AWS 이론 5
- Auto Scaling

Sung-Dong Kim,

School of Computer Engineering,

Hansung University

가용성/확장성

- Availability
  - 시스템이나 서비스가 가동 및 실행되는 시간의 비율
- Scalability
  - 서비스나 응용 프로그램이 증가하는 성능 요구에 맞게 향상될 수 있는 정도
  - scale up, scale out

What

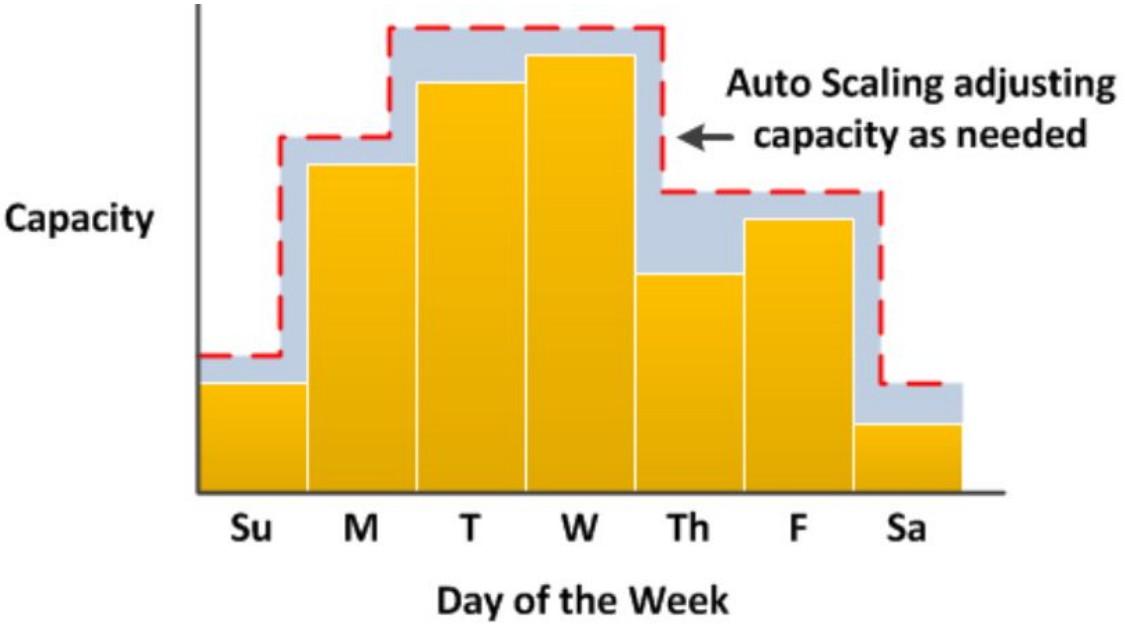
- Application의 로드를 처리할 수 있는 정확한 수의 EC2 instance를 보유하도록 보장
- Auto scaling group
  - EC2 instance 모음
  - 최대, 최소, 목표 instance 개수 지정
- Application의 가용성을 간편하게 관리: 조정 정책을 지정하여 application 수요에 따라 instance 시작/종료

이점 (1)

- 내결함성 향상: 비정상적 instance를 탐지하여 정상적 instance로 대체
- 가용성 향상: 트래픽을 처리할 수 있는 적절한 용량
- 비용 관리 개선: 동적으로 instance 시작/종료

