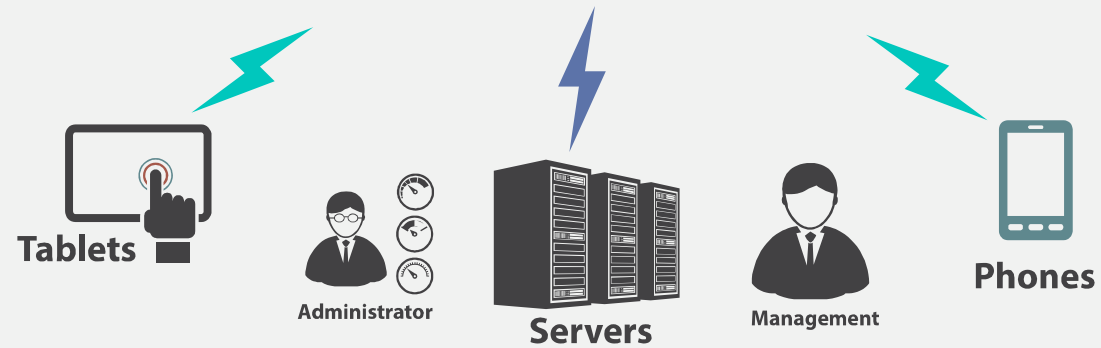


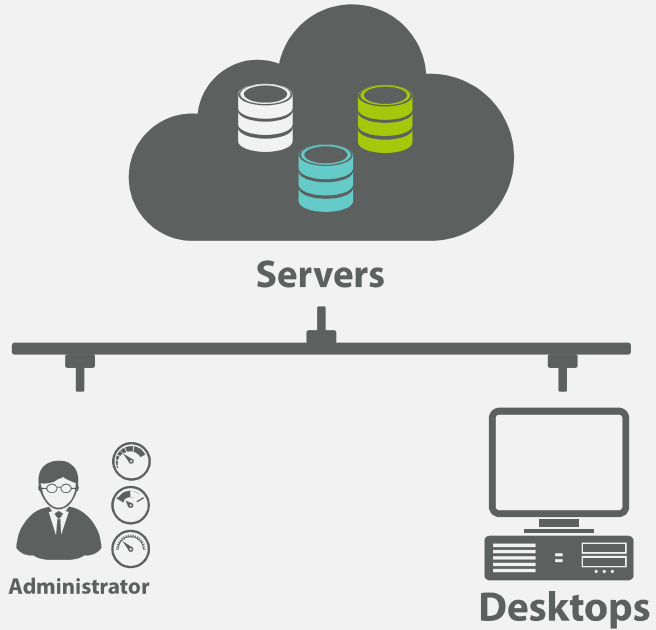


클라우드 아키텍처 구조

탄력적인 Resource 운영을 위한 Service







Index

01. 수업 목표

02. AWS ELB

03. AWS Auto Scaling

개요

The image shows two overlapping screenshots of the AWS Management Console. The top screenshot is for 'Elastic Load Balancing' and the bottom one is for 'AWS Auto Scaling'.

Elastic Load Balancing

- Navigation: 개요, 기능, 요금, 시작하기, FAQ, 파트너, 고객
- Header: « 네트워킹 및 콘텐츠 전송
- Section: Elastic Load Balancing
- Text: 네트워크 트래픽을 분산하여 애플리케이션 확장성 개선
- Button: Elastic Load Balancing 시작하기
- Callout: 매월 무료 750시간
Network Load Balancer 및 Application Load Balancer 간 제공 - AWS 프리 티어 사용 혜택

