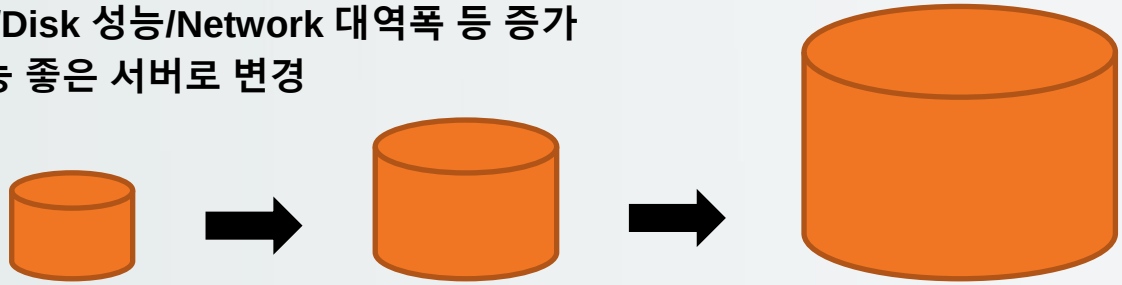
EC2 인스턴스 자동 확장(Auto Scaling)

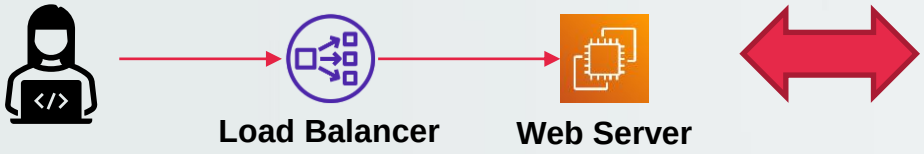
- 웹 트래픽 증가에 따른 대응 방법
  - 1. Scale Up
    - CPU/RAM/Disk 성능/Network 대역폭 등 증가
    - 비싸고 성능 좋은 서버로 변경

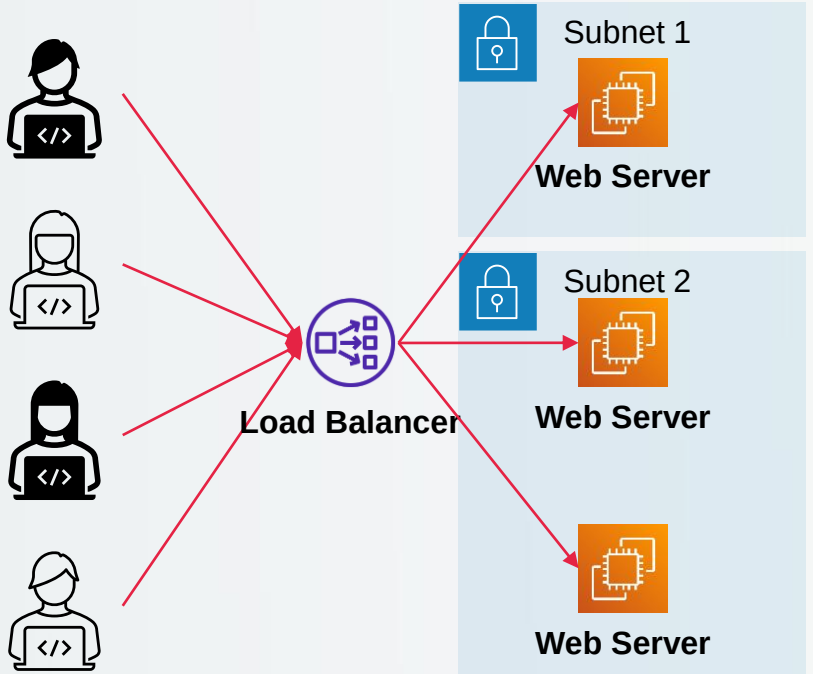


  - 2. Scale Out → 자동 확장(Auto Scaling)
    - 부하를 처리할 서버 대수를 늘림
    - 저렴한 서버 여러 대를 이용해 더 많은 부하를 감당

EC2 인스턴스 자동 확장(Auto Scaling)

- 자동확장 동작 방식
  - 트래픽이 적을 때는 적은 수의 서버를 유지
  - 트래픽이 많을 때는 서버를 실행시켜 안정성 확보
  - 비용 최적화 가능





트래픽이 적을 때

트래픽이 많을 때

EC2 인스턴스 자동 확장(Auto Scaling)

- **자동 확장 그룹(Auto Scaling Group)**
  - 확장/축소에 사용될 EC2 인스턴스의 논리적 집합
  - 자동 확장에 의해 늘어나거나 줄어든 대상들의 집합을 의미

최소 용량 : 항상 유지해야 하는 인스턴스 개수

필요 크기 : 평상시 유지하는 인스턴스 개수
최소 용량보다는 크거나 같아야 함

최대 크기 : 확장했을 때 늘어날 수 있는 인스턴스 개수

EC2 인스턴스 자동 확장(Auto Scaling)

- 자동 확장 그룹(Auto Scaling Group)
  - 인스턴스 개수 조정 옵션
    - 1. 항상 지정된 수의 인스턴스 유지
      - 자동 확장 그룹에서 지정한 인스턴스 개수를 유지
    - 2. 수동 조정
      - 사용자가 직접 그룹의 인스턴스 개수를 조정
    - 3. 일정 기반 조정
      - 특정 시간대에 인스턴스를 확장하고, 그 시간대가 지나면 다시 인스턴스를 축소
    - 4. 온디맨드 기반 조정
      - EC2 지표(Metric)을 기반으로 조정
      - 예시) CPU 사용률이 80% 이상일 때, EC2 인스턴스를 늘림