CSE 120
Day 13 Notes

Elijah Hantman

Caching Review

- Cache holds recently referenced data
- Exploits principal of locality
- Computer systems often use many caches
- Caching is not limited to hardware.
- Memory hierarchy, lots of slow memory, smaller and smaller amounts of faster memory.
- Only adjacent levels in the memory hierarchy can transfer information.

Motivation

- Memory is too slow
- Caches speed up access
- What does into a cache?
- How do caches work?
- Why do caches work?
- What are the tradeoffs when designing a cache?

How do we transfer data?

- Every pair of levels can be thought of as an upper and lower level. Usually we transfer an entire block when something is copied between levels.
- Block (or line)
  the minimum unite of information that can be moved in a cache
- Hit Rate
  The fraction of memory access found in a level of the memory hierarchy
- Miss Rate
  The fraction of memory accesses not found in a level of the memory hierarchy
- Hit Time
  The time required to access a level of the memory hierarchy, including time required to determine whether it is a hit or a miss.
- Miss Penalty
  The time required to fetch a block into a level of the memory hierarchy from a lower level, including time taken to access the block, transmit the block, and insert the block into the new level, and then passing the block to the requestor.

What is a cache?

- Cache holds recently referenced data
  functions as a buffer for larger, slower components
- Exploits locality
  - Provide as much inexpensive storage space as possible
  - Offer access speed equivalent to the fastest memory
    * For data in the cache
    * key is to have the right data cached
- Computer systems extensively use caching and caches
- Cache ideas are not limited to hardware.

Direct Mapped Cache

- Assume single cache level
- All data is aligned, there are no gaps between any data.
- Assume 64 bit address for caching.
- 2^{10} bytes = 1024 Bytes = 1KiB

> The little "i" in the unit indicates the base is 2 rather than 10.

- Location of Block in Cache determined by (main) memory address
- Direct Mapped means only one choice.
- Two end of a spectrum, Direct Mapped caches, Full Associative cache. Between these two there are n-set associative caches.
- Direct Mapped means that each block must go into a single cache line, Full associative means the block can go into any place in the cache. N-Set Associative mappings can place a given block into N different locations.
- Direct mapped cache $\rightarrow$ 1 way associative
  n-set associative where n is the number of blocks in the cache.
- Full Associative cache $\rightarrow$ n way associative where n is the number of blocks in the cache.
  1-set associative cache.

- Cache location is determined by either modulo the address, or masking out the upper bits.
- Modern L1 cache uses lower 12 bits for determining set.
- Rest of address is used as a tag for validation.
- has the disadvantage of having a high miss rate. This is because many memory addresses map to a single block in cache.
- Secondary disadvantage of direct mapped cache is that for virtual memory many bits tend to be unusable until converted via a TLB or memory mapper.
- Valid bit which is only set when real data is loaded into the cache. Initialized to 0, only set when a read happens.

Hardware Design

- 64 bit address
- Smallest addressable unit is 1 bytes
- Cache size is 2^n blocks
- Block size is 2^m words
- No. of bits used to reference cache is $n + m + 2$
- Size of tag field is $64 - (n + m + 2)$

Direct Mapped Cache, and Cache Writes

- How does a CPU know which block index to access?
  - Formula is: Block Address % Number of blocks
  - This requries conversion from byte to block address.
  - Address is $\lfloor \text{Byte address} \,/\, \text{Bytes per block} \rfloor$
- Impact of Block Size
  - Large the block size, the greater the miss penalty
  - Large blocks favor spatial locality
  - However large blocks decrease number of blocks in a cache, and cause more collisions and greater miss rate. In addition, a larger block requires more time to search through for the correct address.

-