로드 밸런싱

> 둘 이상의 컴퓨터 자원들에게 자원을 나누는 것

트래픽 증가 핸들링

1. Scale - up : 하드웨어의 설능을 올린다.
2. Scale - out: 여러대의 서버가 나눠서 일하도록 한다.

비용이 저렴하고, 무중단 서비스를 제공하는 환경 구성이 용이하므로 Scale -out이 효과적이다.

여러 **서버에게 균등하게 트래픽을 분산시켜주는 것을 로드 밸런싱**이라고 한다.

로드 밸런서

클라이언트와 서버에서, 트래픽을 여러 서버에 분산시켜 준다.

서버를 증설하면서 로드 밸런서로 관리해주면 웹 서버의 부하를 해결할 수 있다.

로드 밸런서의 동작

1. 라운드 로빈: 클라이언트로부터 받은 요청을 로드밸런싱 대상 서버에 순서대로 할당하는 방식. 서버의 성능이 동일하고 처리 시간이 짧은 어플리케이션의 경우 사용한다.
2. 가중 라운드 로빈: 서버마다 가중치를 매기고, 가중치가 높은 서버에 더 많은 비율로 요청을 할당한다. 서버마다 트래픽 처리 능력이 다른 경우 사용하기 적합하다.
3. IP Hash방식: 사용자 IP를 해싱해서 분배한다. (특정 사용자가 항상 같은 서버로 연결되는 것을 보장한다)
4. 최소 연결 방식(Least Connections): 연결 개수가 가장 적은 서버 선버를 선택해 요청을 할당한다.
5. 최소 응답시간 (Least Response Time) : 현재 연결 상태, 응답 시간을 고려하여 요청을 할당한다. (가장 연결 개수가 적고 평균 응답시간이 가장 적은 서버에게 할당)

로드 밸런서 장애 대비

서버를 분배하는 로드 밸런서에 문제가 생길 수 있기 때문에, 로드 밸런서를 이중화하여 대비한다.