[2장] 개략적인 규모 추정

가상 면접 사례로 배우는 대규모 시스템 설계 기초

이인호 / mystyle2006

개략적인 규모 추정

- 보편적으로 통용되는 성능 수치상에서 사고 실험을 행하여 추정치를 계산하는 행위

쉽게 이야기해서 **소프트웨어 프로젝트의 규모를 예측하고 이해하는 과정**을 뜻한다.

이 과정들을 잘 이해하기 위해 아래 내용들을 잘 이해하고 있어야 한다.

- 2의 제곱 수
- 응답지연(Latency)
    - 컴퓨터 네트워크나 시스템에서 발생하는 현상으로, 요청을 보낸 후에 응답을 받기까지 걸리는 시간을 말합니다. 일반적으로 응답지연은 네트워크 지연, 서버 부하, 클라이언트 레이턴시 등 다양한 요인에 의해 발생할 수 있다.
- 가용성에 관련된 수치
    - 주로 시스템이나 서비스가 얼마나 신뢰할 수 있는지를 나타내는 지표로 사용된다.

2의 제곱 수

- 데이터 볼륨의 단위를 2의 제곱수로 표현했을 때 어떤 값이 나오는지 알아야 한다.
- 최소 단위: 1바이트(8비트)

2의 제곱	근사치	이름	축약형
10	1천(Thousand)	1킬로바이트	1KB
20	1백만(million)	1메가바이트	1MB
30	10억(billion)	1기가바이트	1GB
40	1조(trillion)	1테라바이트	1TB
50	1000조(quadrillion)	1페타바이트	1PB

1GB? 1GiB?

1GB와 1GiB는 둘 다 데이터 저장 용량을 나타내는 단위입니다. 그러나 이 둘은 다른 단위이며, 사용되는 문맥에 따라 차이가 있습니다.

- **1GB (Gigabyte):** GB는 10의 9승 바이트를 의미합니다. 이는 일반적으로 하드 디스크 용량, 메모리 용량 등의 실제 데이터 저장 용량을 나타낼 때 사용됩니다. 예를 들어, 하드 디스크 용량이 1GB인 경우에는 약 1,000,000,000 바이트의 데이터를 저장할 수 있음을 의미합니다.
- **1GiB (Gibibyte)**: GiB는 2의 30승 바이트(1,073,741,824 바이트)를 의미합니다. 이는 주로 컴퓨터 메모리 용량을 나타낼 때 사용됩니다. 컴퓨터의 이진수 체계에 따라 계산되며, 정확한 1,073,741,824 바이트의 데이터를 나타냅니다.

[Tips]

예전에는 데이터 저장 용량을 나타내는 단위로 GB(Gigabyte)가 널리 사용되었습니다. 그러나 컴퓨터 과학 분야에서는 2의 거듭제곱 단위를 사용하는 이진수 체계에 따라 바이트를 나타내는 데에 GiB(Gibibyte)를 사용하는 것이 더 정확하다고 판단되었습니다.

이러한 변화는 1998년에 국제 표준화 기구(ISO)와 국제 전기통신 연합(ITU)에 의해 채택된 새로운 표준인 "국제 단위 체계"에 따라 이루어졌습니다. 이 표준에서는 이진수 체계에 따라 데이터 용량을 나타낼 때에는 GiB, TiB(Tebibyte), PiB(Pebibyte) 등의 단위를 사용하도록 권장했습니다.

응답지연은 왜 발생할까?

- **네트워크 지연:** 데이터가 전송되는 동안 발생하는 시간적 지연으로, 데이터가 송신자와 수신자 간의 네트워크를 통해 이동하는 동안 발생합니다. 네트워크 지연은 대기시간, 전송시간, 처리시간 등으로 구성될 수 있습니다.

- **서버 부하:** 서버가 요청을 처리하기 위해 필요한 시간이 증가함에 따라 응답지연이 발생할 수 있습니다. 서버 부하는 CPU, 메모리, 디스크 I/O 등의 자원 사용량이 높을 때 발생할 수 있습니다.

- **클라이언트 레이턴시:** 클라이언트가 요청을 보내고 응답을 받는 데 걸리는 시간으로, 클라이언트의 네트워크 연결 상태 및 처리 속도에 따라 변할 수 있습니다.

- **처리 시간:** 서버에서 요청을 처리하는 데 걸리는 시간으로, 데이터베이스 쿼리 실행, 계산 작업, 파일 입출력 등의 작업에 의해 영향을 받을 수 있습니다.

프로그래머가 알아야 하는 응답지연 값

ns: nanosecond(나노초), μs: microsecond(마이크로초), ms: millisecond(밀리초)

- 메모리는 빠르지만 디스크는 아직도 느리다.
  - 메모리에서 1MB 순차적으로 읽었을 때, 250,000ns = 250μs
  - 디스크에서 1MB 순차적으로 읽었을 때, 30,000,000ns = 30ms (120배 느림)
- 디스크 탐색(seek)은 가능한 한 피하라.
  - 디스크 탐색, 10,000,000ns = 10ms
- 단순한 압축 알고리즘은 빠르다.
  - Zippy로 1KB 압축, 10,000ns = 10μs
  - 데이터를 인터넷으로 전송할 때 가능하면 압축하라

가용성과 관련된 수치 - SLA

SLA(서비스 수준 계약, Service Level Agreement)은 서비스 제공자와 서비스 이용자 간의 계약으로, 서비스 제공자가 제공해야 하는 서비스 수준에 대한 규정과 이를 준수하기 위한 조건을 명시하고 있다.

