CS 61C: Great Ideas in Computer Architecture (Machine Structures)

Instructors:
Randy H. Katz
David A. Patterson
http://inst.eecs.Berkeley.edu/~cs61c/fa10

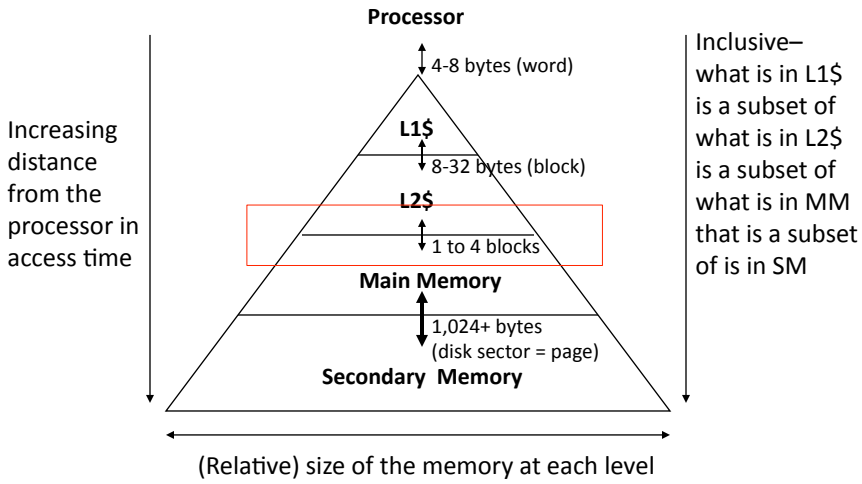
Agenda

- Direct Mapped Cache (continued)
- Administrivia
- Technology Break
- Cache-Memory Interface

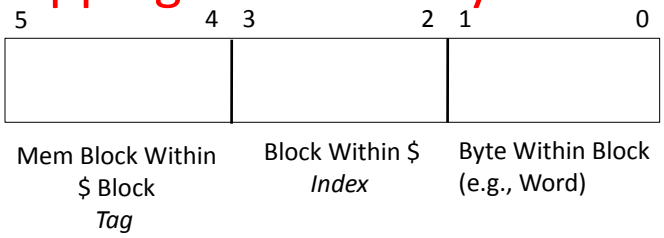
Agenda

- Direct Mapped Cache
- Administrivia
- Technology Break
- Cache-Memory Interface

Characteristics of the Memory Hierarchy



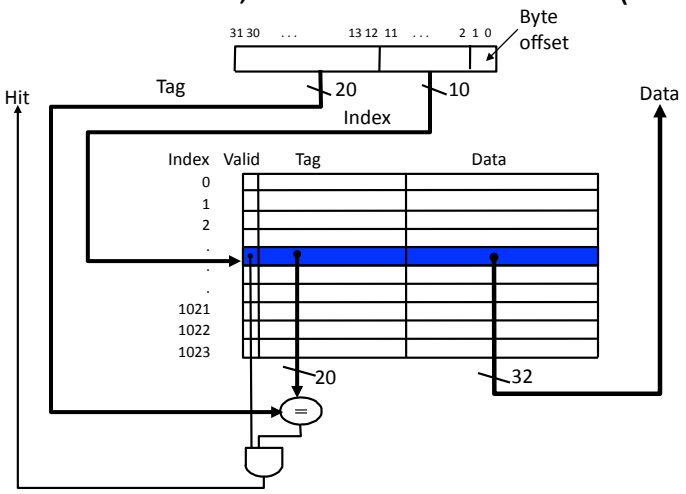
Mapping the Memory Address



- Note: $ = Cache
- In example, block size is 4 bytes/1 word (it could be multi-word)
- Memory and cache blocks are the same size, unit of transfer between memory and cache
- # Memory blocks >> # Cache blocks
  - 16 Memory blocks/16 words/64 bytes/6 bits to address all bytes
  - 4 Cache blocks, 4 bytes (1 word) per block
  - 4 Memory blocks map to each cache block
- Byte within block: low order two bits, ignore! (nothing smaller than a block)
- Memory block to cache block, aka *index*: middle two bits
- Which memory block is in a given cache block, aka *tag*: top two bits

