

COMPUTER ARCHITECTURE (263-2210-00L), FALL 2019

HW 4: MEMORY INTERFERENCE AND QOS

Instructor: Prof. Onur Mutlu

TAs: Mohammed Alser, Rahul Bera, Geraldo Francisco De Oliveira Junior, Can Firtina, Juan Gomez Luna, Jawad Haj-Yahya, Hasan Hassan, Konstantinos Kanellopoulos, Jeremie Kim, Nika Mansouri Ghiasi, Lois Orosa Nogueira, Jisung Park, Minesh Hamenbhai Patel, Abdullah Giray Yaglikci

Assigned: Monday, Nov 18, 2019

Due: **Sunday, Dec 1, 2019**

- **Handin - Critical Paper Reviews (1).** You need to submit your reviews to <https://safari.ethz.ch/review/architecture19/>. Please check your inbox. You should have received an email with the password you can use to login to the paper review system. If you have not received any email, please contact comparch@lists.ethz.ch. In the first page after login, you should click in "Architecture - Fall 2019 Home", and then go to "any submitted paper" to see the list of papers.
- **Handin - Questions (2-5).** Please upload your solution to the Moodle (<https://moodle-app2.let.ethz.ch/>) as a single PDF file. **Please use a typesetting software (e.g., LaTeX) or a word processor (e.g., MS Word, LibreOfficeWriter) to generate your PDF file. Feel free to draw your diagrams either using an appropriate software or by hand, and include the diagrams into your solutions PDF.**

1 Critical Paper Reviews [400 points]

Please read the guidelines for reviewing papers and check the sample reviews. You may access them by *simply clicking on the QR codes below or scanning them*. We will give out extra credit that is worth 0.5% of your total grade for each good review.



Guidelines



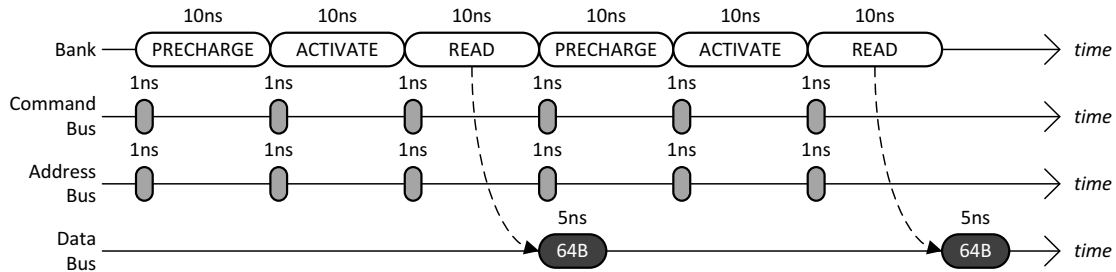
Sample reviews

Write an approximately one-page critical review for each of the following papers. A review with bullet point style is more appreciated. Try not to use very long sentences and paragraphs. Keep your writing and sentences simple. Make your points bullet by bullet, as much as possible.

- Veynu Narasiman, Chang Joo Lee, Michael Shebanow, Rustam Miftakhutdinov, Onur Mutlu, and Yale N. Patt, "Improving GPU Performance via Large Warps and Two-Level Warp Scheduling" MICRO 2011. https://people.inf.ethz.ch/omutlu/pub/large-gpu-warps_micro11.pdf
- O. Mutlu and T. Moscibroda, "Parallelism-Aware Batch Scheduling: Enhancing both Performance and Fairness of Shared DRAM Systems," ISCA 2008. https://people.inf.ethz.ch/omutlu/pub/parbs_isca08.pdf
- Ebrahimi E, Lee CJ, Mutlu O, Patt YN. "Fairness via source throttling: a configurable and high-performance fairness substrate for multi-core memory systems", ASPLOS 2010 https://people.inf.ethz.ch/omutlu/pub/rlmc_isca08.pdf
- Engin Ipek, Onur Mutlu, José F. Martínez, and Rich Caruana, "Self Optimizing Memory Controllers: A Reinforcement Learning Approach" ISCA 2008. https://people.inf.ethz.ch/omutlu/pub/rlmc_isca08.pdf

