1. 복수 서버로 부하 분산

서버에 액세스가 증가할 때는 고속화한 회선에서 흘러오는 대량의 패킷에 서버의 처리 능력이 따라잡지 못할 수도 있다.

특히 CGI(Common Gateway Interface)등의 애플리케이션에서 페이지의 데이터를 동적으로 만드는 경우에는 서버 머신의 프로세서 파워를 사용하기 때문에 더욱 중요해 진다.

이를 해결하기 위해 복수의 서버를 사용하여 처리를 분담하는 **분산처리** 방법이 있다. 이때 처리를 분담하는 방식은 여러 가지가 있다.

- ➔ 가장 간단한 방법은 단순히 여러 대의 웹 서버를 설치하고 한 대가 담당하는 사용자 수를 줄이는 방법이다.
- ➔ 이런 방법을 취할 경우 클라이언트가 보내는 리퀘스트를 웹 서버에 분배하는 구조가 필요하다.
  - ■ 여러 방법 중 DNS 서버에 분배하는 방법이 가장 간단하다.
  - ■ DNS서버에 같은 이름으로 여러 대의 웹 서버를 등록해 놓으면 DNS 서버는 조회가 있을 때 마다 차례대로 IP주소를 되돌려준다.
  - ■ 예를 들어 192.0.2.60, 192.0.2.70, 192.0.2.80의 IP주소를 대응시키면 60부터 70, 80의 순서로 순회하면서 회답한다.
  - ■ 이러한 방식을 **라운드 로빈**이라고 한다.
  - ■ 단 이러한 방법은 고장 난 웹 서버에도 상관하지 않고 IP주소를 회답한다는 것이다.
  - ■ 따라서 웹 서버가 변하면 대화가 도중에 끊길 수도 있다.

2. 부하 분산 장치로 복수의 웹 서버로 분할

라운드 로빈 방식의 문제점을 피하기 위해 **부하 분산 장치** 또는 **로드 밸런서** 등으로 부르는 기기가 고안되었다.

- **DNS에서 웹 서버에 IP주소를 보내기 위해 여러 서버를 판단하는 방법**

➔ 판단 근거는 여러가지가 있다.

➔ 대화가 복수 페이지에 걸쳐 있지 않은 단순한 액세스라면 웹 서버의 부하 상태가 판단 근거가 된다.

- 웹 서버와 정기적으로 정보를 교환하여 CPU나 메모리의 사용률 등을 수집하고, 이것을 바탕으로 어느 웹 서버의 부하가 낮은지 판단하거나, 시험 패킷을 웹 서버에 보내 응답 시간으로 부하를 판단한다.
- 단 웹 서버의 부하는 단 시간에 증가하거나 감소하므로 꼼꼼히 상황을 조사하지 않으면 정확한 곳까지 파악할 수 없게 되고, 또 너무 자세한 상황을 조사하면 그 동작 자체로 웹 서버의 부하가 증가하게 된다.

➔ 대화가 복수 페이지에 걸쳐 있을 때는 웹 서버의 부하에 관계없이 이전의 리퀘스트와 같은 웹 서버에 전송해야 한다.

- 이를 위해 대화가 복수의 페이지에 걸쳐 있는지 먼저 판단해야 한다.
- HTTP의 기본동작은 리퀘스트 메시지를 보내기 전에 TCP의 접속 동작을 하고, 응답 메시지를 반송하면 연결을 끊는다.
- 이후 다음 웹 서버에 액세스 할 때는 TCP접속 동작부터 다시 수행하므로 웹 서버 측에서 보면 HTTP의 대화는 1회씩 전혀 다른 것으로 보여 받은 리퀘스트가 이전 리퀘스트와 연결된 것인지 아닌지를 판단하기 어렵다.
- 이러한 전후 관계를 판단하기 위해 양식에 입력한 데이터를 보낼 때 그 안에 전후의 관련을 나타내는 정보를 부가하거나, HTTP헤더 필드에 부가하는 방법 등이 고안되었다.
- 부하 분산 장치는 이러한 정보를 조사하여 일련의 동작이라면 이전과 같은 웹 서버에 리퀘스트를 전송하고, 그렇지 않으면 부하가 적은 웹 서버에 전송하도록 동작한다.