWS 명령줄 인터페이스(AWS CLI) 이용 가능