2 Memory Interference and QoS [100 points]

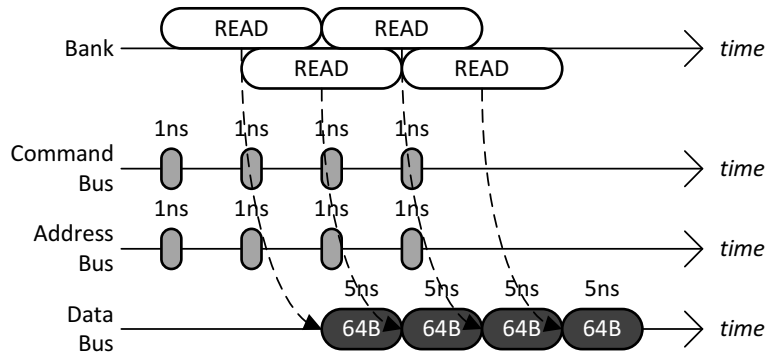
Row-Buffer Conflicts. The following timing diagram shows the operation of a single DRAM channel and a single DRAM bank for two back-to-back reads that conflict in the row-buffer. Immediately after the bank has been busy for 10ns with a READ, data starts to be transferred over the data bus for 5ns.



- (a) Given a long sequence of back-to-back reads that always conflict in the row-buffer, what is the data throughput of the main memory system? Please state your answer in **gigabytes/second**.

- (b) To increase the data throughput, the main memory designer is considering adding more DRAM banks to the single DRAM channel. Given a long sequence of back-to-back reads to all banks that always conflict in the row-buffers, what is the minimum number of banks that is required to achieve the maximum data throughput of the main memory system?

Row-Buffer Hits. The following timing diagram shows the operation of the single DRAM channel and the single DRAM bank for four back-to-back reads that hit in the row-buffer. It is important to note that row-buffer hits to the same DRAM bank are pipelined: while each READ keeps the DRAM bank busy for 10ns, up to at most **half** of this latency (5ns) can be overlapped with another read that hits in the row-buffer.



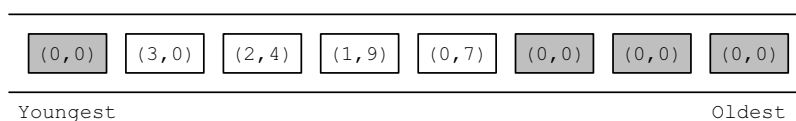
(c) Given a long sequence of back-to-back reads that always hits in the row-buffer, what is the data throughput of the main memory system? Please state your answer in **gigabytes/second**.

(d) When the maximum data throughput is achieved for a main memory system that has a single DRAM channel and a single DRAM bank, what is the bottleneck that prevents the data throughput from becoming even larger? **Circle** all that apply.

BANK
COMMAND BUS
ADDRESS BUS
DATA BUS

Memory Scheduling Policies. The diagram below shows the memory controller's *request queue* at time 0. The shaded rectangles are read requests generated by thread T_0 , whereas the unshaded rectangles are read requests generated by thread T_1 . Within each rectangle, there is a pair of numbers that denotes the request's (*BankAddress*, *RowAddress*). Assume that the memory system has a **single** DRAM channel and four DRAM banks. Further assume the following.

- All the row-buffers are **closed** at time 0.
- Both threads start to stall at time 0 because of memory.
- A thread continues to stall until it receives the data for all of its requests.
- Neither thread generates more requests.



(f) For the *FCFS* scheduling policy, calculate the memory stall time of $T0$ and $T1$.

T0:

T1:

(g) For the *RR – FCFS* scheduling policy, calculate the memory stall time of $T0$ and $T1$.

T0:

T1:

- (h) For the *PAR* – *BS* scheduling policy, calculate the memory stall time of $T0$ and $T1$. Assume that all eight requests are included in the same batch.

T0:

T1:

3 Memory Scheduling [50 points]

In class, we covered "parallelism-aware batch scheduling," which is a memory scheduling algorithm that aims to reduce interference between threads in a multi-core system.

- (a) What benefit does request batching provide in this algorithm?

- (b) How does the algorithm preserve intra-thread bank parallelism?

- (c) If thread ranking was formed in a "random manner" (i.e., threads were assigned a random rank), would each thread's parallelism be preserved? Why or why not? Explain.

