**Homework Assignment #2**

***Due Date: 10/11, Tuesday, 11:59 p.m. Please submit via Blackboard. Late submissions are accepted till 10/16, 11:59 p.m, with 10% penalty each day. For all questions, please note that you need to show the steps how you obtain your result and please do NOT just provide the final answer.***

***Please name your submission file starting with “LastName\_FirstName\_HW2”.***

**Q1.** **(6 points)** Please briefly explain: 1) what the “memory wall” challenge refers to in computer architecture; and 2) why do we need the memory hierarchy design.

**Q2.** **(6 points)** Please briefly explain what the temporal locality and spatial locality are. Please give an example for each of them, respectively.

**Q3.** **(10 points)** Assume we have a direct-mapped cache. The cache size is 2n blocks (thus *n* bits are used for the index), and the block size is 2m words (2m+2 bytes). Assume we have 32-bit addresses.

a. What is the size (the number of bits) of each tag field?

b. What is the total number of bits needed for this direct-mapped cache, including valid field, tag field, and data field?

**Q4. (15 points)** Caches are important to providing a high-performance memory hierarchyto processors. Below is a list of 32-bit memory address references, given as **word** addresses (as a word is 4 bytes, the word addresses are byte addresses shifted right by 2 bits, i.e., byte addresses divided by 4)

0x03, 0xb4, 0x2b, 0x02, 0xbf, 0x58, 0xbe, 0x0e, 0xb5, 0x2c, 0xba, 0xfd

a. For each of these references, identify the binary address, the tag, and the index given a **direct-mapped cache with** **16 one-word blocks**. Also list if each reference is a hit or a miss, assuming the cache is initially empty.

b. For each of these references, identify the binary address, the tag, and the index given a **direct-mapped cache with two-word blocks** and **a total size of 8 blocks**. Also list if each reference is a hit or a miss, assuming the cache is initially empty.

**Q5. (12 points)** Using the sequence of references from Q4, show the final cache contents for a three-way set associative cache with two-word blocks (i.e., the block size is 8 bytes) and a total size of 24 blocks. Use the LRU (least-recently used) replacement policy. For each reference, identify the index bits, the tag bits, and if it is a hit or a miss.

**Q6. (12 points)** Assume that main memory accesses take 70 ns and that memory accesses are 36% of all instructions. The following table shows data for L1 caches attached to each of two processors, P1 and P2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | L1 size | L1 miss rate | L1 hit time |
| P1 | 2 KiB | 8.0% | 0.66ns |
| P2 | 4 KiB | 6.0% | 0.90ns |

a. What is the Average Memory Access Time for P1 and P2?

b. Assuming that the L1 hit time determines the cycle times for P1 and P2, i.e., the clock rates are 1.52GHz and 1.11 GHz for P1 and P2, respectively. Assuming a base CPI of 1.0 without any memory stalls, what is the total CPI for P1 and P2, respectively?

c. Assuming P1 now is added with an additional L2 cache, with the following table shows data for this L2 cache, what is the AMAT for P1 with the addition of an L2 cache?

|  |  |  |
| --- | --- | --- |
| L2 size | L2 miss rate | L2 hit time |
| 1 MiB | 20% | 5.62ns |

**Q7. (10 points)** Given three variables: the cache capacity, cache block size, and cache associativity, and assume one variable changes and the other two are fixed, please fill in the below table with “increase”, “decrease”, or “same”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Compulsory misses | Capacity Misses | Conflict misses | Hit time | Miss penalty |
| Larger block size, same cache capacity, same associativity |  |  |  |  |  |
| Larger cache capacity, same block size,  same associativity |  |  |  |  |  |
| Higher associativity, same cache capacity, same block size |  |  |  |  |  |

**Q8. (6 points)** Please list two sample compiler optimization techniques that can reduce the cache miss rate and also briefly explain why.

**Q9. (8 points)** Please explain what virtual memory is. Please briefly explain these terminologies: virtual address, physical address, page, page fault, page table, and Translation Look-aside Buffer (TLB).

**Q10. (15 points)** Please pick one of the papers listed below (the paper PDF file can be downloaded from Blackboard), read the paper in detail, and write a short summary of the paper you studied. Please limit to maximum 400 words, and please focus on what problem is studied in the paper and what are the key conclusions.

[1] J. Leidel and Y. Chen. HMC-Sim-2.0: A Co-Design Infrastructure for Exploring Custom Memory Cube Operations. The International Journal of Parallel Computing (ParCo), Volume: 68, Pages: 77 - 88, 2017.

[2] J. Leidel and Y. Chen. HMC-SIM: A Simulation Framework for Hybrid Memory Cube Devices. Journal of Parallel Processing Letters, Volume: 24, Issue: 04, Pages: 1465 - 1474, December 2014.

[3] W. Xie, Y. Chen and P. Roth. Parallel-DFTL: A Flash Translation Layer that Exploits Internal Parallelism in Solid State Drives. In Proceedings of the 11th IEEE International Conference on Networking, Architecture, and Storage (NAS'16), Pages: 1 - 10, 2016. DOI Best Paper Nominee.

[4] K. Zhang, Z. Wang, Y. Chen, H. Zhu and X.-H. Sun. PAC-PLRU: A Cache Replacement Policy to Salvage Discarded Predictions from Hardware Prefetchers. In the Proc. of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'11), 2011.

THE END.