Caching

4-9-2003

Opening Discussion

- What did we talk about last class?
- Have you seen anything interesting in the news?
- What are the advantages and disadvantages of VLIW compared to dynamic superscalar pipelining? Which results in a more flexible architecture?

The Ideal Memory

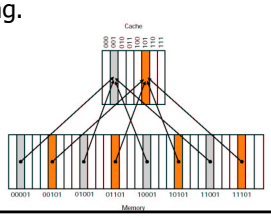
- We would like to have unlimited fast memory, but that doesn't happen.
- As processors have increased in speed, latency of RAM has become more significant.
  - Current SDRAM has a latency of ~20ns. The first read latency can easily be 50 clock cycles though burst reads will let sequential data be read in consecutive clock cycles

Cache and Memory Hierarchies

- Faster memory typically costs more and even if it didn't, we couldn't get that much of it "close" to the processor for fast access.
- Instead we put in multiple levels of memory with smaller, faster memory closer to the processor.
- Spatial and temporal locality allow this to work well.
- Hits and Misses

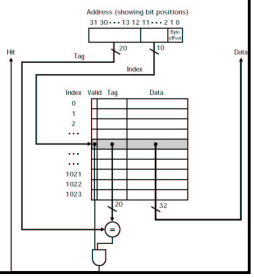
Direct Mapped Caches

- A memory address maps to exactly 1 cache block.
- Use low bits for position in cache and high bits for tag.



Reading from a Cache

- Given an address we use the low bits to determine cache block. We see if that block is valid and if the tag on it matches the upper bits on our address.
- They they match we can just read from cache (a hit).



What Happens When We Miss?

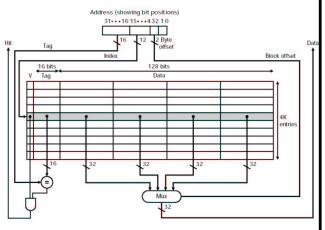
- What if the block is invalid or the tag doesn't match? This is called a miss and it means we have to go out to main memory. We can't do that in one clock cycle.
- We need to decrement the PC by 4 and stall the entire pipeline while waiting for memory to return the value to the cache where it will be found when the instruction is restarted.

Writing to Memory

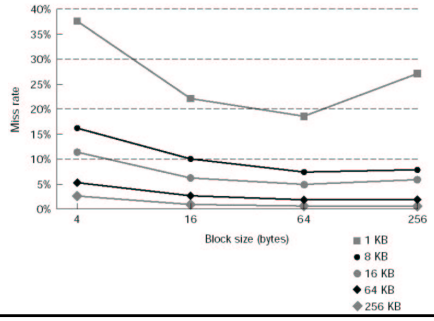
- Writing to memory is more complex because we can't just write to cache. It needs to hit memory at some point.
- First we always write to the cache and set the tag and valid bit.
  - write-through: always write to memory
  - write buffers: puts write in buffer so we can continue pipeline. Stalls if buffer is full.
  - write-back: only write when we overwrite that block.

Using Spatial Locality

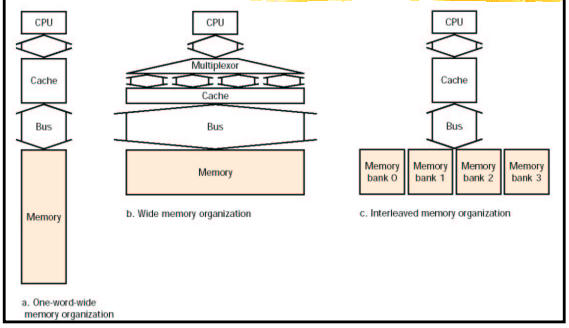
- Using multiple word blocks we reduce the number of misses if there is spatial locality.
- Reading many words uses burst modes.
- Write misses require reading a full block.



Miss Rates



Memory System Design



Minute Essay

- Compare the delays we had from things like branch misprediction and data hazards to those that we get from cache misses. Can you begin to see why some "new" chips don't upgrade the core much, but instead improve the memory subsystem?
- Remember that assignment #6 is due on Friday.