CSE1400 - Computer Organisation

Self-Study: Week 8 - Caching and Pipelining

Delft University of Technology

2021/2022 Q1

Special thanks to Sára Juhošová, Ana Băltăreţu and Kiril Vasilev
for helping with the compilation of this set of questions.

Important information:

1. If any question is unclear please consult Answers. For more specific questions, you can use the Queue during lab hours.
2. The average time for solving this self study is **3** hours, and **1** hour is allocated to giving feedback. Timings are included for each exercise to give you a more clear overview of how much time you should be spending on them.
3. The maximum amount of points for this self study is 200 points. To get the points you should submit a serious attempt on Peer and **properly review** your peers' submissions (100 points per full cycle, including review evaluation).
4. Answers will be provided during the weekly tutorial sessions.

1 Caching

1. (3 mins) Imagine you have some data that is larger than the size of all of your registers added together, where would you save this data if:

   (a) you use it very often, almost for every operation?

   Cache, since data is easily accessable and safe in the short-run

   (b) it is very important, but you do not use it that often?

   Main memory, since it is safer for data in the long-run, even though its access is harder.

2. (3 mins) Explain the following terms:

   (a) Spatial Locality: when things are close in space (consequitive bytes)

   (b) Temporal Locality: when things are close in time (last used element)

3. (8 mins) A computer uses 32-bit words and a word-addressable main memory of 16 MiB. It also uses a direct-mapped cache of 256 KiB. The block size is 128 bytes.

   (a) How many words fit in a block?

   (a) ______________

   $$\frac{\text{block size}}{\text{word size}} = \frac{128\,\text{B}}{32\,\text{b}} = \frac{\overset{4}{\cancel{128}} \times 8\text{b}}{\cancel{32}\,\text{b}} = 32\ \frac{\text{words}}{\text{block}}$$

   (b) How many blocks fit in the cache?

   (b) ______________

   $$\frac{\text{cache size}}{\text{block size}} = \frac{\overset{2}{\cancel{256}}\,\text{Ki}\cancel{\text{B}}}{\cancel{128\text{B}}} = 2\text{Ki} = 2048\ \frac{\text{blocks}}{\text{cache}}$$

   (c) How many bits address a word inside a block?

   (c) ______________

   $32 = 2^5 \Rightarrow$ 5-bit addressing

   (d) How many bits address a block inside the cache?

   (d) ______________

   $2048 = 2^{11} \Rightarrow$ 11-bit addressing

4. (8 mins) You have been gifted an 8-way set-associative cache, that has 512 blocks of 128 bytes each. You also know that your main memory has 8 GiB in size and that all memories are byte addressable. Now you are wondering:

   (a) how many bits are needed to address each byte within a block?

   (a) ______________

   128 B byte-addressable block $= 2^7$ B → 7-bit addressing

   (b) how many bits are required to address each set?

   (b) ______________

   $\frac{512}{8} = 64$ sets $= 2^6 \Rightarrow$ 6-bit addressing

   (c) how many bits are required for the tag?

   (c) ______________

   8 GiB byte-addressable $= 2^{33}$ B → 33 bits total

   tag $= 33 - 6 - 7 =$ 20 bits

tag	set	byte
	6	7

   33

5. (10 mins) Consider a cache which uses an LRU cache replacement algorithm. The table below shows the age counters for all blocks in one of the sets in this set-associative cache with 8 blocks per set. The following occurs within this set of the cache:

   1. A cache hit occurs on block 6.
   2. A cache hit occurs on block 4.
   3. A cache miss occurs and the oldest block is replaced with new data.
   4. A cache hit occurs on block 5.

   You can use the following table to fill in your solution:

Block	age before	age after (1)	age after (2)	age after (3)	age after (4)
0	100	100	101	110	110
1	011	011	100	101	101
2	001	010	011	100	100
3	000	001	010	011	011
4	101	101	000	001	001
5	111	111	111	000	000
6	010	000	001	010	010
7	110	110	110	111	111

2 Pipelining

1. (30 mins) Ana wants to keep Answers clean of !smart questions so she decides to gather an army of students to downvote posts that are not relevant. Now all she needs is something to bribe them with. She decides to ask her Among Us crew to help her make some brownies for the students. They divide the tasks like this:

   - Nathan and Cas buy and carry the required ingredients from the store, which is located 10 minutes away and it takes them 5 more minutes to pick out all the ingredients for 3 brownie mixes.
   - Iarina makes one brownie mix every 10 minutes (only if she has all the required ingredients, initially there are no supplies at the place where they cook).
   - Tony watches over the brownies while they cook. It takes them 30 minutes to get cooked perfectly and Tony can fit 2 mixes in the oven at one time (one mix results in 8 individual brownies).
   - Eames packs the 4 brownies into one box and puts a ribbon on top, which in total takes him 5 minutes per box.
   - Ana makes sure everything runs smoothly, but does no task. → that's who we aspire to be ☺

   They decide that to optimize their process, they need to work together and in a pipeline-like manner.

