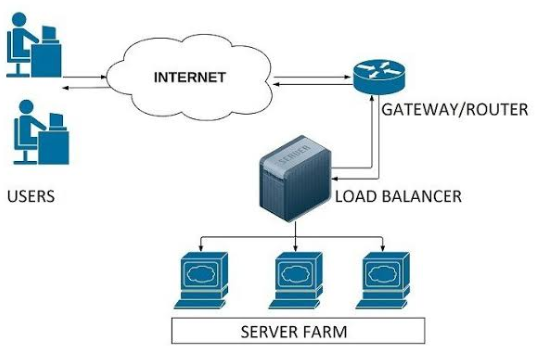
로드 밸런싱(부하 분산)이란?



네트워크/서버에 가해지는 부하를 분산해주는 기술. 많은 트래픽을 감당하기 위해서 한 서버의 성능을 높이는 작업(Scale-Up)은 한계가 있기 때문에, 그 트래픽들을 여러 대의 서버(Scale-Out)로 분산하여 처리하는 기술을 로드 밸런싱이라고 한다.

이 기술을 제공하는 서비스/장치를 로드밸런서라고 하는데, 이는 서버들 혹은 네트워크 허브 사이에 위치한다.

로드밸런서의 기본 기능

Health Check – 서버들에 대한 주기적인 Health Check를 통해 서버들의 장애 여부를 판단한다. 이로 인해 서버 일부에 이상이 생겨도 다른 정상적으로 동작하는 서버로 트래픽을 보내줄 수 있다.

L3 – ICMP를 이용하여 서버의 IP주소가 통신 가능한 상태인지.

L4 – TCP의 3,4-way handshake를 바탕으로 각 포트 상태 체크

L7 – 어플리케이션 계층에서 체크를 수행. 실제 웹페이지에 통신을 시도

터널링 – 눈에 보이지 않는 통로를 만들어 통신할 수 있게 하는 개념. 로드밸런서는 클라이언트-서버 간 중간에서 제공한다. 연결된 상호 간에만 캡슐화된 패킷을 구별해 캡슐화를 해제하게 한다.

NAT(Network Address Translation) - IP주소를 변환해주는 기능.

SNAT – 내부->외부로 트래픽이 나가는 경우. 내부 사설 IP주소를 외부의 공인 IP주소로 변환해주는데, 공유기가 대표적. 부족한 공인 IP를 효과적으로 사용 가능.

DNAT – 외부->내부. 로드밸런서가 이 대표적인 예시. 로드밸런서의 관점에서 여러 개의 호스트가 하나의 공인IP주소를 통해 접속하는 것이 주 목적

DSR(Direct Server Routing) – 서버에서 클라이언트로 되돌아가는 경우, 네트워크 장비나 로드밸런서로 거치지 않고 바로 클라이언트로 찾아가서 로드밸런서의 부하를 줄여준다.

로드 밸런싱의 종류와 방법.

L4 로드밸런싱 – 네트워크 계층(IP), 트랜스포트 계층(TCP, UDP)정보를 바탕으로 부하를 분산합니다.

라운드로빈 – 세션을 각 서버에 순차적으로 맺어주는 방식(담당자가 있으니 짧게)

가중치 및 비율 할당 방식 – 서버마다 비율을 설정해 두고 해당 비율만큼 세션을 맺어주는 방식. Ex) 서버의 성능이 좋다면 비율을 높인다.

최소 연결 – 가장 적은 세션을 가진 서버로 트래픽을 보낸다.(가장 많이 사용)

응답 시간 – 가장 빠른 응답 시간을 보내는 서버로 트래픽을 우선 보내준다. 각 서버들의 가용한 리소스/성능/처리중인 데이터의 양이 다를 경우 적합한 방식.

해시 기반 – 특정 클라이언트는 특정 서버로만 할당시키는 방식. 경로가 보장되며, 접속자 수가 많을수록 분산 및 효율이 뛰어나다.

대역폭 기반 – 서버들과의 대역폭을 고려하여 트래픽을 분산하는 방식.

L7 로드밸런싱 – L4의 기능을 포함하는 것 뿐만 아니라, OSI 7계층의 프로토콜(HTTP, SMTP, FTP)등을 바탕으로도 분산 처리가 가능하다.

URL스위칭 방식 – 특정 하위 URL들은 특정 서버로 처리하는 방식.

컨텍스트 스위칭 방식 – 클라이언트가 요청한 특정 리소스에 대해 특정 서버 등으로 연결을 해줄 수 있다.

쿠키 지속성 – 쿠키 정보를 바탕으로 클라이언트가 연결했었던 동일한 서버에 계속 할당해주는 방식.

L4 vs L7 장/단점

L4 로드밸런서

패킷의 내용을 확인하지 않고 부하를 분산 -> 속도가 빠르고 효율이 높음. But 섬세한 라우팅 불가.

데이터의 내용을 복호화할 필요가 없어 안전하다.

가격이 L7보다 저렴하다.

사용자의 IP가 수시로 바뀐다면 연속적인 서비스 제공이 힘들다.

L7 로드밸런서

상위 계층에서 부하를 분산 -> 섬세한 라우팅. 캐싱 기능을 제공

비정상적인 트래픽을 사전에 필터링 -> 서비스 안정성 높음.

L4에 비해 가격이 비쌈. 패킷 복호화 -> 더 높은 비용

클라이언트가 로드밸런서와 인증서 공유 -> 이를 통한 보안상의 위험

로드 밸런서의 주요 성능 지표 – 초당 연결 수, 동시 연결 수, 처리 용량