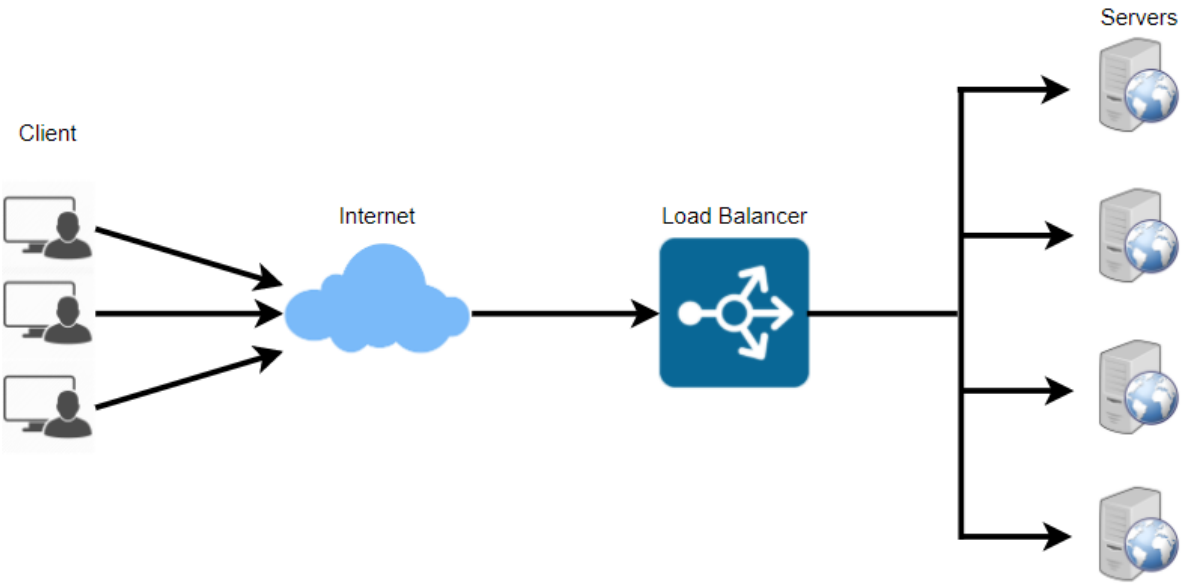
로드밸런싱



둘 이상의 CPU 나 저장장치와 같은 컴퓨터 자원들에게 작업을 나누는 것

웹사이트에 접속하는 인원이 급격하게 증가함에 따라, 이 모든 인원들에 대해 트래픽을 감당하기엔 1대의 서버로는 부족해짐

대응 방안으로 하드웨어의 성능을 올리거나(Scale-up),

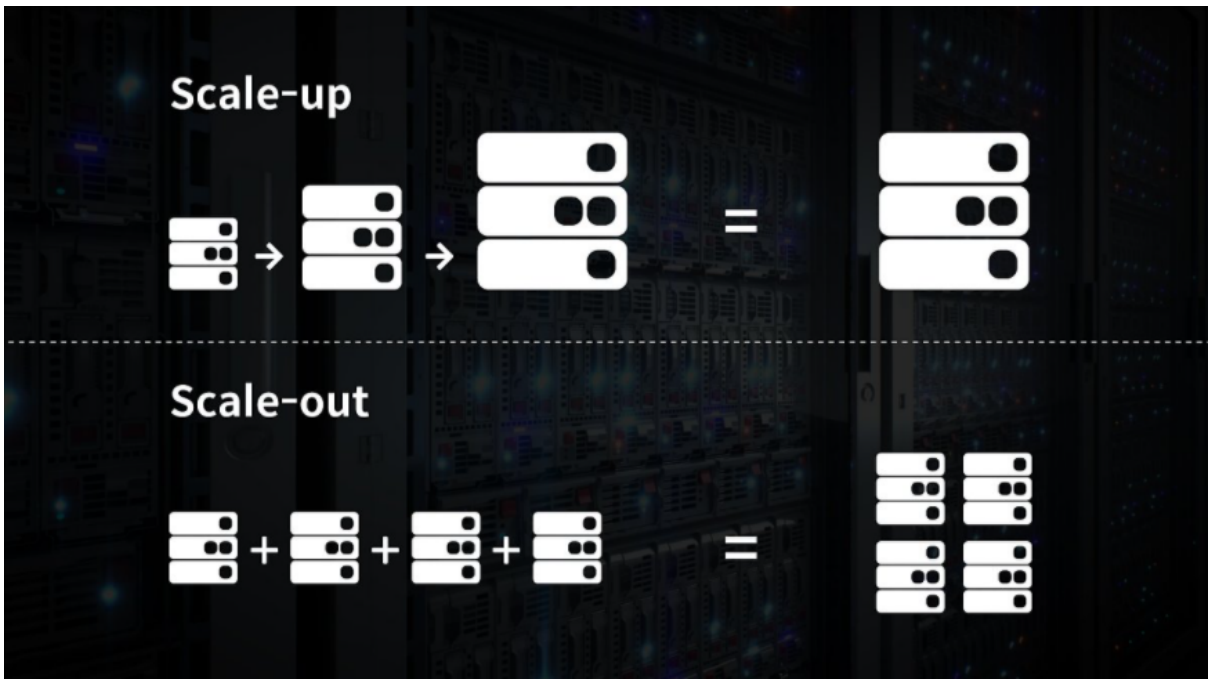
여러대의 서버가 나눠서 이를 감당하도록 하는 것 (Scale-out) 이 있음

