Latency Numbers Everyone Should Know

Operation	Time in ns	Time in ms (1ms = 1,000,000 ns)
L1 cache reference	1	
Branch misprediction	3	
L2 cache reference	4	
Mutex lock/unlock	17	
Main memory reference	100	
Compress 1 kB with Zippy	2,000	0.002
Read 1 MB sequentially from memory	10,000	0.010
Send 2 kB over 10 Gbps network	1,600	0.0016
SSD 4kB Random Read	20,000	0.020
Read 1 MB sequentially from SSD	1,000,000	1
Round trip within same datacenter	500,000	0.5
Read 1 MB sequentially from disk	5,000,000	5
Read 1 MB sequentially from 1Gbps network	10,000,000	10
Disk seek	10,000,000	10
TCP packet round trip between continents	150,000,000	150

Therefore, it is possible to read:

- sequentially from HDD at a rate of ~200MB per second
- sequentially from SSD at a rate of ~1 GB per second
- sequentially from main memory at a rate of ~100GB per second (burst rate)
- sequentially from 10Gbps Ethernet at a rate of ~1000MB per second

No more than 6-7 round trips between Europe and the US per second are possible, but approximately 2000 per second can be achieved within a datacenter.

Back of the Envelope Calculations

Quick tips: Use numbers based on the decimal system to run numbers in your head.

Sample calculation:
What is the overall latency of retrieving 30 256kB images from one server?

Naïve design: do all the work on one machine - dominated by disk seek time.

Reads required to generate page:

30 images / 2 disks per machine = 15

Time to read one image from HDD:

(256KB / 1MB) * 5 ms + 10 ms seek = 11.28 ms

Approximate time to generate results:

15 reads * 11.28 ms = 169.2 ms

One HDD-based server can generate 1000 ms / 169.2 ms ~= 5 result pages per second.