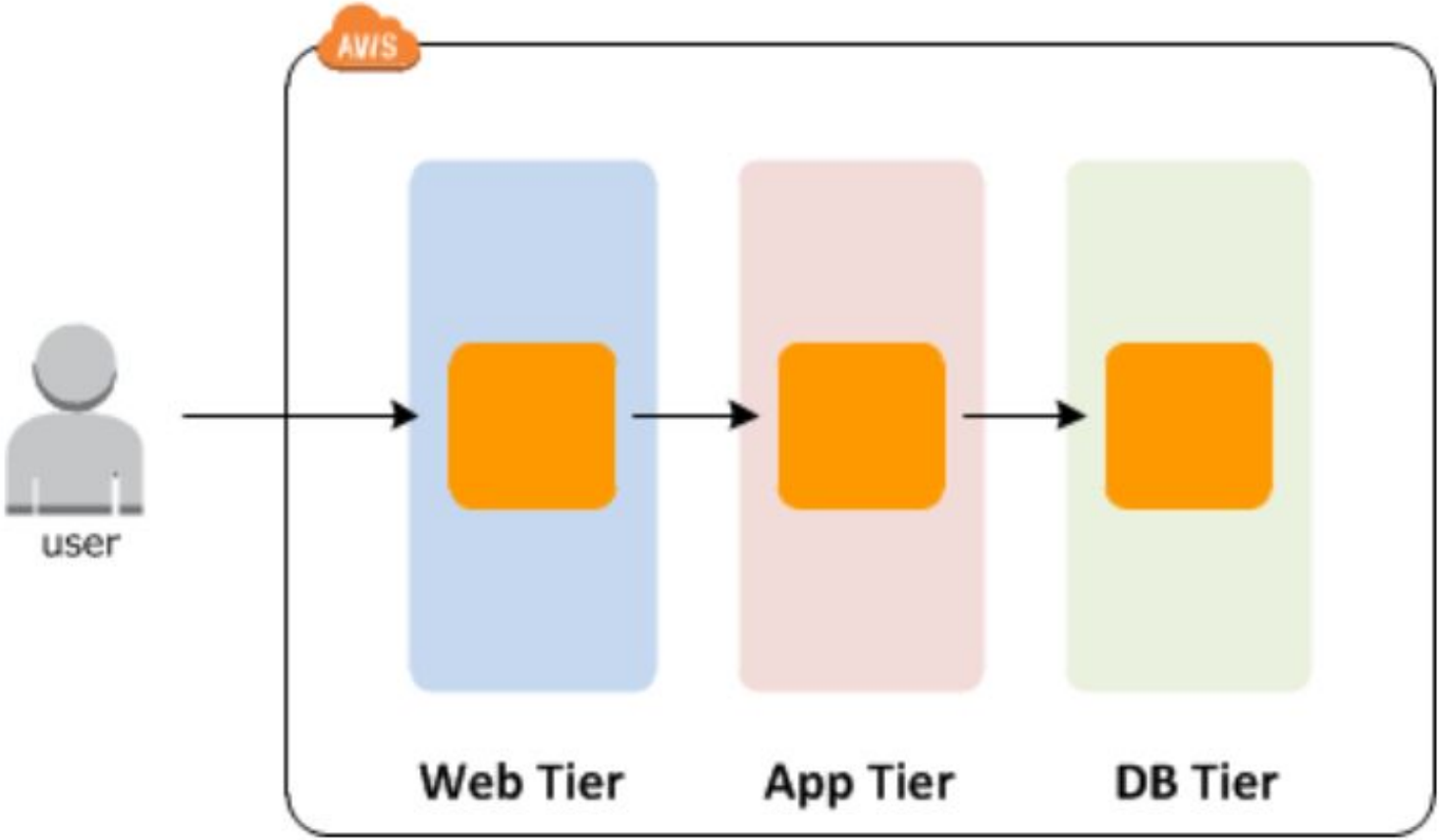
이점 (2)