하드웨어 향상 비용이나, 무중단 서비스 제공 환경 구성에 대한 용이한 **Scale-out** 이 효과적임

Scale-out 의 방식으로 서버 증설을 할 경우 로드밸런싱이 필요함

여러개의 서버에게 균등하게 트래픽을 분산시켜주는 것이 바로 **로드밸런싱**

Scale-up 과 Scale-out



- Scale-up : 기존 서버의 사양을 업그레이드해 시스템을 확장하는 것
- Scale-out : 서버를 여러대 추가하여 시스템을 확장하는 것

로드 밸런싱

분산식 웹 서비스로, 여러 서비스에 부하를 나눠주는 역할을 수행함

로드 밸런서(Load Balancer)를 클라이언트와 서버 사이에 두고, 부하가 일어나지 않도록 여러 서버에 분산시켜주는 방식

서비스를 운영하는 사이트의 규모에 따라 웹 서버를 추가로 증설하면서 로드 밸런서로 관리해주면 웹 서버의 부하를 해결할 수 있음

로드 밸런서가 서버를 선택하는 방식

- 라운드 로빈(Round Robin) : CPU 스케줄링의 라운드 로빈 방식 활용
    - 프로세스들 사이에 우선순위를 두지 않고, 순서대로 시간 단위로 할당

- Least Connections : 연결 개수가 가장 적은 서버를 선택 (트래픽으로 인해 세션이 길어지는 경우 권장)
- Source : 사용자 IP 를 해싱하여 분배 (특정 사용자가 항상 같은 서버로 연결되는 것을 보장)

로드 밸런서의 종류

부하 분산에는 L4 로드밸런서와 L7 로드밸런서가 가장 많이 활용됨

L4 로드밸런서부터 포트 정보를 바탕으로 로드를 분산하는 것이 가능하기 때문

한 대의 서버에 각기 다른 포트 번호를 부여하여 **다수의 서버 프로그램을 운영하는 경우**, **최소 L4** 로드밸런서나 그 이상의 로드밸런서를 사용해야함

