2. Memory Scheduling [40 points]

We have a byte-addressable processor that is connected to a single memory channel that has a single rank of DRAM. The physical address space is 32 bits, and the processor uses the following mapping shown in Table 1 to index the DRAM. Each DRAM row has a certain number of columns, where a column has the same size of a cache line. The processor uses 64-byte cache lines.

MSB			LSB
Rows	Banks	Columns	Cache Line Offset

Table 1: Mapping from the physical address to DRAM.

(a) Table 2 shows the snapshot of the memory request queue that has 6 pending memory requests. Assume all the rows are closed when the snapshot was taken. With the FR-FCFS scheduling policy, issuing these 6 requests results in a row buffer hit rate of 1/3 (i.e., 1/3 of the requests retrieve data from the row buffers directly).

Request	Physical Address
A (oldest)	0x0000_0000
B	0x0000_1000
C	0x0000_4000
D	0x0000_0040
E	0x0000_0800
F (youngest)	0x0010_0040

Table 2: The state of the memory request queue.

What is the row size in KB? Show your work.

Since the hit rate is 1/3, there are two row hit requests. With the amount of information we have now, we know that there are 6 bits for the cache line offset. So requests A and D must go to the same row. A accesses row0, bank0, and col0. D accesses the same row and bank, but it goes to col1. Since A results in a miss and D is a hit, we need to find another hit request. The best candidate is request E.
By observing the bit-pattern difference between B and E, one can find out that the column bits end at the 12th bit, thus there are 6 bits for the columns as there are 6 bits for the cache line offset. Therefore, the row size is $2^{12} = 4KB$.
Summary: B goes to bank 1, C goes to bank 0 (row 1), and F goes to bank 0 (row 64). Therefore, B, C, and F are row misses.

(b) Table 3 shows the memory request queue that has 4 pending memory requests at time 0. Assume the following:

- A row buffer hit takes **50 cycles**.
- A row buffer conflict takes **250 cycles**.
- Requests going to different banks can be processed by the banks in parallel.
- All the row buffers are closed at time 0.
- The controller cannot issue two requests at the same time. Each request takes **10** cycles to process, so it takes 10 cycles between issuing two separate requests to the memory.
- The controller employs FR-FCFS scheduling policy.

Request	Physical Address
A (oldest)	0x0000_4000
B	0x0000_1040
C	0x0000_3040
D (youngest)	0x0000_4a00

Table 3: The state of the memory request queue at time 0.

If it takes **320 cycles** to finish processing all four requests in the memory, **at least** how many banks does this rank of DRAM have? Show your work.

320 = 10 + 250 (miss) + 10 + 50 (hit) So there must be one bank serving one miss and one hit, which is on the critical path. In parallel, there are two other banks serving requests.
With the information from part a, we know that A and D go to the same bank and same row. Therefore, B and C must go to two different banks. If they were to the same bank, that would incur 510 cycles of latency. By comparing the bit patterns between A and C, one can find that there are two bits for the banks. Thus, there are at least four banks in the rank.