Direct Mapped Cache Example

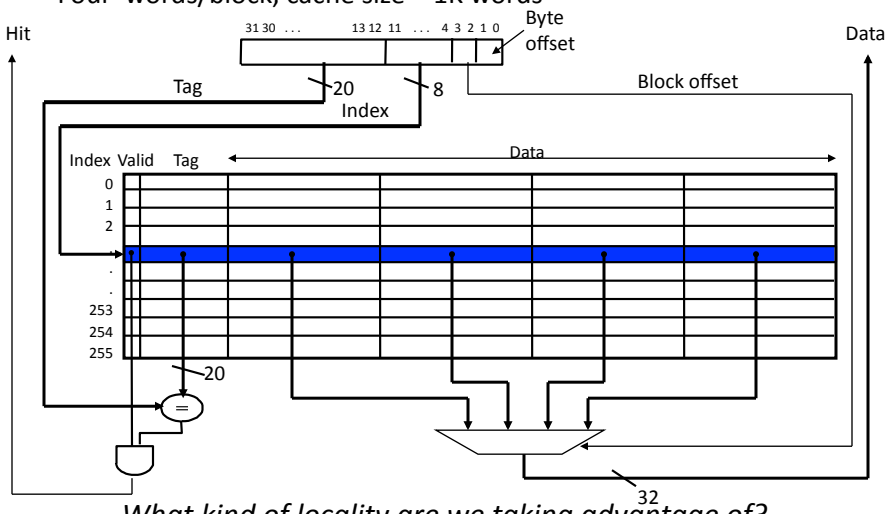
- One word blocks, cache size = 1K words (or 4KB)



What kind of locality are we taking advantage of?

Multiword Block Direct Mapped Cache

- Four words/block, cache size = 1K words



What kind of locality are we taking advantage of?

Taking Advantage of Spatial Locality

- Let cache block hold more than one word

Start with an empty cache - all blocks initially marked as not valid

0 1 2 3 4 3 4 15

0 miss

00	Mem(1)	Mem(0)

1 hit

00	Mem(1)	Mem(0)

2 miss

00	Mem(1)	Mem(0)
00	Mem(3)	Mem(2)

3 hit

00	Mem(1)	Mem(0)
00	Mem(3)	Mem(2)

4 miss

01 5 4

~~00~~	~~Mem(1)~~	~~Mem(0)~~
00	Mem(3)	Mem(2)

