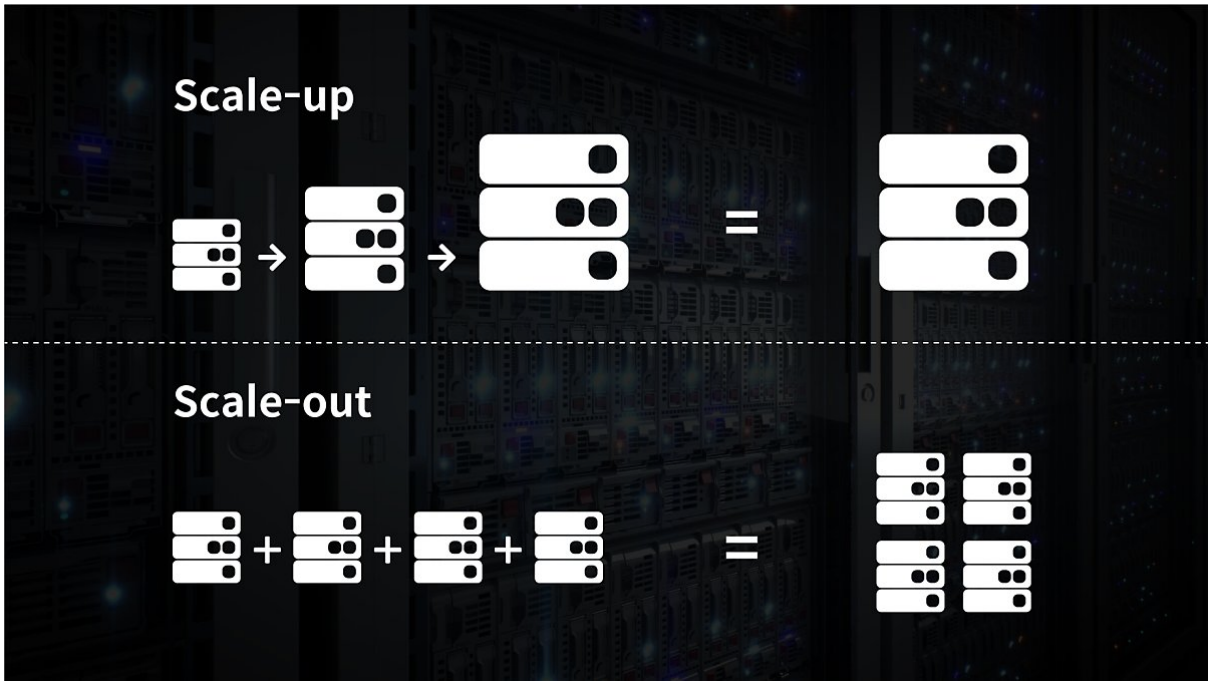
로드 밸런싱 (Load Balancing)

로드 밸런싱은 **네트워크 또는 서버에 가해지는 부하(Load)를 분산(Balancing) 해주는 기술**을 말한다. 로드 밸런싱 기술을 제공하는 서비스 또는 장치는 (로드 밸런서, Load Balancer)는 클라이언트와 네트워크 트래픽이 집중되는 서버들(Server Pool) 또는 네트워크 허브 사이에 위치한다.

특정 서버 또는 네트워크 허브에 부하가 집중되지 않도록 트래픽을 다양한 방법으로 분산하여 서버나 네트워크 허브들의 성능을 최적인 상태로 유지할 수 있도록 한다.

로드 밸런싱의 필요성

로드 밸런싱은 여러 대의 서버를 두고 서비스를 제공하는 **분산 처리 시스템**에서 필요한 기술이다. 서비스 제공의 초기 단계라면 유저가 적어 서버로의 요청도 적을 것이므로 서버 한 대로도 요청에 응답하는 것이 가능할 것이다. 그러나 사업이 커지고 클라이언트 유저 수가 많아지면 기존 서버만으로 대처하기 힘들어지게 된다. 증가한 트래픽에 대처하는 방법은 크게 두 가지이다.



1) Scale-up : **서버의 성능 자체를 향상**시키는 방법. 서버 CPU를 i3에서 i7로 업그레이드 하는 것

2) Scale-out : 기존 서버 방식와 동일하거나 낮은 성능의 **서버를 두 대 이상을 운영하는 것.** CPU가 i3인 컴퓨터를 여러 대 추가 구입하는 것. Scale-out의 방식으로 서버를 증설하기로 했다면 여러 대의 서버로 트래픽을 균등하게 분산시켜주는 **로드 밸런싱**이 필요하다.

