Caches
What are three qualities of an ideal memory unit?

Fast
Large
Cheap
What is the memory hierarchy?

A hierarchy of memory devices, each one larger but slower than the last

- L1 cache: 0.5 ns
- L2 cache reference: 7 ns
- Main memory reference: 100 ns
- Read 4K randomly from SSD: 150,000 ns, 0.15 ms
- Read 1 MB sequentially from memory: 250,000 ns, 0.25 ms
- Round trip within same datacenter: 500,000 ns, 0.5 ms
- Read 1 MB sequentially from SSD*: 1,000,000 ns, 1 ms
- Disk seek: 10,000,000 ns, 10 ms
- Read 1 MB sequentially from disk: 20,000,000 ns, 20 ms (80x memory, 20X SSD)

https://gist.github.com/jboner/2841832
How do we predict what will be used?

Locality of Reference
  Temporal Locality
  Spatial Locality
How do we know what’s in the cache?

Split the memory address into tag (upper bits), index (middle bits) and offset (lower bits)
There are a number of slots for each index - match the tags to validate address
How much memory does a 4KB cache with 8B blocks take up? (assume 32 bit memory address)

5.25 KB
What happens when we miss?

Stall the processor and read from lower level memory
In which slot does each index go?

Direct Mapped
Set Associative
Fully Associative
How do we find data in an associative cache?
Use the index bits to find the proper set
Match against the tags for that set
How does associativity change the partitioning of an address into tag and index for a given # of frames?

More associativity means shorter indexes and longer tags.
What do we (ideally) want in our cache?

Data that is going to be read next
How do we replace data to get close?

Random
FIFO
Least Recently Used
Not Most Recently Used (cop out)
How do we store to a single cache?

Same way as reading, but being very careful not to write until a hit is guaranteed.
How do we write to a cache hierarchy?

Write-through
Write-back
What happens when a write misses?

Write-allocate
Write-non-allocate

Don’t stall
What are the types of misses?

Compulsory
Capacity
Conflict
How can we reduce conflict misses?

Victim buffer
What is inclusion / exclusion?

Inclusion - get block from memory, goes in L1 and L2
Exclusion - block goes only in L1, L2 acts as victim buffer
How do we ensure cache coherence in multicore setting?

MSI and MESI Cache Coherence Protocol
How does caching affect performance?