- **L4? L7?**

  네트워크 통신 시스템은 개방형 통신을 위한 국제 표준 모델인 OSI 7 Layer 를 사용함

  각각의 계층이 L1, L2, …, L7 에 해당함

  상위 계층에서 사용되는 장비는 하위 계층의 장비가 갖고 있는 기능을 모두 가지고 있으며, 상위 계층으로 갈 수록 더욱 정교한 로드밸런싱이 가능함

  AWS 로드 밸런서 종류

  - **클래식 로드 밸런서(ELB) - L4**

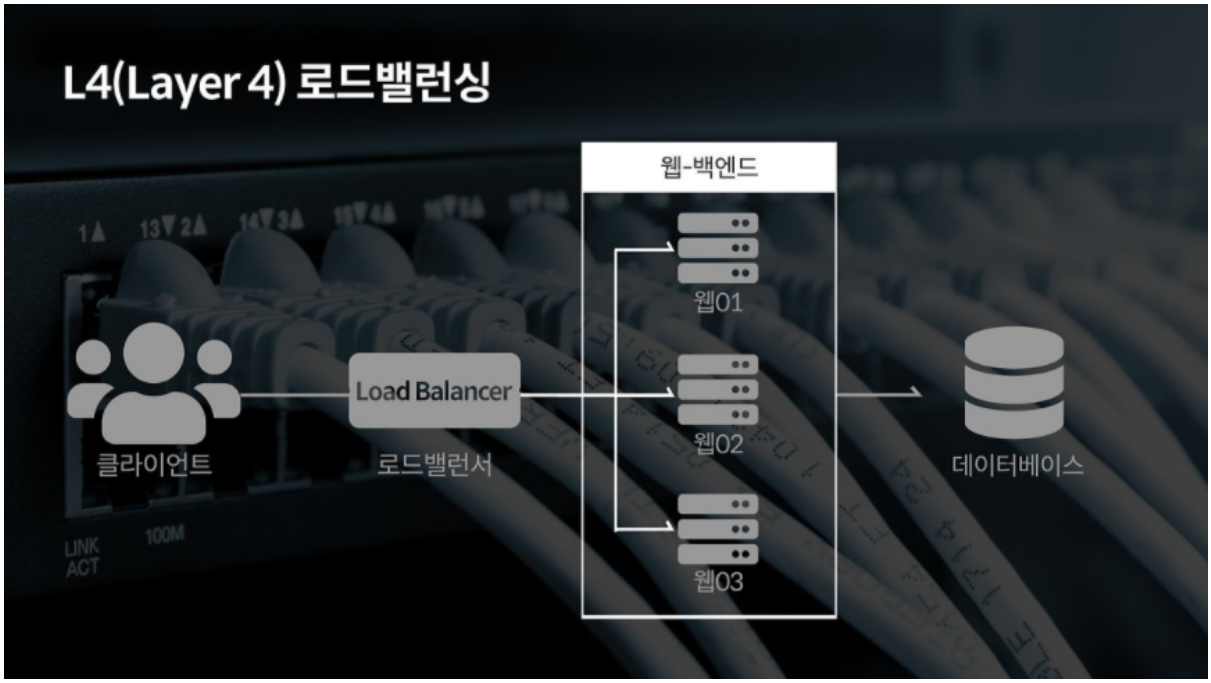
    라우터 스위치 등 물리적인 하드웨어 영역으로, 데이터를 변경하거나 수정할 수 없음