   (a) Draw how the pipeline would look for 85 minutes:

ingredients (in) = 25 min (10 min to shop, 5 min to pick, 10 min to return)
output: in for 3 mixes

brownie mix (mix) = 10 min

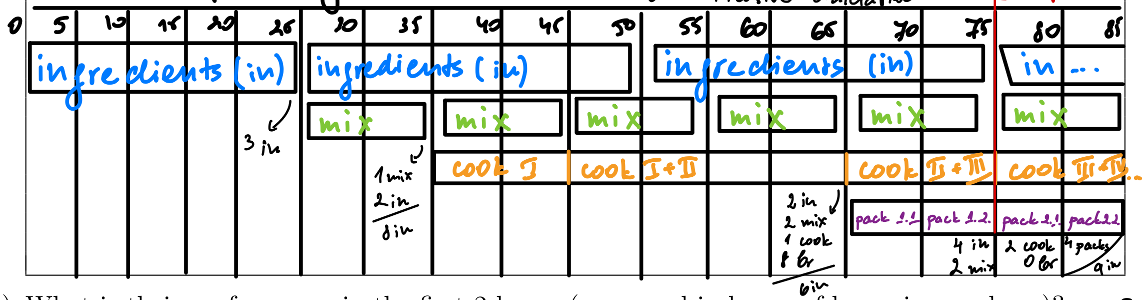
brownie cooking (cook) = 30 min (max 2 mix)
Output: 1 mix = 8 brownies

brownie packing (pack) = 5 min

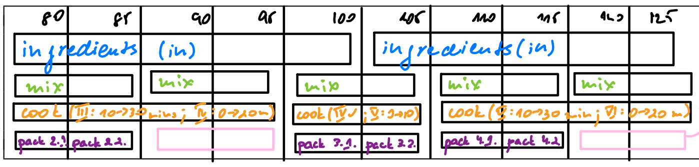
input	current
in (3 every 25 min)	in + mix + cook + $\frac{br}{8}$ + $\frac{packs}{4}$

cannot identify clear bottle neck (might be cooking, but we need more extensive evaluation

← stop for now



   (b) What is their performance in the first 2 hours (measured in boxes of brownies per hour)?



cooking is the bottle neck

packing 5

✓ 3×3 in = 9 in = 4 + 2 + 2 + 1

+3

1/4

85: 4 in + 2 mix + 2 cook + 0 br + 4 packs
+ 0 in + 0 mix + 2 cook + 0 br + 4 packs

120: 4 in + 2 mix + 4 cook + 0 br + 8 packs

12 = 4 + 2 + 4 + 2

$\frac{8 \text{ boxes}}{2h} = 4 \frac{\text{boxes}}{\text{hour}}$

   (c) What is their performance after these first two hours (measured in boxes of brownies per hour)?

bottle neck → cooking ⇒ $4 \frac{\text{cooks}}{\text{hour}} = 32 \frac{\text{brownies}}{\text{hour}} = 8 \frac{\text{boxes}}{\text{hour}}$

1h: cook [x: 10→30; x+1: 0→20 min] 20 min + cook [x+1: 20→30 min; x+2: 0→10] 10 min + cook [x+2: 10→30; x+3: 0→20 min] 20 min + cook [x+3: 20→30 min; x+4: 0→10 min] 10 min

(d) Ana realizes that since they have been playing so much Among Us, some members of the crew members might start acting like Impostors in real life, which means that their task would take double the time it normally would (for Tony, doubling the time is equal to only putting 1 mix into the oven), so she decides to step in. Whose place should Ana take in order to guarantee that they get the maximum performance (without knowing the actual impostor)? Explain why.

Since we concluded that cooking is the bottle neck step even when we have double performance (being able to bake two mixes at a time), cooking must be the place Ana steps in (we don't aspire to be her now :)

2. (10 mins) Consider the following program:

```
1 Add R1, R2, R3        R3 = R1+R2
2 Move (R4), R5         R5 = (R4)
3 Move #10, R6          R6 = 10
4 Add R3, R4, R7        R3+R4 = R7
```

A processor with a four-stage pipeline (**F**etch, **D**ecode, **E**xecute, **W**rite) has to execute a small program. Each stage takes one cycle to complete. Reading an operand from memory causes a stall and adds a two-cycle delay.

Pmemory = 2 cycle

(a) How long does it take to execute the program?

(a) ________________

I1	F1	D1	E1	W1	4 5	6	7	✓	
I2	1	F2	D2			E2	W2	8	
I3		2	F3	D2 2cycle memory delay		D3	E3	W3	9
I4			3	F4	D2 2 cycle memory delay		D4	E4	W4

9 cycles

(b) A second decode unit is added. A fetched instruction can go to either decode unit. If two instructions finish decoding simultaneously, the oldest instruction is executed first. How long does it take to execute the program with this improved processor?

(b) ________________

I1	F1	D1	E1	W1	✓			
I2		F2	D2			E2	W2	
I3			F3	D3	E3	W3		
I4				F4	second decoding unit	D4	E4	W4

8 cycles (1 cycle improvement; should be 2 from memory delay but we have a decoding overlap).