AWS Auto Scaling

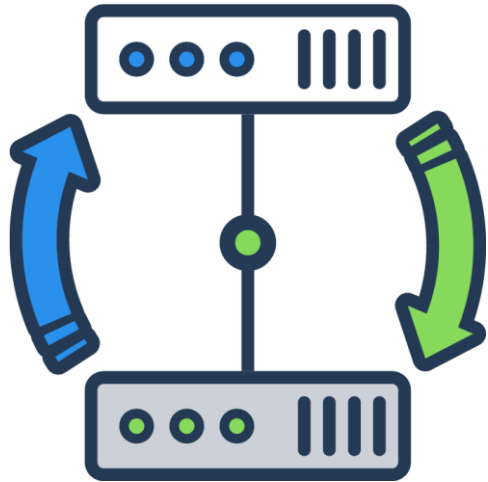
- Navigation: 개요, 기능, 요금, 리소스, FAQ
- Section: AWS Auto Scaling
- Text: 성능과 비용을 최적화하도록 애플리케이션 규모 조정
- Button: AWS Auto Scaling 시작하기
- Text: 통합된 인증서 관리, 사용자 인증, SSL/TLS 복호화를 통해 애플리케이션을 보안합니다.
- Section: 작동 방식
- Text: Elastic Load Balancing(ELB)은 하나 이상의 가용성 영역에 걸쳐 애플리케이션 트래픽을 분산합니다.
- Text: AWS Auto Scaling은 애플리케이션을 모니터링하고 용량을 자동으로 조정하여, 최대한 저렴한 비용으로 안정적이고 예측 가능한 성능을 유지합니다. AWS Auto Scaling을 사용하면 몇 분 만에 손쉽게 여러 서비스 전체에서 여러 리소스에 대해 애플리케이션 규모 조정을 설정할 수 있습니다. 이 서비스는 간단하면서도 강력한 사용자 인터페이스를 제공하므로 이를 사용하여 Amazon EC2 인스턴스와 스팟 플릿, Amazon ECS 작업, Amazon DynamoDB 테이블 및 인덱스, Amazon Aurora 복제본 등 리소스에 대한 규모 조정 계획을 수립할 수 있습니다. AWS Auto Scaling을 사용하면 성능과 비용을 최적화하거나 둘 사이의 적절한 균형을 유지하기 위한 권장 사항을 활용해 간단하게 규모를 조정할 수 있습니다. 이미 Amazon EC2 Auto Scaling을 사용하여 Amazon EC2 인스턴스의 규모를 동적으로 조정하고 있는 경우, 이제 AWS Auto Scaling과 결합하여 다른 AWS 서비스의 추가 리소스를 조정할 수 있습니다. AWS Auto Scaling을 사용하면 항상 적시에 올바른 리소스가 애플리케이션에 할당됩니다.
- Text: AWS Management Console, 명령줄 인터페이스(CLI) 또는 SDK를 사용하면 AWS Auto Scaling을 손쉽게 시작할 수 있습니다. AWS Auto Scaling은 추가 요금 없이 사용할 수 있습니다. 애플리케이션을 실행하는 데 필요한 AWS 리소스와 Amazon CloudWatch 모니터링 요금만 지불하면 됩니다.
- Section: Predictive Scaling 소개
- Text: ML 기술을 사용하여 트래픽이 변경되기 전에 컴퓨팅 용량을 자동으로 조정합니다.
- Link: 블로그 읽기>>
- Section: Amazon EC2 Auto Scaling 이전
- Text: AWS에서 규모를 조정할 수 있는 몇 가지 옵션이 있습니다. Amazon EC2 인스턴스만 관리하면 됩니까?
- Link: Amazon EC2 Auto Scaling 페이지로 이동하기 >

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

AWS ELB



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.



