**Chapter 2**

Back-of-the-envelope Estimation

Back-of-the-envelope estimations – made from though experiments and common performance numbers for which design best fit the requirements

**Power of two**

* A byte is a sequence of 8 bits
* ASCII character uses 1 byte of memory

|  |  |  |  |
| --- | --- | --- | --- |
| **Power** | **Approximate value** | **Full name** | **Short name** |
| 10 | 1 thousand | 1 kilobyte | 1 KB |
| 20 | 1 million | 1 megabyte | 1 MB |
| 30 | 1 billion | 1 gigabyte | 1 GB |
| 40 | 1 trillion | 1 terabyte | 1 TB |
| 50 | 1 quadrillion | 1 petabyte | 1 PB |

Latency numbers – length of typical computer operations

* Disk seek – amount of time that a disk drive’s head takes to move to a specific location on a disk

2010 from Google

|  |  |
| --- | --- |
| **Operation name** | **Time** |
| L1 cache reference | 0.5 ns |
| Branch mispredict | 5 ns |
| L2 cache reference | 7 ns |
| Mutex lock/unlock | 100 ns |
| Main memory reference | 100 ns |
| Compress 1K bytes with Zippy | 10,000 ns = 10 us |
| SSD random read | 16,000 ns = 16 us |
| Send 2K bytes over 1 Gbps network | 20,000 ns = 20 us |
| Read 1 MB sequentially from memory | 250,000 ns = 250 us |
| Round trip within the same datacenter | 10,000,000 ns = 10 ms |
| Disk seek | 10,000,000 ns = 10 ms |
| Read 1 MB sequentially from the network | 10,000,000 ns = 10 ms |
| Read 1 MB sequentially from disk | 30,000,000 ns = 30 ms |
| Send packet CS (California) -> Netherlands -> CA | 150,000,000 ns = 150 ms |

General conclusions

* Memory is fast, disk is slow
* Avoid disk seeks if possible
* Simple compression algorithms are fast
* Compress data before sending it over the internet if possible
* Data centers in different regions takes more time to send data between them

Availability numbers – the ability of a system to be continuously operational for a desirably long period of time

* 100% means a service that zero downtime
* Most services fall between 99-100%
* Service level agreement (SLA) – agreement between the service provider and customer formally defining the level of uptime the service will deliver
  + Amazon, Google, Microsoft set their SLAs at 99.9% or above
* Uptime is traditionally measured in nines

|  |  |  |
| --- | --- | --- |
| **Availability %** | **Downtime per day** | **Downtime per year** |
| 99% | 14.40 minutes | 3.65 days |
| 99.9% | 1.44 minutes | 8.5 hours |
| 99.99% | 8.64 seconds | 52 minutes |
| 99.999% | 864 milliseconds | 5 minutes |
| 99.9999% | 86.4 milliseconds | 31 seconds |

Estimating QPS and storage requirements

* QPS – queries per second
* DAU – daily active users
* Commonly asked back-to-the-envelope estimations: QPS, peak QPS, storage, cache, number of services

**Tips for estimations**

* Use rounding & approximations
* Lay down your assumptions
* Label units