4 Memory System [80 points]

A machine with a 4 GB DRAM main memory system has 4 channels, 1 rank per channel and 4 banks per rank. The cache block size is 64 bytes.

(a) You are given the following byte addresses and the channel and bank to which they are mapped:

Byte: 0x0000 ⇒ Channel 0, Bank 0
 Byte: 0x0100 ⇒ Channel 0, Bank 0
 Byte: 0x0200 ⇒ Channel 0, Bank 0
 Byte: 0x0400 ⇒ Channel 1, Bank 0
 Byte: 0x0800 ⇒ Channel 2, Bank 0
 Byte: 0x0C00 ⇒ Channel 3, Bank 0
 Byte: 0x1000 ⇒ Channel 0, Bank 1
 Byte: 0x2000 ⇒ Channel 0, Bank 2
 Byte: 0x3000 ⇒ Channel 0, Bank 3

Determine which bits of the address are used for each of the following address components. Assume row bits are higher order than column bits:

- Byte on bus
Addr [2 : 0]
- Channel bits (channel bits are contiguous)
Addr [_____ : _____]
- Bank bits (bank bits are contiguous)
Addr [_____ : _____]
- Column bits (column bits are contiguous)
Addr [_____ : _____]
- Row bits (row bits are contiguous)
Addr [_____ : _____]

(b) Two applications App 1 and App 2 share this memory system (using the address mapping scheme you determined in part (a)). The memory scheduling policy employed is FR-FCFS. The following requests are queued at the memory controller request buffer at time t . Assume the first request (A) is the oldest and the last one ($A + 15$) is the youngest.

A B A + 1 A + 2 A + 3 B + 10 A + 4 B + 12 A + 5 A + 6 A + 7
 A + 8 A + 9 A + 10 A + 11 A + 12 A + 13 A + 14 A + 15

These are cache block addresses, not byte addresses. Note that requests to $A + x$ are from App 1, while requests to $B + x$ are from App 2. Addresses A and B are row-aligned (i.e., they are at the start of a row) and are at the same bank but are in different rows.

Assuming row-buffer hits take T time units to service and row-buffer conflicts/misses take $2T$ time units to service, what is the slowdown (compared to when run alone on the same system) of

- App 1?

- App 2?

(c) Which application slows down more?

Why?

(d) In class, we discussed memory channel partitioning and memory request scheduling as two solutions to mitigate interference and application slowdowns in multicore systems. Propose another solution to reduce the slowdown of the more-slowed-down application, without increasing the slowdown of the other application? Be concrete.

5 Vector Processing [200 points]

You are studying a program that runs on a vector computer with the following latencies for various instructions:

- VLD and VST: 50 cycles for each vector element; fully interleaved and pipelined.
- VADD: 4 cycles for each vector element (fully pipelined).
- VMUL: 16 cycles for each vector element (fully pipelined).
- VDIV: 32 cycles for each vector element (fully pipelined).
- VRSHF (right shift): 1 cycle for each vector element (fully pipelined).

Assume that:

- The machine has an in-order pipeline.
 - The machine supports chaining between vector functional units.
 - In order to support 1-cycle memory access after the first element in a vector, the machine interleaves vector elements across memory banks. All vectors are stored in memory with the first element mapped to bank 0, the second element mapped to bank 1, and so on.
 - Each memory bank has an 8 KB row buffer. Vector elements are 64 bits in size.
 - Each memory bank has two ports (so that two loads/stores can be active simultaneously), and there are two load/store functional units available.
- (a) What is the minimum power-of-two number of banks required in order for memory accesses to never stall? (Assume a vector stride of 1.)

- (b) The machine (with as many banks as you found in part a) executes the following program (assume that the vector stride is set to 1):

```
VLD V1 = A
VLD V2 = B
VADD V3 = V1, V2
VMUL V4 = V3, V1
VRSHF V5 = V4, 2
```

It takes 111 cycles to execute this program. What is the vector length?

If the machine did not support chaining (but could still pipeline independent operations), how many cycles would be required to execute the same program?