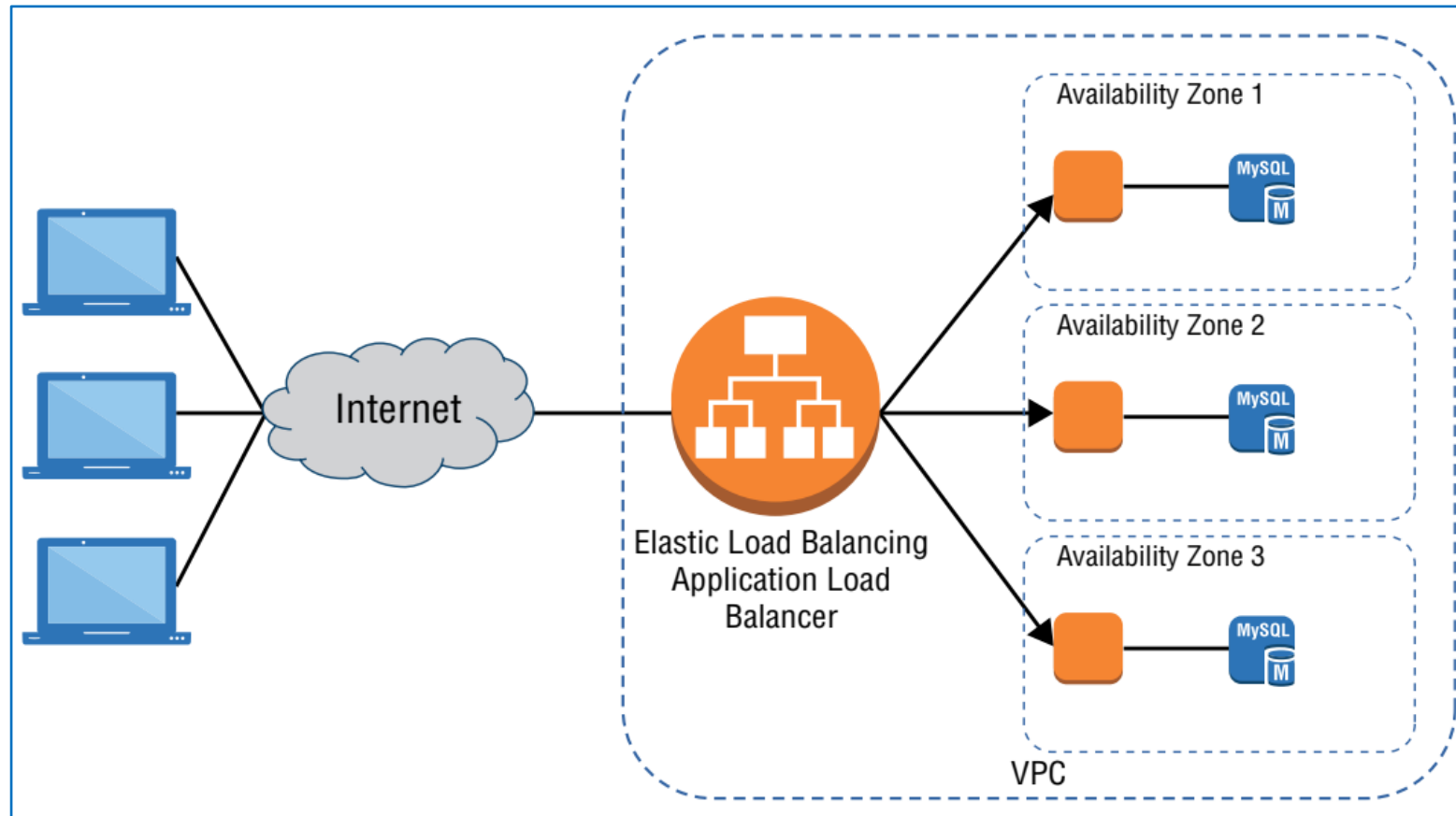
Availability(HA)