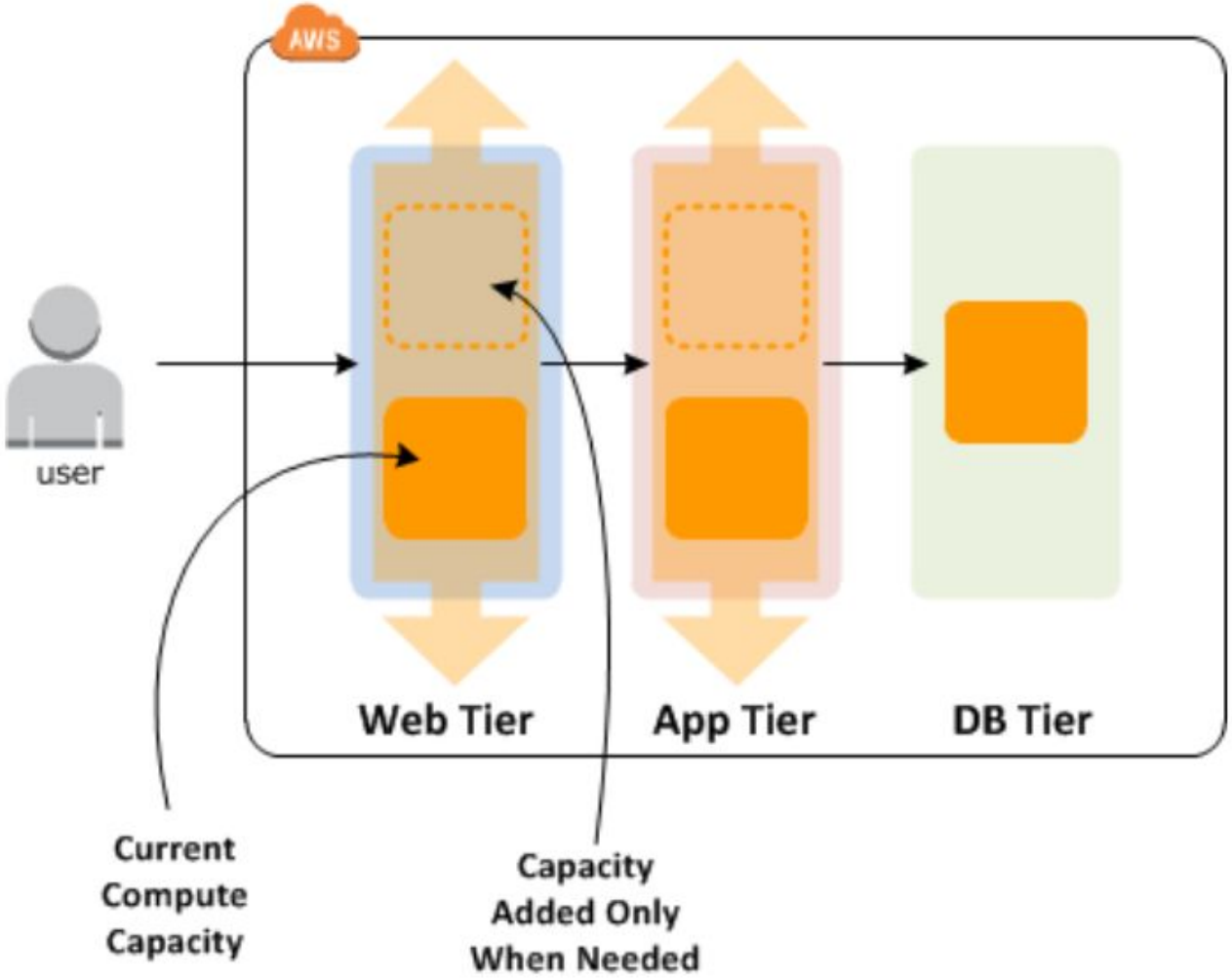
- 가변 수요에 대처


Auto Scaling adjusting
capacity as needed
Capacity
Su
M
T
W
Th
F
Sa
Day of the Week

이점 (3)

- 웹 앱 아키텍처
  - 고객의 트랙픽을 처리하기 위해 여러 개의 app 사본을 하나의 EC2 instance (cloud server)에서 호스팅, 각각에서 고객 요청이 처리됨
  - Auto scaling은 EC2 instane 시작/종료를 관리
  - 시작/종료 시기를 결정하는 조건(CloudWatch 경보 등) 집합 정의
  - application의 가용성, 내결함성 향상
  - ELB: 인스턴스 간 트래픽 분산