정적 로드 밸런싱

> 정적 로드 밸런싱 알고리즘은 고정된 규칙을 따르며 현재 서버 상태와 무관하다.

라운드 로빈 방식 (Round Robin Method)

서버에 들어온 요청을 순서대로 돌아가며 배정한다. 클라이언트의 요청을 순서대로 배분하기 때문에 여러 대의 서버가 동일한 스펙을 가지고 있고 서버와의 연결(세션)이 오래 지속되지 않는 경우에 활용하기 적합하다.

라운드 로빈 방식에서는 권한 있는 이름 서버가 특수 하드웨어나 소프트웨어 대신 로드 밸런싱을 수행한다. 이름 서버는 서버 팜에 있는 여러 서버의 IP 주소를 차례대로 또는 라운드 로빈 방식으로 반환한다.

가중 라운드 로빈 방식 (Weighted Round Robin Method)

각각의 서버마다 가중치를 가지고 가중치가 높은 서버에 클라이언트를 우선적으로 배분한다. 주로 서버의 트래픽 처리 능력이 상이한 경우 사용되는 부하 분산 방식이다. 예를 들어 A라는 서버가 가중치 5를 갖고 B라는 서버가 가중치 2를 갖는다면 로드 밸런서는 라운드 로빈 방식으로 A 서버에 5개 B 서버에 2개 요청을 전달한다.

IP 해시 방식 (IP Hash Method)

클라이언트의 IP 주소를 특정 서버로 매핑하여 요청을 처리하는 방식이다. 사용자의 IP를 해싱해 로드를 분배하기 때문에 사용자가 항상 동일한 서버로 연결되는 것을 보장한다.

동적 로드 밸런싱

> 동적 로드 밸런싱 알고리즘은 트래픽을 배포하기 전에 서버의 현재 상태를 검사한다.

최소 연결 방식 (Least Connection Method)

요청이 들어온 시점에 가장 적은 연결상태를 보이는 서버에 우선적으로 트래픽을 배분한다. 자주 세션이 길어지거나 서버에 분배된 트래픽이 일정하지 않은 경우에 적합한 방식이다.

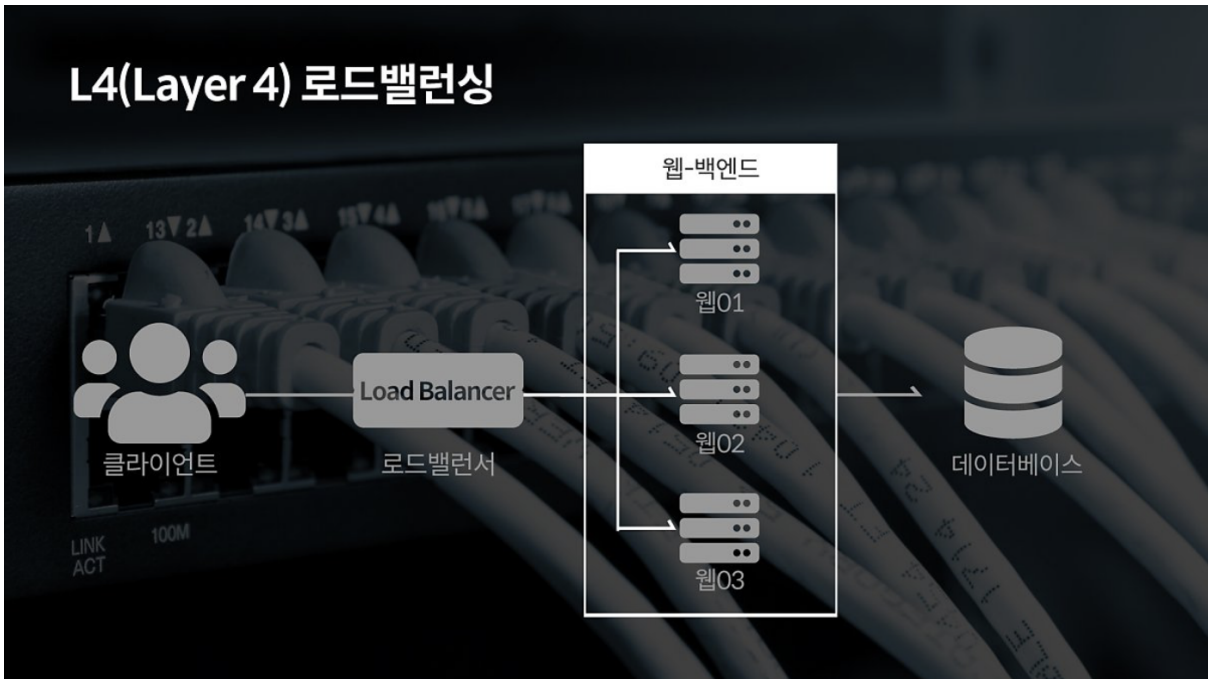
최소 리스폰 타임 (Least Response Time Method)