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

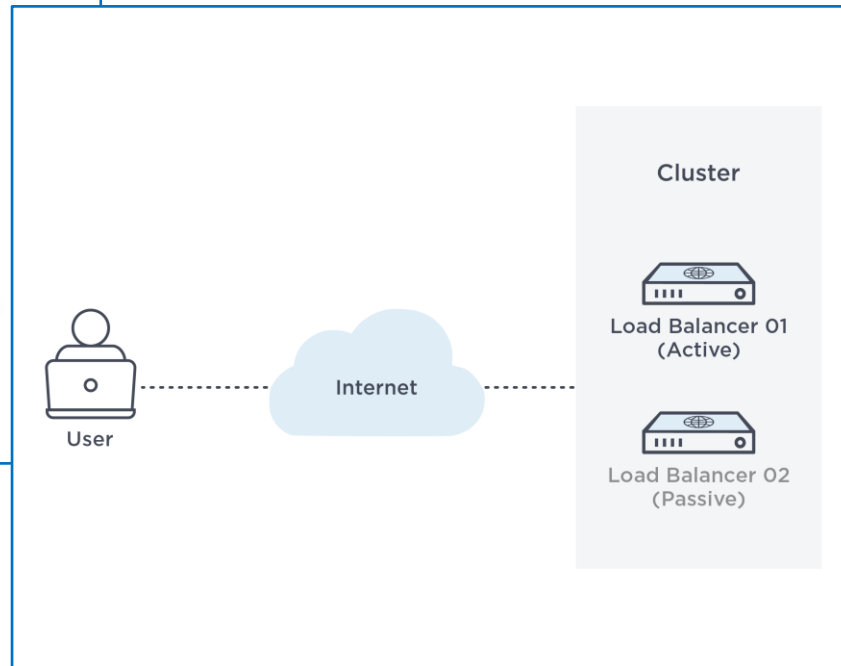
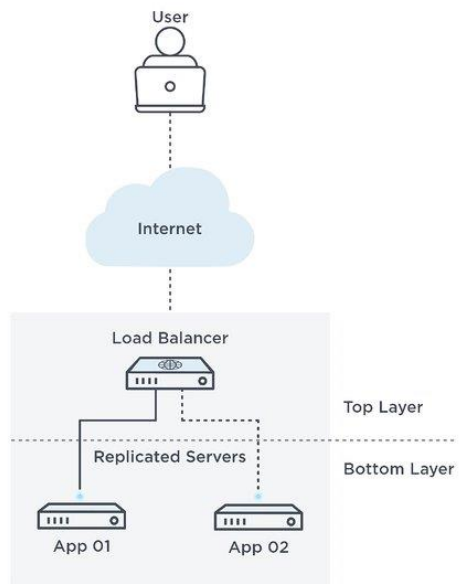
<https://www.nojitter.com/slas-burden-enterprise>

- 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.

Availability(HA)



Availability(HA)



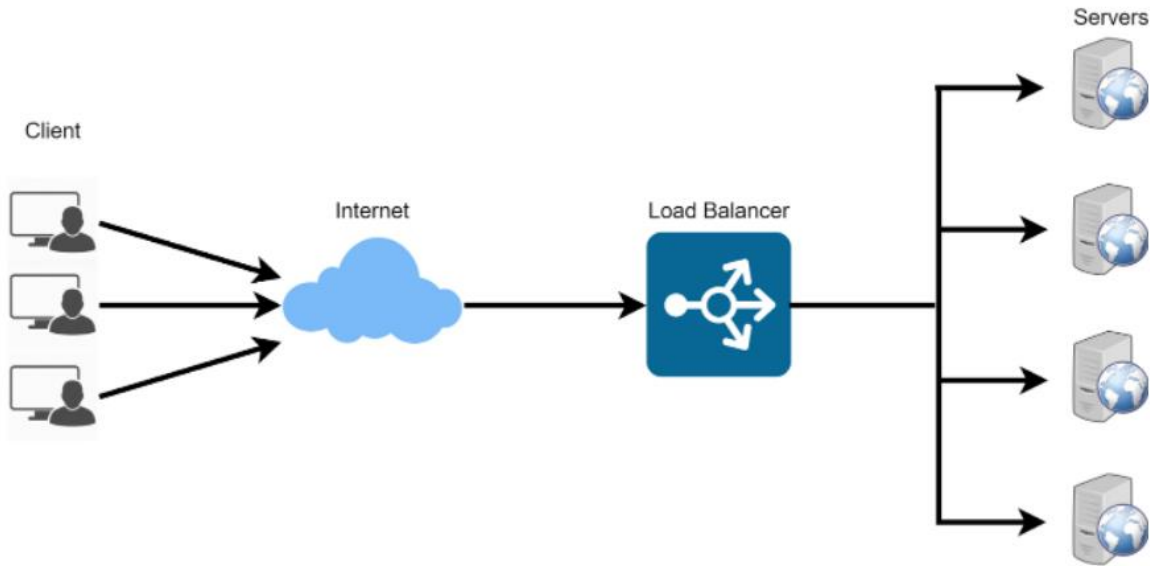
- The basic elements of HA
 - Redundancy
 - Monitoring
 - Failover
- Technical components enabling HA
 - Data backup and recovery
 - Load balancing
 - Clustering