AWS
user
Web Tier
App Tier
DB Tier

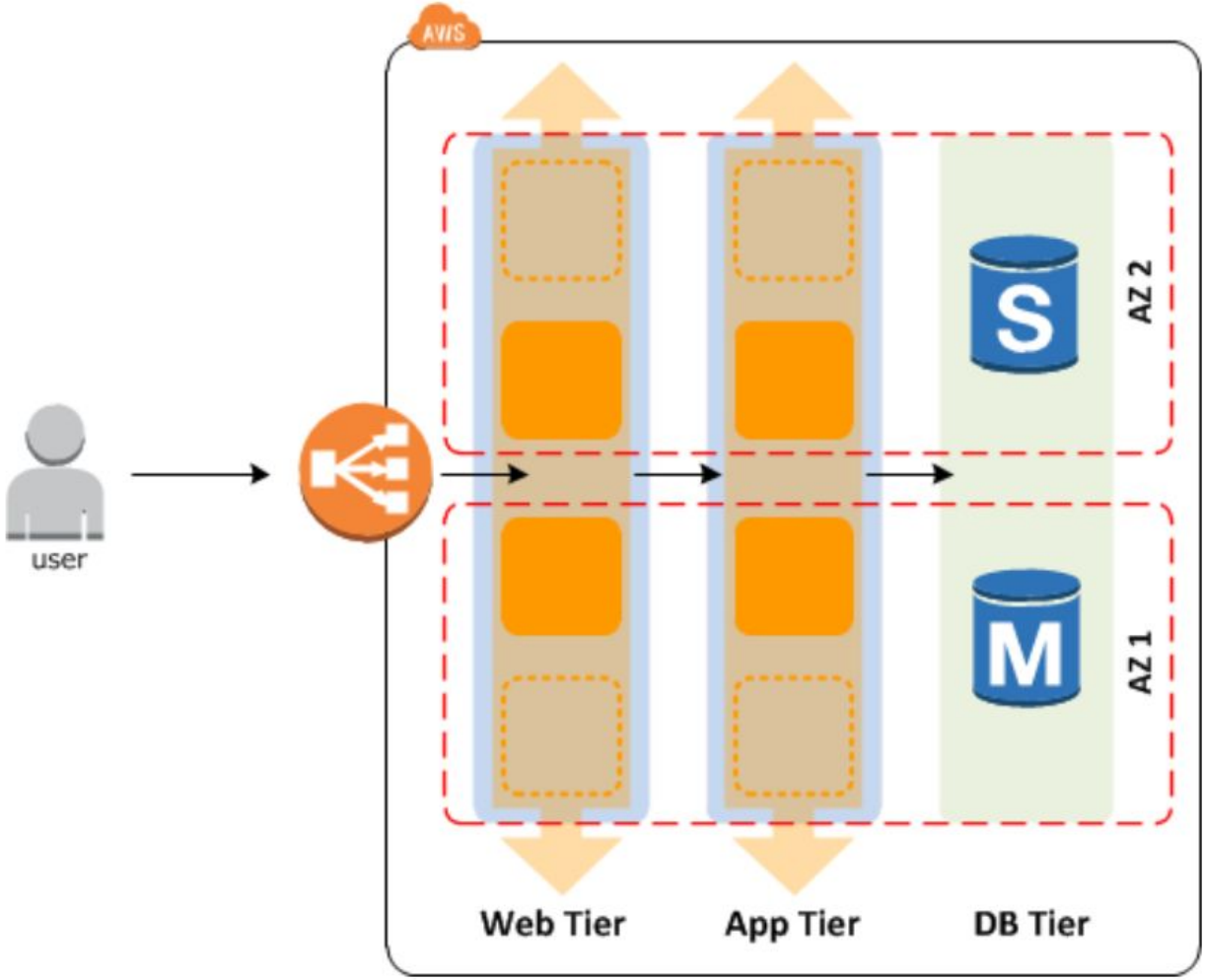

AWS
user
Web Tier
App Tier
DB Tier
Current
Compute
Capacity
Capacity
Added Only
When Needed

