Chapter4.

- 본 장에서는 3장에서 다뤘던 방식대로 처리율 제한 장치 설계 문제를 접근하며, 문제 해결 방식과 처리율 제한 장치에 대한 내용을 전달한다.

개요

- 처리율 제한 장치란 일정 시간 동안 요청 횟수 기준을 초과하는 요청을 무시하거나, 일정 시간 이후 처리하도록 핸들링하는 객체를 의미한다.
- 클라이언트에 있을 수도 있고, 서버에 있을 수도 있지만 클라이언트 요청은 지나치게 조작이 쉬우므로 적합하지 않아 서버 혹은 클라이언트와 서버 사이의 미들웨어에서 처리하는 것이 일반적이다.
- 처리율 제한 장치는 요구사항에 따라 까다롭지만 `공격에 의한 자원 고갈 공격 방지, 비용 절감, 서버 과부하 방지` 등의 장점을 갖는다.

문제 해결 절차

1. 문제 이해 및 설계 범위 확정

- 이 단계에서
  - 어떤 종류의 처리율 제한 장치여야 하는지
  - 규모는 어느 정도를 염두에 두는지
  - 시스템이 단일 서버 환경을 기준으로 동작하는지 혹은 분산 환경을 기준으로 동작하는지
  - 어떤 기준을 통해 처리율을 제한할 것인지

  등 설계 범위 및 동작 방식을 확정한다.

- 설정된 처리율을 초과하는 요청은 정확하게 제한한다.
- 낮은 응답시간: 이 처리율 제한 장치는 HTTP 응답시간에 나쁜 영향을 주어서는 곤란하다.
- 가능한 한 적은 메모리를 써야 한다.
- 분산형 처리율 제한(distributed rate limiting): 하나의 처리율 제한 장치를 여러 서버나 프로세스에서 공유할 수 있어야 한다.
- 예외 처리: 요청이 제한되었을 때는 그 사실을 사용자에게 분명하게 보여주어야 한다.
- 높은 결함 감내성(fault tolerance): 제한 장치에 장애가 생기더라도 전체 시스템에 영향을 주어서는 안 된다.