서버의 현재 연결 상태와 응답 시간(Response Time, 서버에 요청을 보내고 최초 응답을 받을 때까지 소요되는 시간)을 모두 고려하여 트래픽을 배분한다. 가장 적은 연결상태와 가장 짧은 응답시간을 보이는 서버에 로드를 배분하는 방식이다.

로드 밸런싱 종류

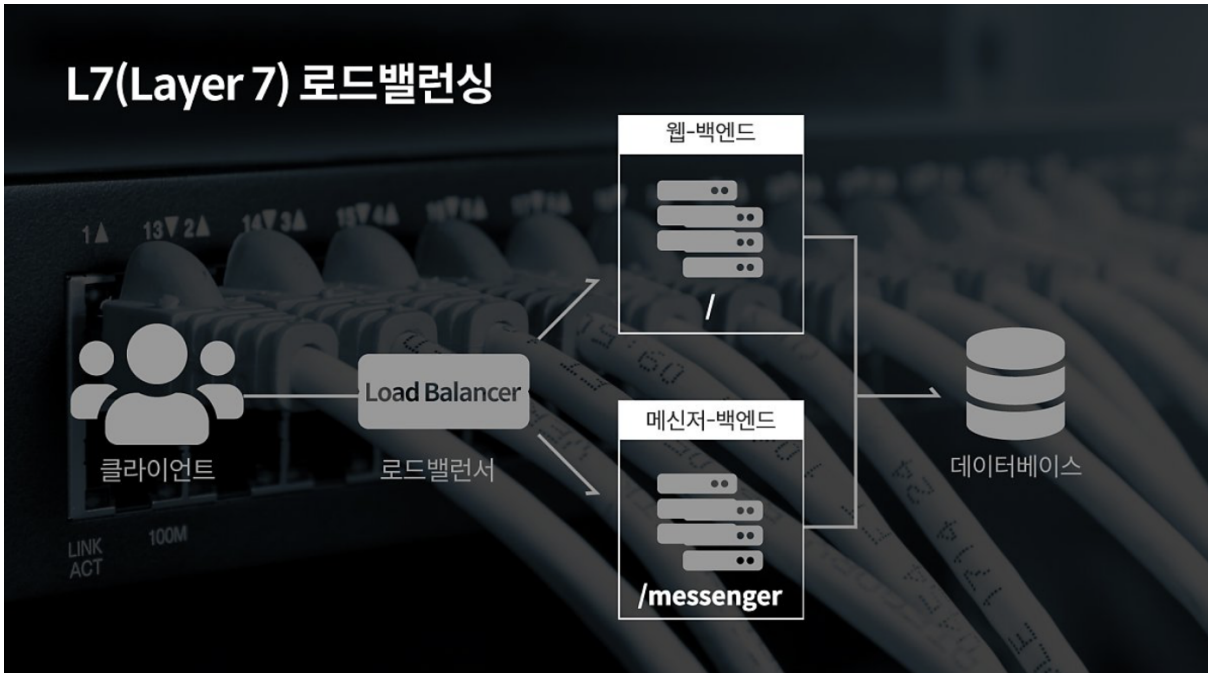
부하 분산에는 L4 로드밸런서와 L7 로드밸런서가 많이 사용된다. L4부터 포트를 다룰 수 있기 때문이다.

L4 로드밸런싱



L4 로드 밸런서는 **IP, 포트**를 기준으로 스케줄링 알고리즘을 통해 부하를 분산한다. 클라이언트에서 로드 밸런서로 요청을 보냈을 때 최적의 서버로 요청을 전송하고 결과를 클라이언트로 보낸다. 즉, 요청하는 서비스에 상관없이 서버를 돌린다.

L7 로드밸런싱



L7 로드 밸런서는 **애플리케이션 계층(HTTP, FTP, SMTP)에서 로드를 분산**하기 때문에 HTTP 헤더, 쿠키 등과 같은 사용자의 요청을 기준으로 특정 서버에 트래픽을 분산하는 것이 가능하다. 즉, **패킷의 내용을 확인하고 그 내용에 따라** 로드를 특정 서버에 배분하는 것이 가능하다.