3 hit

01	Mem(5)	Mem(4)
00	Mem(3)	Mem(2)

4 hit

01	Mem(5)	Mem(4)
00	Mem(3)	Mem(2)

15 miss

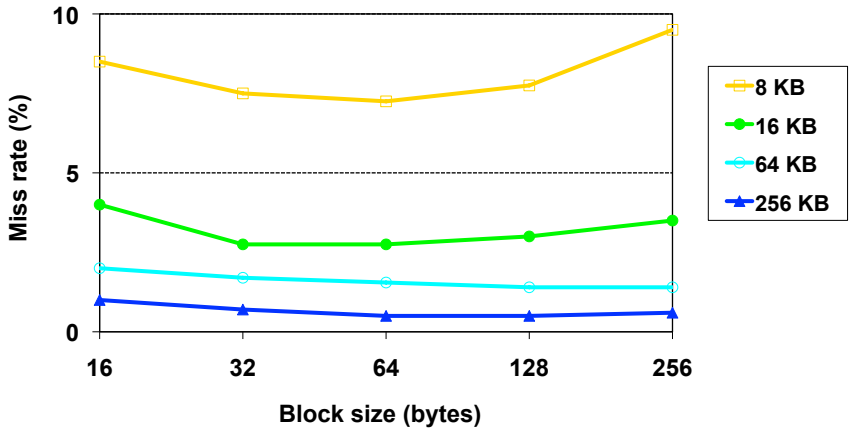
11

01	Mem(5)	Mem(4)
~~00~~	~~Mem(3)~~	~~Mem(2)~~

15 14

- 8 requests, 4 misses

Miss Rate vs Block Size vs Cache Size



- Miss rate goes up if the block size becomes a significant fraction of the cache size because the number of blocks that can be held in the same size cache is smaller (increasing capacity misses)

Agenda

- Direct Mapped Caches
- Administrivia
- Technology Break
- Cache-Memory Interface

Midterm!

- HW #3: Posted, Due SUNDAY@23:59:59
- Exam Review on Monday, 100 GPB 6-8 PM
- NO lecture next Wednesday, 6 October
- Exam, 6-9 PM, 1 Pimentel
  - Closed book, notes
  - No calculator
  - One 8.5” x 11” crib sheet

Agenda

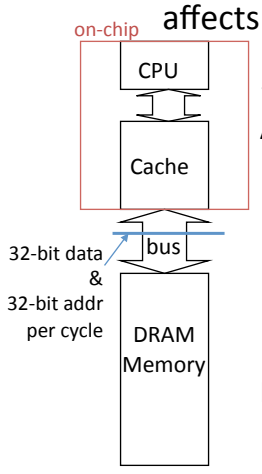
- Direct Mapped Caches
- Administrivia
- Technology Break
- Caches-Memory Interface

Agenda

- Cache Hits and Misses
- Administrivia
- Technology Break
- Caches-Memory Interface

Memory Systems that Support Caches

- The off-chip interconnect and memory architecture affects overall system performance in dramatic ways



One word wide organization (one word wide bus and one word wide memory)

Assume

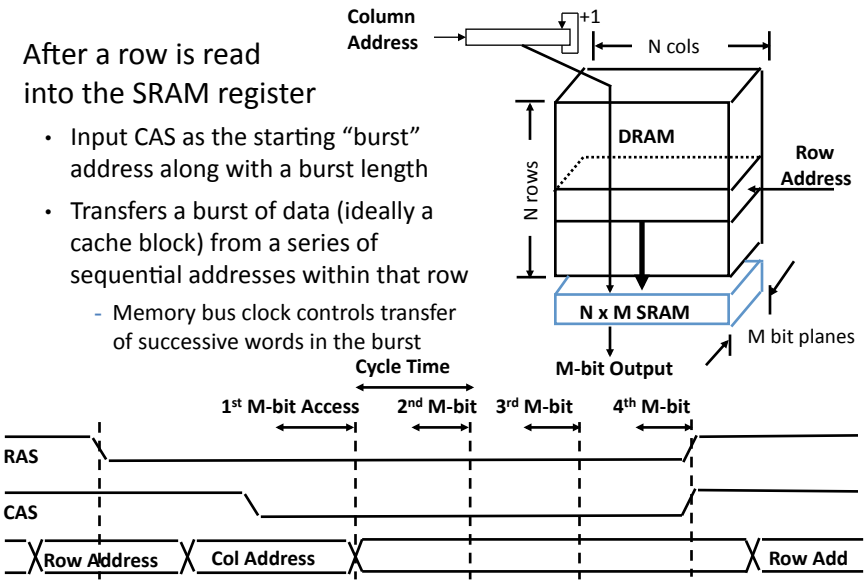
- 1 memory bus clock cycle to send address
- 15 memory bus clock cycles to get the 1st word in the block from DRAM (row cycle time), 5 memory bus clock cycles for 2nd, 3rd, 4th words (subsequent column *access* time)—note effect of latency!
- 1 memory bus clock cycle to return a word of data

