CMU 18-447 Introduction to Computer Architecture, Spring 2012

Handout 12/ HW 6: Caches and Main Memory

Prof. Onur Mutlu, Instructor
Chris Fallin, Lavanya Subramanian, Abeer Agrawal, TAs

Given: Wednesday, Apr 4, 2012
Due: **Monday, Apr 16, 2012**

1 Caches and Virtual Memory

A 2-way set associative write back cache with perfect LRU replacement requires 15×2^9 bits of storage to implement its tag store (including bits for valid, dirty and LRU). The cache is virtually indexed, physically tagged. The virtual address space is 1 MB, page size is 2 KB, cache block size is 8 bytes.

(a) What is the size of the data store of the cache in bytes?

(b) How many bits of the virtual index come from the virtual page number?

(c) What is the physical address space of this memory system?

2 Interleaving and Row-Buffer Locality

A machine has a main memory of 4 KB, organized as 1 channel, 1 rank and N banks (where N > 1). The system does not have virtual memory.

- Data is interleaved using a cache block interleaving policy, as described in lecture, where consecutive cache blocks are placed on consecutive banks.
- The size of a cache block is 32 bytes. Size of a row is 128 bytes.
- 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.
- A row-buffer hit is an access to a row that is present in the row-buffer. A row-buffer miss is an access to a row that is not present in the row-buffer.

(a) For a program executing on the machine, accesses to the following bytes miss in the on-chip caches and go to memory.

0, 32, 320, 480, 4, 36, 324, 484, 8, 40, 328, 488, 12, 44, 332, 492
The row-buffer hit rate is 0%, i.e., all accesses miss in the row-buffer.

What is the minimum value of N - the number of banks?

(b) If the row-buffer hit rate for the same sequence were 75%, what would be minimum value of N - the number of banks?

(c) i) Could the row-buffer hit rate for the sequence be 100%? Why or why not? Explain.
ii) If yes, what is the minimum number of banks required to achieve a row-buffer hit rate of 100%?

3 Memory Scheduling

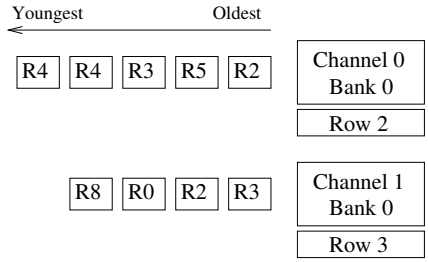
A machine has a DRAM main memory of 4 KB, organized as 2 channels, 1 rank/channel and 4 banks/rank. The system does not have virtual memory.

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.

Following are the commands issued to DRAM, to access data.

- ACTIVATE: Loads the row (that needs to be accessed) into the bank's row-buffer. This is called opening a row. (Latency: 15ns)
- PRECHARGE: Restores the contents of the banks row-buffer back into the row. This is called closing a row. (Latency: 15ns)
- READ/WRITE: Accesses data from the row-buffer. (Latency: 15ns)

(a) Application A runs alone on this machine. The following figure shows a snapshot of the request buffers at time t. Each request is tagged with the index of the row it is destined to. Row 2 is currently open in bank 0 of channel 0 and row 3 is currently open in bank 0 of channel 1.

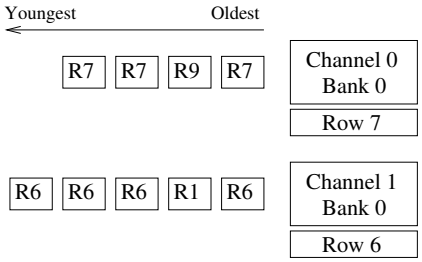


Application A is stalled until all of these memory requests are serviced and does not generate any more requests.

What is the stall time of application A using i) an FCFS scheduling policy and ii) an FR-FCFS scheduling policy?

(b) Now, application B runs alone on this machine. The following figure shows a snapshot of the request buffers at time t. Each request is tagged with the index of the row it is destined to. Row 7 is currently open in bank 0 of channel 0 and row 6 is currently open in bank 0 of channel 1.

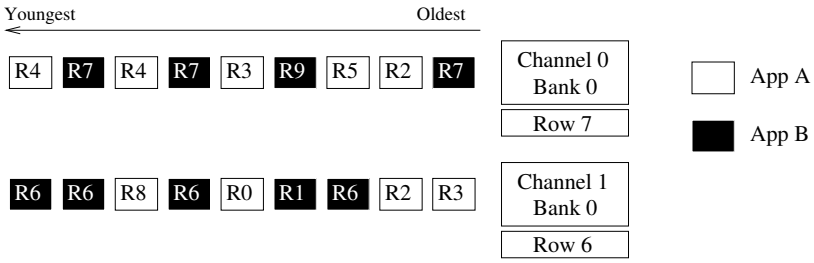
Application B is stalled until all of these memory requests are serviced and does not generate any more requests.



What is the stall time of application B using i) an FCFS scheduling policy and ii) an FR-FCFS scheduling policy?

(c) Applications A and B are run on the same machine. The following is a snapshot of the request buffers at time t. 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 7 is currently open in bank 0 of channel 0 and row 6 is currently open in bank 0 of channel 1.

An 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 i) an FCFS scheduling policy and ii) an FR-FCFS scheduling policy?
What is the stall time of application B using i) an FCFS scheduling policy and ii) an FR-FCFS scheduling policy?

4 Main Memory Potpourri

A machine has a 4 KB DRAM main memory system. Each row is refreshed every 64 ms.

(a) The machine's designer runs two applications A and B (each run alone) on the machine. Although applications A and B have a similar number of memory requests, application A spends a surprisingly larger fraction of cycles stalling for memory than application B does? What might be the reasons for this?

(b) Application A also consumes a much larger amount of memory energy than application B does. What might be the reasons for this?

(c) When applications A and B are run together on the machine, application A's performance degrades significantly, while application B's performance doesn't degrade as much. Why might this happen?

(d) The designer decides to use a smarter policy to refresh the memory. A row is refreshed only if it has not been accessed in the past 64 ms. Do you think this is a good idea? Why or why not?

(e) The refresh energy consumption when application B is run, drops significantly when this new refresh policy is applied, while the refresh energy when application A is run reduces only slightly. Is this possible? Why or why not?