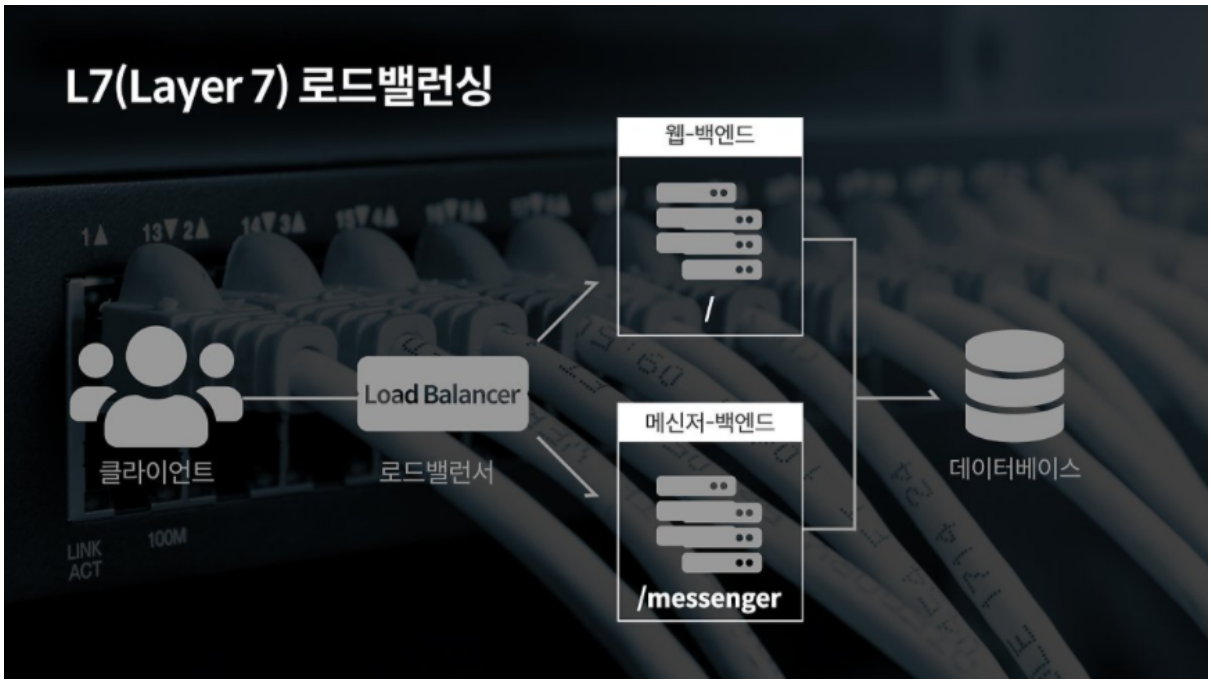
    단점 : 서버의 기본 주소가 바뀌면 로드 밸런서를 새로 생성

  - **애플리케이션 로드 밸런서(ALB) - L7**

    포트나 헤더 등의 수정이 가능함

L4 로드밸런싱과 L7 로드밸런싱


L4(Layer 4) 로드밸런싱
웹-백엔드
웹01
웹02
웹03
Load Balancer
로드밸런서
클라이언트
데이터베이스


L7(Layer 7) 로드밸런싱
웹-백엔드
/
메신저-백엔드
/messenger
Load Balancer
로드밸런서
클라이언트
데이터베이스

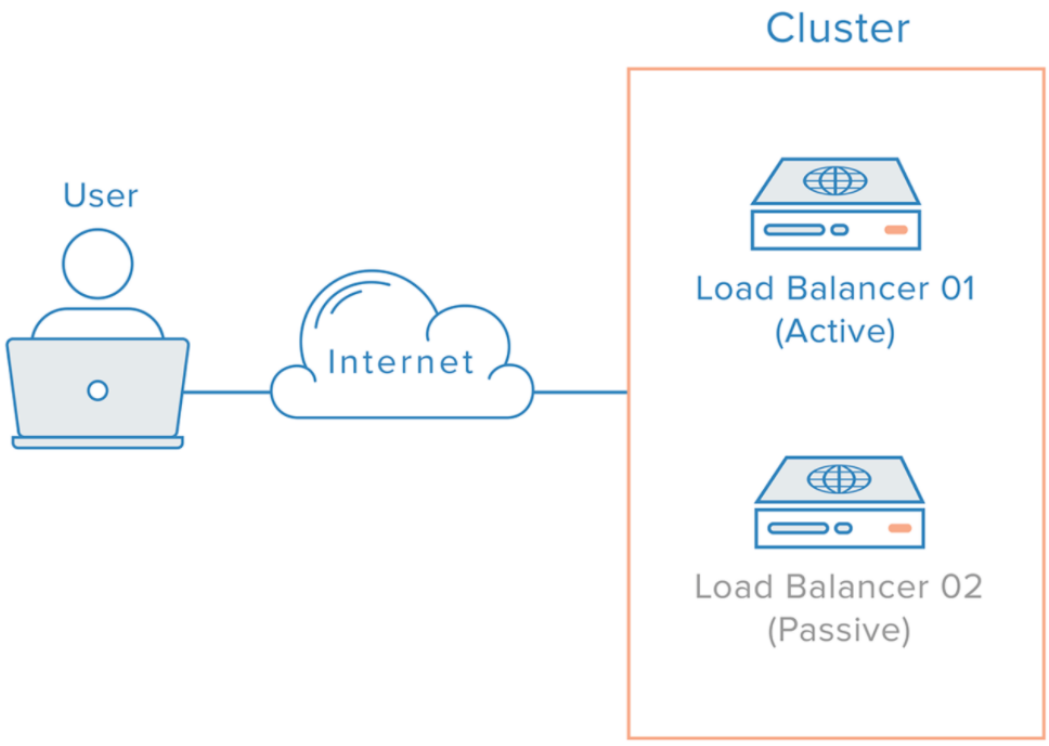
L4 와 L7 비교 표

	L4 로드밸런서	L7 로드밸런서
네트워크 계층	Layer 4 전송계층(Transport layer)	Layer 7 응용계층(Application layer)
특징	> TCP/UDP 포트 정보를 바탕으로 함	> TCP/UDP 정보는 물론 HTTP의 URI, FTP의 파일명, 쿠키 정보 등을 바탕으로 함
장점	> 데이터 안을 들여다보지 않고 패킷 레벨에서만 로드를 분산하기 때문에 속도가 빠르고 효율이 높음 > 데이터의 내용을 복호화할 필요가 없기에 안전함 > L7 로드밸런서보다 가격이 저렴함	> 상위 계층에서 로드를 분산하기 때문에 훨씬 더 섬세한 라우팅이 가능함 > 캐싱 기능을 제공함 > 비정상적인 트래픽을 사전에 필터링할 수 있어 서비스 안정성이 높음