SLA는 서비스 제공자와 이용자 간의 신뢰를 구축하고 서비스 수준을 보장하기 위한 중요한 도구입니다. 서비스 제공자는 SLA를 준수하여 고객 만족도를 높이고 신뢰를 유지할 수 있으며, 서비스 이용자는 SLA를 통해 서비스 수준을 신뢰하고 요구 사항을 충족시킬 수 있습니다.

그렇다면 어떤 명시들이 존재할까?

가용성과 관련된 수치 - SLA

서비스 수준 목표(SLO, Service Level Objective): 서비스 제공자가 제공해야 하는 서비스 수준에 대한 목표와 성능 지표를 정의합니다. 예를 들어, 서비스의 가용성, 응답 시간, 처리량 등이 SLO의 일부가 될 수 있습니다.

서비스 수준 지표(SLI, Service Level Indicator): 서비스의 상태를 측정하고 모니터링하기 위한 지표로, 실제 서비스의 동작을 측정하여 SLA의 이행 여부를 확인하는 데 사용됩니다.

보상 및 제재 조항: 서비스 제공자가 SLA를 위반할 경우 이를 보상하거나 제재를 가하는 조항을 정의합니다. 이는 서비스 이용자가 제공자가 약속한 서비스 수준을 충족하지 않을 경우에 대비한 보호 수단입니다.

서비스 수준 목표 달성 방법: 서비스 제공자가 어떻게 서비스 수준 목표를 달성할 것인지에 대한 방법과 절차를 설명합니다. 이는 서비스 제공자와 이용자 간의 협력과 역할 분담을 명확히 합니다.

가용성과 관련된 수치 - SLA

서비스 수준 목표(SLO, Service Level Objective): 서비스 제공자가 제공해야 하는 서비스 수준에 대한 목표와 성능 지표를 정의합니다. 예를 들어, 서비스의 가용성, 응답 시간, 처리량 등이 SLO의 일부가 될 수 있습니다.

서비스 수준 지표(SLI, Service Level Indicator): 서비스의 상태를 측정하고 모니터링하기 위한 지표로, 실제 서비스의 동작을 측정하여 SLA의 이행 여부를 확인하는 데 사용됩니다.

보상 및 제재 조항: 서비스 제공자가 SLA를 위반할 경우 이를 보상하거나 제재를 가하는 조항을 정의합니다. 이는 서비스 이용자가 제공자가 약속한 서비스 수준을 충족하지 않을 경우에 대비한 보호 수단입니다.

서비스 수준 목표 달성 방법: 서비스 제공자가 어떻게 서비스 수준 목표를 달성할 것인지에 대한 방법과 절차를 설명합니다. 이는 서비스 제공자와 이용자 간의 협력과 역할 분담을 명확히 합니다.

가용성과 관련된 수치 - QPS

"Queries Per Second"의 약어로, 초당 쿼리 수를 의미합니다. 주로 데이터베이스나 웹 서버와 같은 시스템에서 사용되며, 특정 시간 동안 수행된 쿼리 또는 요청의 개수를 나타낸다.

예를 들어, 웹 서버가 초당 100개의 HTTP 요청을 처리하는 경우 해당 서버의 QPS는 100이 된다. 데이터베이스에서 초당 100개의 데이터베이스 쿼리를 처리하는 경우 해당 데이터베이스의 QPS는 100이 된다.

QPS는 시스템의 부하나 처리량을 측정하고 모니터링하는 데 사용된다. 높은 QPS는 시스템이 더 많은 요청을 처리할 수 있다는 것을 의미하지만, 과도한 QPS는 시스템에 부하를 일으키고 성능을 저하시킬 수 있다. 따라서 QPS를 관리하고 최적화하는 것은 시스템의 성능을 향상시키는 데 중요하다.

가용성과 관련된 수치 - QPS 사례

가정

- 월간 능동 사용자가 1억명
- 30%의 사용자가 매일 서비스를 이용
- 각 사용자는 10건의 채팅을 항상 이용

위 사례에서 QPS를 계산하기 위해 필요한 정보는 아래와 같다.

- 매일 서비스를 이용하는 사용자의 수
- 각 사용자가 채팅을 하는 빈도 (채팅당 QPS)

가용성과 관련된 수치 - QPS 사례

위 가정을 통해

- 사용자의 수는 월간 능동 사용자의 30%인 1억 * 0.3 = 3천만명
- 사용자당 10건의 채팅을 이용하기에 각 사용자당 채팅당 QPS는 10건

따라서 이 서비스의 QPS는

- 매일 서비스를 이용하는 사용자의 수 * 채팅당 QPS = 3천만명 * 10건 3억 QPS 이다.
- 여기에 채팅 한 건을 저장하는데 100바이트가 필요하다고 했을 때 저장소 요구량은
    - 3천만 * 10 * 100 = 30,000,000,000 바이트 (약 30GB)

감사합니다