이점 (4)

- 가용 영역 전반에 instance 분산
  - region내 여러 AZ에 걸쳐 auto scaling 그룹 확장 → 지리적 이중화를 통한 보안 및 안정성 확보
  - instance를 AZ 간에 고르게 분산하려고 시도
  - VPC 내의 auto scaling group
    - subnet에서 EC2 instance가 시작
    - 한 AZ에 여러 subnet이 있으면 subnet을 무작위로 선택하여 시작

이점 (4)

- 가용 영역 전반에 instance 분산
  - 재분배 활동
    - AZ 간에 불균형시, 재분배 작업 수행
    - 이전 instance 종료 전, 새 instance 시작

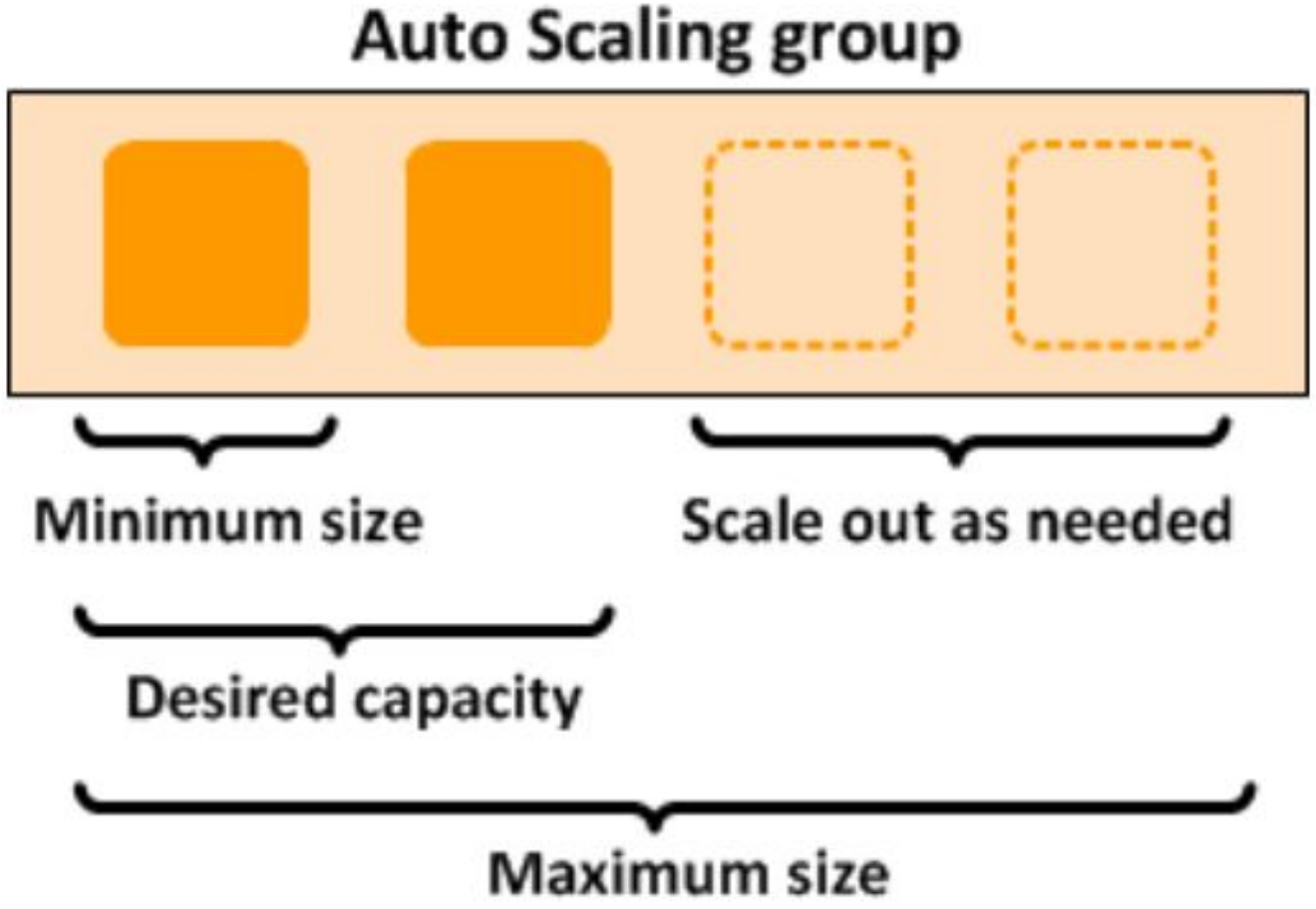

AWS
user
S
M
AZ 2
AZ 1
Web Tier
App Tier
DB Tier

구성 요소 (1)

- Amazon Auto Scaling Group
- 구성 템플릿 (configuration template)
- 조정 옵션

구성 요소 (2)

- Amazon Auto Scaling Group
  - 논리적 단위로 처리되는 EC2 인스턴스의 모음
  - 인스턴스의 조정 및 관리 목적으로 구성된 논리적 그룹
  - 인스턴스의 수를 조건에 따라 자동 조정 및 관리하는 auto scaling의 핵심 기능
  - 그룹의 최소, 최대 및 원하는 용량 정의


Auto Scaling group
Minimum size
Scale out as needed
Desired capacity
Maximum size

구성 요소 (3)

- 구성 템플릿 - 시작 구성
  - 인스턴스를 시작하는 데 사용하는 템플릿
  - auto scaling group이 생성되기 위해서 EC2 instance를 어떻게 만들 것인가를 설정
  - AMI, instance type, key pair, 보안 그룹, EBS 등 인스턴스에 대한 정보를 지정

구성 요소 (4)

- 조정 옵션 - auto scaling group 조정
  - 인스턴스의 수를 늘리거나 줄이는 기능