Memory-Bus to Cache bandwidth

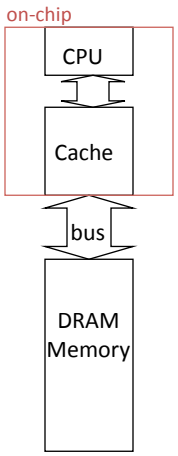
- Number of bytes accessed from memory and transferred to cache/CPU per memory bus clock cycle

(DDR) SDRAM Operation

- After a row is read into the SRAM register
  - Input CAS as the starting "burst" address along with a burst length
  - Transfers a burst of data (ideally a cache block) from a series of sequential addresses within that row
    - Memory bus clock controls transfer of successive words in the burst



One Word Wide Bus, One Word Blocks

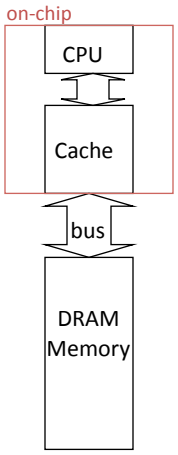


- If block size is one word, then for a memory access due to a cache miss, the pipeline will have to *stall* for the number of cycles required to return one data word from memory

1	memory bus clock cycle to send address
15	memory bus clock cycles to read DRAM
1	memory bus clock cycle to return data
17	total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is
  - 4/17 = 0.235 bytes per memory bus clock cycle

One Word Wide Bus, Four Word Blocks



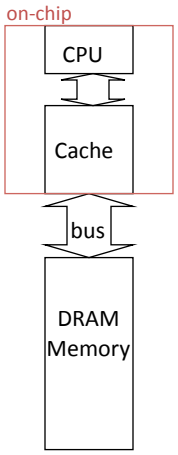
- What if block size is four words and each word is in a different DRAM row?

	1	cycle to send 1st address
4 x 15 =	60	cycles to read DRAM
	1	cycles to return last data word
	62	total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

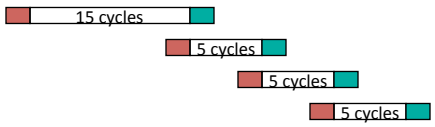
(4 x 4)/62 = 0.258 bytes per clock

One Word Wide Bus, Four Word Blocks



- What if the block size is four words and all words are in the same DRAM row?

	1	cycle to send 1st address
15 + 3*5 =	30	cycles to read DRAM
	1	cycles to return last data word
	32	total clock cycles miss penalty

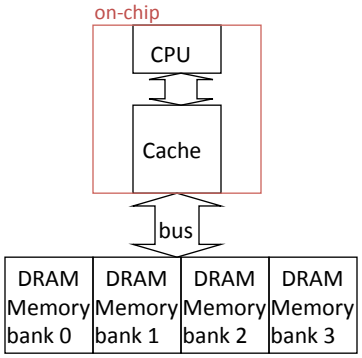


- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

(4 x 4)/32 = 0.5 bytes per clock

Interleaved Memory, One Word Wide Bus

- For a block size of four words



1	cycle to send 1st address
15	cycles to read DRAM banks
4*1 = 4	cycles to return last data word
20	total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

(4 x 4)/20 = 0.8 bytes per clock

DRAM Memory System Observations

- Its important to match the cache characteristics
  - Caches access one block at a time (usually more than one word)
- With the DRAM characteristics
  - Use DRAMs that support fast multiple word accesses, preferably ones that match the block size of the cache
- With the memory-bus characteristics
  - Make sure the memory-bus can support the DRAM access rates and patterns
  - With the goal of increasing the Memory-Bus to Cache bandwidth

Summary

- Hits in caches hide the long access latencies to main memory
- Misses suffer from high latency of going to main memory—processor may have to wait for memory in these cases unless other tricks are used (future parallelism discussions!)
- Mitigate the effect of the latency by exploiting memory bandwidth and parallelism to move a block from memory to cache in the hope that locality will increase future hits