CS 3330 Computer Architecture, Spring 2020
HW 5: Caching and Main Memory

Instructor: Prof. Samira Khan
TAs: Amel Fatima, Sihang Liu, Korakit Seemakhupt, Yasas Senevirathne, Yizhou Wei
Min Jae Lee, Nikita Semichev, Yuying Zhang.

Assigned: Apr 15, 2020
Due: **Apr 22, 2020**

1 Caching [40 points]

Below, we have given you four different sequences of addresses generated by a program running on a processor with a data cache. Cache hit ratio for each sequence is also shown below. Assuming that the cache is initially empty at the beginning of each sequence, find out the following parameters of the processor's data cache:

- Associativity (1, 2 or 4 ways)
- Block size (1, 2, 4, 8, 16, or 32 bytes)
- Total cache size (256 B, or 512 B)
- Replacement policy (LRU or FIFO)

Assumptions: all memory accesses are one byte accesses. All addresses are byte addresses.

Sequence No.	Address Sequence	Hit Ratio
1	0, 2, 4, 8, 16, 32	0.33
2	0, 512, 1024, 1536, 2048, 1536, 1024, 512, 0	0.33
3	0, 64, 128, 256, 512, 256, 128, 64, 0	0.33
4	0, 512, 1024, 0, 1536, 0, 2048, 512	0.25

- **Cache block size?**
- **Associativity?**
- **Total cache size?**
- **Replacement policy?**

2 Memory Scheduling [40 points]

A machine has a DRAM main memory organized as 2 channels, 1 rank and 2 banks/channel. An open row policy is used, i.e., a row is retained in the row-buffer after an access until an access to another row is made. The following commands (defined as we discussed in class) can be issued to DRAM with the given latencies:

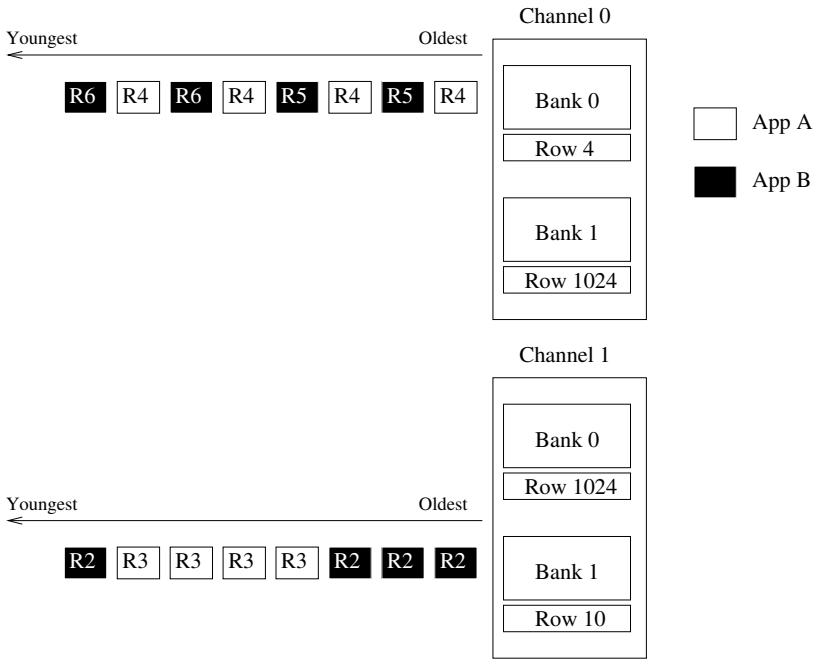
Command	Latency
ACTIVATE	15 ns
PRECHARGE	15 ns
READ/WRITE	15 ns

Assume the bus latency is 0 cycles.

(a) Two applications A and B are run on the machine. The following is a snapshot of the request buffers at time t_0. Requests are tagged with the index of the row they are destined to. Additionally, requests of applications A and B are indicated with different colors. Row 4 is initially open in bank 0 of channel 0 and row 10 is initially open in bank 1 of channel 1.

Consider that all requests at channel 0 are mapped to bank 0. And all accesses at channel 1 are mapped to bank 1.

Each application is stalled until all of its memory requests are serviced and does not generate any more requests.



- What is the stall time of application A using an FR-FCFS scheduling policy?
- What is the stall time of application B using an FR-FCFS scheduling policy?

(b) Consider a simple channel partitioning scheme where application A's data is mapped to channel 0 and application B's data is mapped to channel 1. When data is mapped to a different channel, only the channel number changes; the bank number does not change. For instance, requests of application A that were mapped to bank 1 of channel 1 would now be mapped to bank 1 of channel 0.

- What is the stall time of application A using this channel partitioning mechanism and an FR-FCFS memory scheduler?
- What is the stall time of application B using this channel partitioning mechanism and an FR-FCFS memory scheduler?

3 Handin

You should electronically hand in your homework (in pdf format) to Collab.