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.

Load Balancing Algorithms



- Round Robin
- Hash
- Least Connection
- Response Time

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

ELB's Types

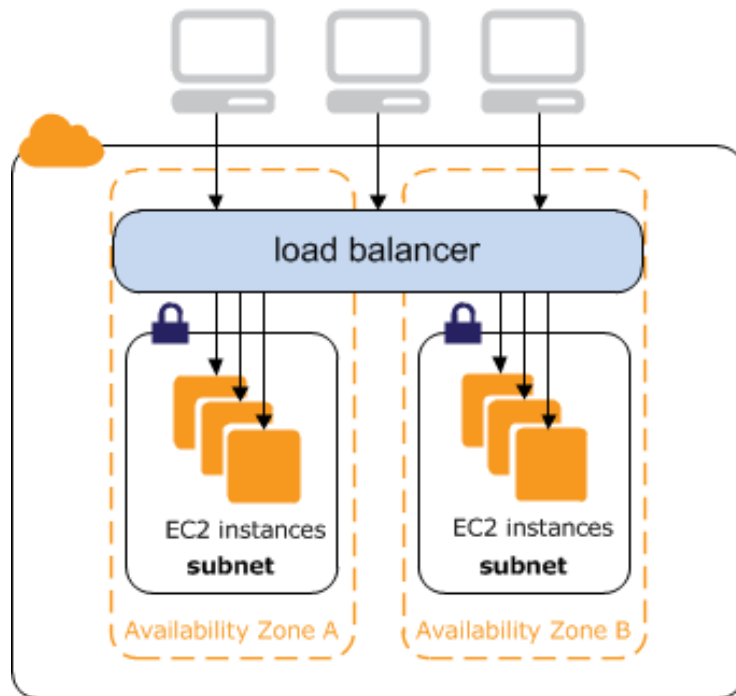
High availability percentages of SLAs	
PERCENTAGE	YEARLY DOWNTIME*
99.9	8hr 45m 57s
99.99	52m 35.7s
99.999	5m 15.6s
99.9999	31.6s
99.99999	3.2s
99.999999	0.3s
99.9999999	31.6 ms

*APPROXIMATE VALUES; SOURCE: [HTTPS://UPTIME.IS/](https://uptime.is/)
©2019 TECHTARGET. ALL RIGHTS RESERVED

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

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

ELB Key Features



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

<https://docs.aws.amazon.com/ko-kr/elasticloadbalancing/latest/classic/elb-internet-facing-load-balancers.html>