단점	> 패킷의 내용을 살펴볼 수 없기 때문에 섬세한 라우팅이 불가능함 > 사용자의 IP가 수시로 바뀌는 경우라면 연속적인 서비스를 제공하기 어려움	> 패킷의 내용을 복호화해야 하기에 더 높은 비용을 지불해야 함 > 클라이언트가 로드밸런서와 인증서를 공유해야하기 때문에 공격자가 로드밸런서를 통해서 클라이언트에 데이터에 접근할 보안 상의 위험성이 존재함

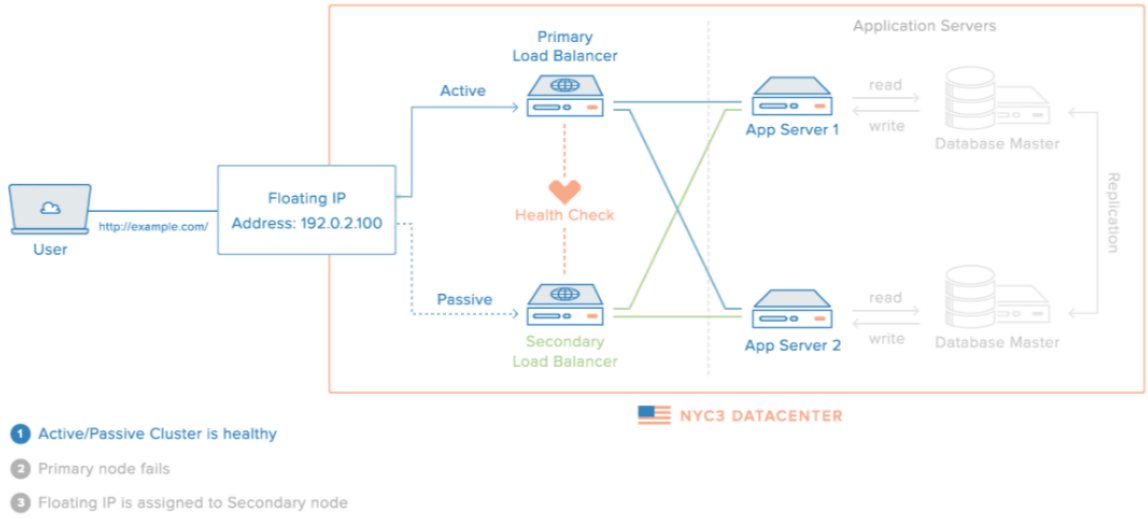
로드 밸런서 장애 대비

서버를 분배하는 로드 밸런서에 문제가 생길 수 있으므로, 로드 밸런서를 이중화하여 대비함

Active 상태와 Passive 상태



장애 발생 시나리오



1. 이중화된 로드 밸런서들은 서로 Health Check 를 함
2. 메인 로드 밸런서가 동작하지 않으면, 가상 IP (Virtual IP, VIP)는 여분의 로드 밸런서로 변경됨
3. 여분의 로드 밸런서로 운영함

- 로드밸런서(Load Balancer)의 개념과 특징
- [AWS]비전공자도 이해할 수 있는 로드밸런싱
- 로드 밸런서(Load Balancer)란?
- [AWS] 가장쉽게 VPC 개념잡기