  - 이벤트와 함께 시작 또는 auto scaling group의 조정 작업과 함께 시작
  - 조정 옵션: 그룹 조정 방법/정책
    - 현재 인스턴스 수준 유지
    - 수동 조정 / 일정 기반 조정
    - 온디맨드 기반 조정 - 사용률에 따른 조정

수명 주기 (1)

- auto scaling group이 instance를 시작하고 서비스에 들어갈 때, 수명 시작
- instance 종료, auto scaling group에서 instance를 제외시키고 종료할 때, 수명 종료

수명 주기 (2)

- 확장
  - EC2 instance를 시작하고, 그룹에 연결하라고 지시
  - 수동 조정, 조정 정책에 따른 동적 조정, 특정 시간에 조정하는 예약/일정 조정
- 축소
  - 수동 조정, 조정 정책에 따른 동적 조정, 특정 시간에 조정하는 예약/일정 조정

관련 서비스

- Amazon EC2 - 클라우드에서 가상 머신 생성/실행
- Amazon CloudWatch
  - 조정 정책 활성화
  - Auto Scaling 그룹과 EC2 인스턴스에 대한 지표 모니터링
- Elastic Load Balancing
  - 수신되는 application traffic을 AS group의 instance에 자동으로 분산시킴

AWS auto scaling

- 인프라의 증설/축소를 손쉽게 구현, 확장성 및 탄력성 높은 시스템을 구축할 수 있음
- 서버나 애플리케이션을 모니터링하고 리소스를 자동으로 조정 (scale in/out)
- 최대한 저렴한 비용으로 안정적이고 예측 가능한 성능 유지

EC2 auto scaling의 동적 조정

- 애플리케이션 수요 곡선에 따름
- 애플리케이션 로드 지표 선택
- 조건부 또는 일정 예약으로 설정
- CloudWatch로 사용 (선택 사양)

EC2 auto scaling을 사용한 플릿 관리

- 중단 없이 손상된 EC2 instance 교체
- 실행중인 인스턴스의 상태 모니터링
- 손상된 인스턴스 자동 교체
- 여러 AZ에서 용량 밸런싱

Source

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