Lab1. Create and Deploy Application Load Balancer



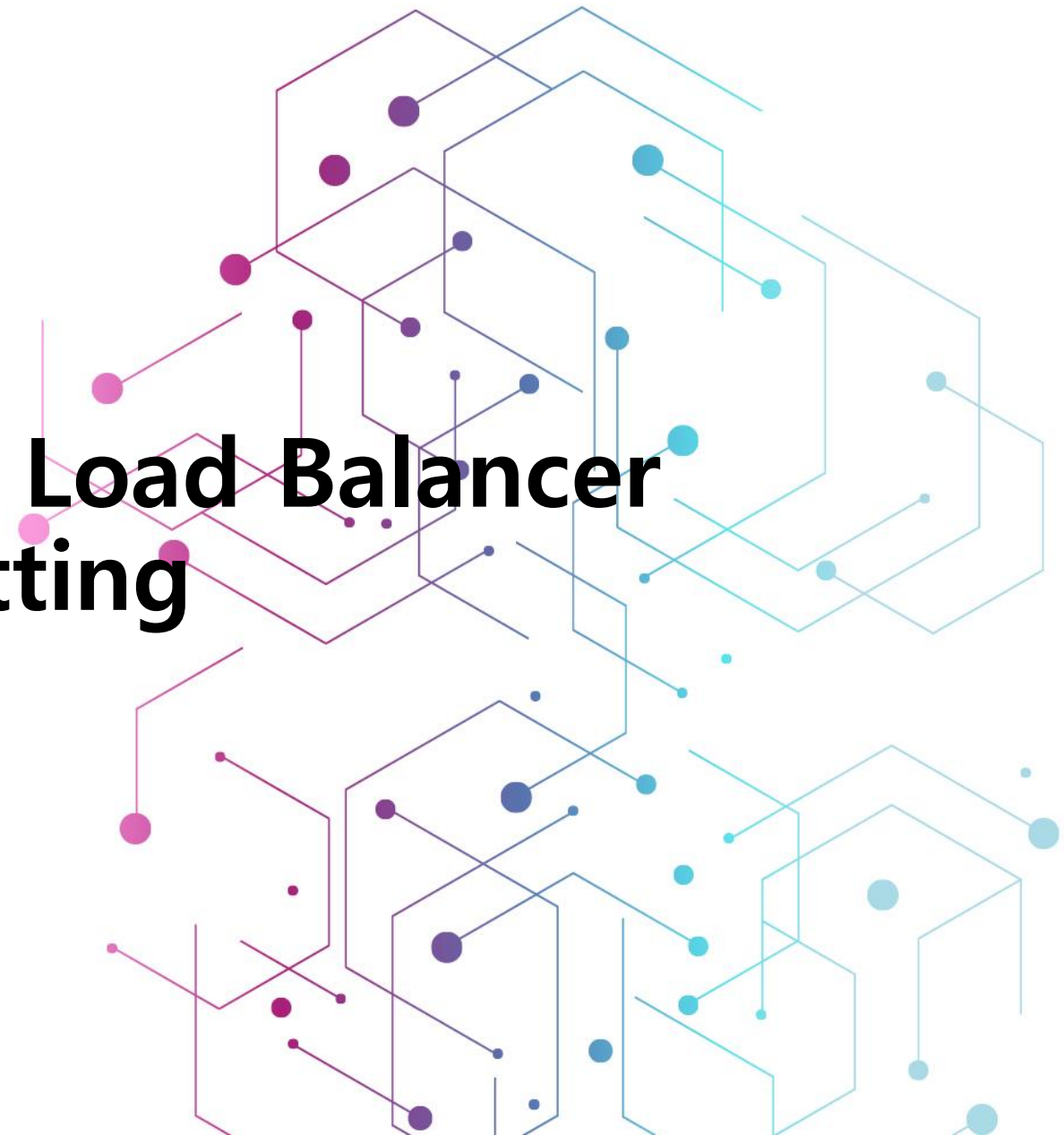


Lab2. Application Load Balancer Test on Failure





Lab3. Application Load Balancer Sticky Session Setting



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.