- (c) The architect of this machine decides that she needs to cut costs in the machine's memory system. She reduces the number of banks by a factor of 2 from the number of banks you found in part (a) above. Because loads and stores might stall due to bank contention, an arbiter is added to each bank so that pending loads from the oldest instruction are serviced first. How many cycles does the program take to execute on the machine with this reduced-cost memory system (but with chaining)?

Now, the architect reduces cost further by reducing the number of memory banks (to a lower power of 2). The program executes in 279 cycles. How many banks are in the system?

- (d) Another architect is now designing the second generation of this vector computer. He wants to build a multicore machine in which 4 vector processors share the same memory system. He scales up the number of banks by 4 in order to match the memory system bandwidth to the new demand. However, when he simulates this new machine design with a separate vector program running on every core, he finds that the average execution time is longer than if each individual program ran on the original single-core system with 1/4 the banks. Why could this be? Provide concrete reason(s).

What change could this architect make to the system in order to alleviate this problem (in less than 20 words), while only changing the shared memory hierarchy?

6 GPUs and SIMD [100 points]

We define the *SIMD utilization* of a program that runs on a GPU as the fraction of SIMD lanes that are kept busy with *active threads* during the run of a program. As we saw in lecture, the SIMD utilization of a program is computed across the *complete run* of the program.

The following code segment is run on a GPU. Each thread executes **a single iteration** of the shown loop. Assume that the data values of the arrays **A** and **B** are already in vector registers so there are no loads and stores in this program. (Hint: Notice that there are 3 instructions in each iteration.) A warp in the GPU consists of 32 threads, and there are 32 SIMD lanes in the GPU.

```
for (i = 0; i < 1025; i++) {
    if (A[i] < 33) {          // Instruction 1
        B[i] = A[i] << 1;    // Instruction 2
    }
    else {
        B[i] = A[i] >> 1;    // Instruction 3
    }
}
```

Please answer the following six questions.

- (a) How many warps does it take to execute this program?

- (b) What is the maximum possible SIMD utilization of this program? (Hint: The warp scheduler does not issue instructions where no threads are active).

- (c) Please describe what needs to be true about array **A** to reach the maximum possible SIMD utilization asked in part (b). (Please cover all cases in your answer.)

- (d) What is the minimum possible SIMD utilization of this program?

- (e) Please describe what needs to be true about array **A** to reach the minimum possible SIMD utilization asked in part (d). (Please cover all cases in your answer.)

- (f) What is the SIMD utilization of this program if $A[i] = i$? Show your work.

7 BossMem [50 points]

A researcher has developed a new type of nonvolatile memory, BossMem. He is considering BossMem as a replacement for DRAM. BossMem is 10x faster (all memory timings are 10x faster) than DRAM, but since BossMem is so fast, it has to frequently power-off to cool down. Overheating is only a function of time, not a function of activity. An idle stick of BossMem has to power-off just as frequently as an active stick. When powered-off, BossMem retains its data, but cannot service requests. Both DRAM and BossMem are banked and otherwise architecturally similar. To the researcher's dismay, he finds that a system with 1GB of DRAM performs considerably better than the same system with 1GB of BossMem.

- (i) What can the researcher change or improve in the core (he can't change BossMem or anything beyond the memory controller) that will make his BossMem perform more favorably compared to DRAM, realizing that he will have to be fair and evaluate DRAM with his enhanced core as well? (15 words or less)

- (ii) A colleague proposes he builds a hybrid memory system, with both DRAM and BossMem. He decides to place data that exhibits low row buffer locality in DRAM and data that exhibits high row buffer locality in BossMem. Assume 50% of requests are row buffer hits. Is this a good or bad idea? Show your work.

- (iii) Now a colleague suggests trying to improve the last-level cache replacement policy in the system with the hybrid memory system. Like before, he wants to improve the performance of this system relative to one that uses just DRAM and he will have to be fair in his evaluation. Can he design a cache replacement policy that makes the hybrid memory system look more favorable? In 15 words or less, justify NO or describe a cache replacement policy that would improve the performance of the hybrid memory system more than it would DRAM.

- (iv) In class we talked about another nonvolatile memory technology, phase-change memory (PCM). Which technology, PCM, BossMem, or DRAM requires the greatest attention to security? What is the vulnerability?

(v) Which is likely of least concern to a security researcher?