로드 밸런싱(Load Balancing)은 여러 대의 서버에 들어오는 트래픽을 분산하여 시스템의 성능을 향상시키고 장애 발생 시에도 서비스의 가용성을 높이는 기술이다.

로드 밸런싱의 개념

1. 로드 밸런서(Load Balancer) :
    1. 클라이언트의 요청을 받아 여러 서버에 분산하는 역할을 하는 장비나 소프트웨어이다.
    2. 하드웨어 로드 밸런서와 소프트웨어 로드 밸런서가 있다.
2. 로드 밸런싱 알고리즘 :
    1. 라운드 로빈(Round Robin) : 요청을 순서대로 각 서버에 분산하는 방식이다.
    2. 최소 연결(Least Connections) : 현재 연결 수가 가장 적은 서버에 요청을 분산하는 방식이다.
    3. IP 해시(IP Hash) : 클라이언트의 IP 주소를 해싱하여 특정 서버에 요청을 전달하는 방식이다.
    4. 가중치 기반(Weighted) : 서버의 능력에 따라 가중치를 부여하고 그 가중치에 따라 요청을 분산한다.
3. 세션 유지(Session Persistence) :
    1. 특정 클라이언트의 요청이 항상 동일한 서버로 전달되도록 하는 기능이다. 이를 통해 세션 상태를 유지할 수 있다. 방법으로는 쿠키 기반, IP 기반 등이 있다.
4. 장애 조치(Failover) :
    1. 서버 중 하나가 다운되었을 때 로드 밸런서가 자동으로 다른 서버로 트래픽을 전환하여 서비스를 계속 유지하는 기능이다.
5. 스케일링(Scaling) :
    1. 수요 증가에 따라 서버를 추가하거나 제거하여 시스템의 성능을 조절할 수 있다. 로드 밸런서가 자동으로 이 작업을 지원하기도 한다.

로드 밸런서의 종류

1. 하드웨어 로드 밸런서 :
    1. 물리적인 장비로 제공되는 로드 밸런서이다. 고성능을 제공하지만 비용이 높은 경우가 많다.
2. 소프트웨어 로드 밸런서 :
    1. 소프트웨어로 구현된 로드 밸런서이다. 유연성과 비용 측면에서 유리하며 클라우드 환경에서 많이 사용된다.
3. 클라우드 기반 로드 밸런서 :
    1. AWS의 Elastic Load Balancing(ELB), Azure의 Azure Load Balancer, Google Cloud의 Cloud Load Balancing 등 클라우드 서비스 제공자가 제공하는 로드 밸런서이다.

로드 밸런싱은 대규모 웹 애플리케이션이나 서비스에서 성능과 신뢰성을 유지하는 데 필수적인 기술이다. 구현 방법과 필요성은 시스템의 요구 사항에 따라 달라질 수 있다.