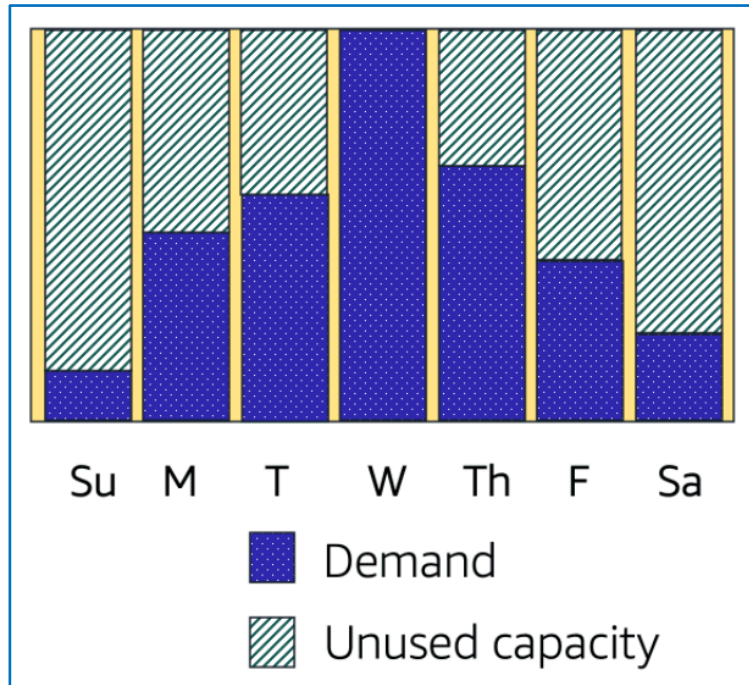
Scalability



<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*.

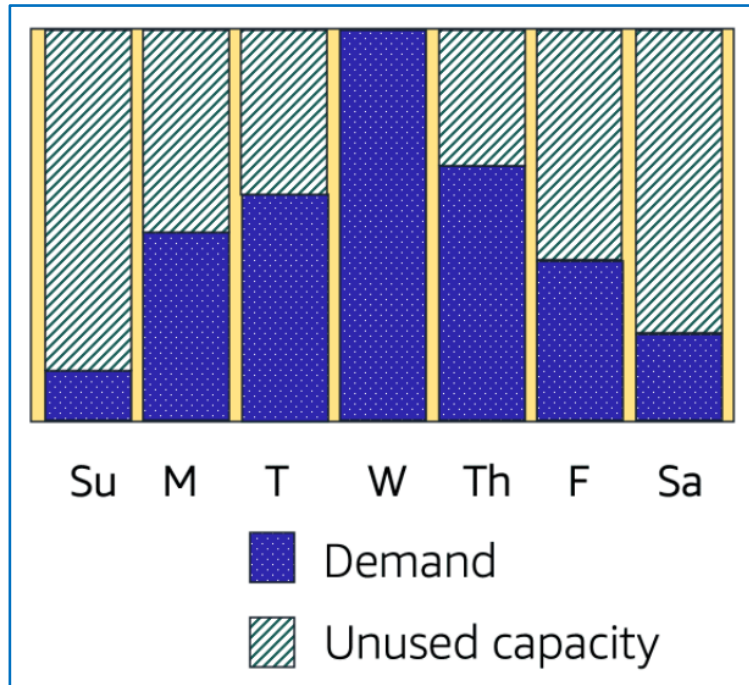
What's Auto Scaling



AWS Cloud Practitioner 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.

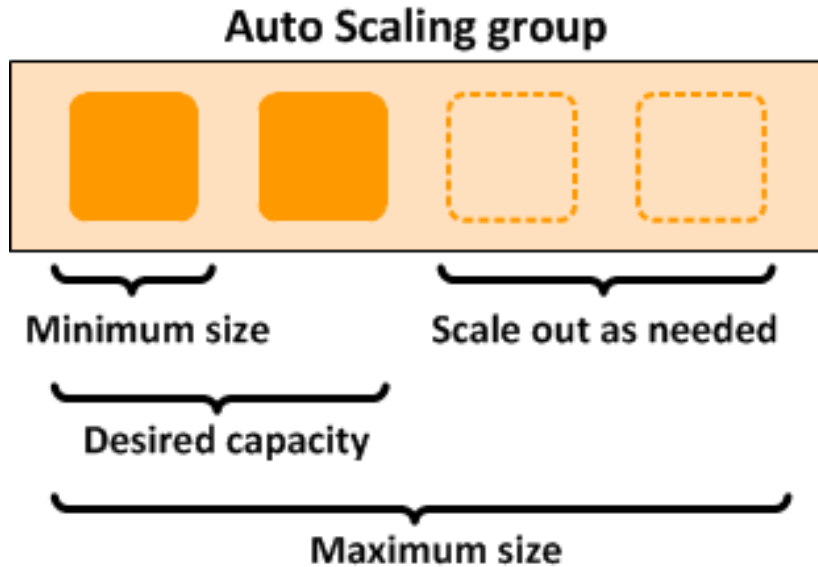
What's Auto Scaling



AWS Cloud Practitioner 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.

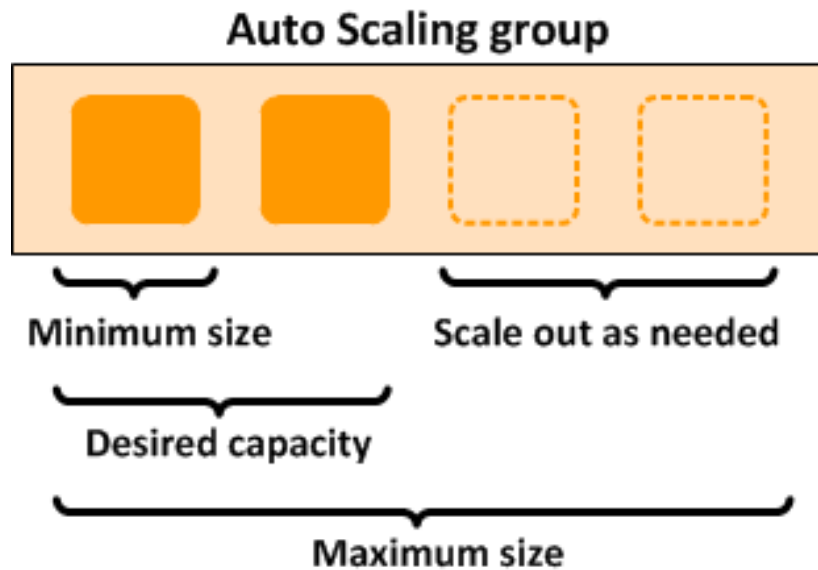
What's Auto Scaling



<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.

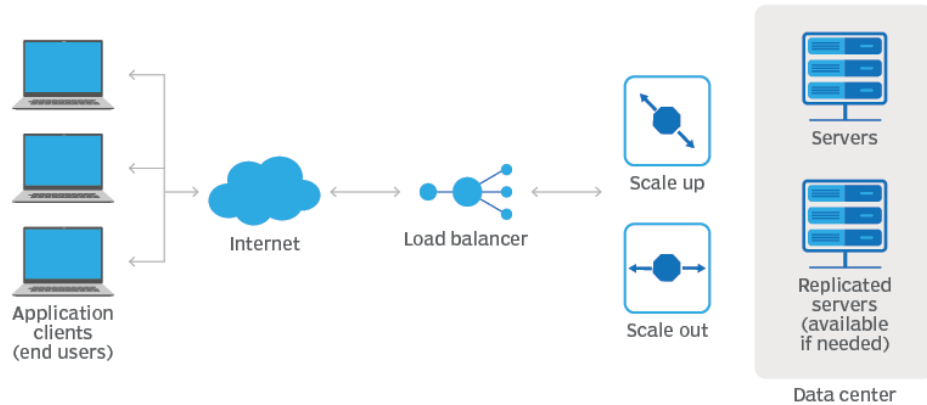
What's Auto Scaling



<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 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





| #1

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

- ① 1년
- ② 2년
- ③ 3년
- ④ 4년
- ⑤ 5년



| #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



| #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



| #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.



| #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