CS M151B Homework 7

Charles Zhang

February 20, 2022

Problem 5.5

For a direct-mapped cache design with a 32-bit address, the following bits of the address are used to access the cache:

Tag	Index	Offset
31-10	9-5	4-0

a) What is the cache block size (in words)?

b) How many entries does the cache have?

c) What is the ratio between total bits required for such a cache implementation over the data storage bits?

Beginning from power on, the following byte-addressed cache references are recorded:

Hex	00	04	10	84	E8	A0	400	1E	8C	C1C	B4	884
Dec	0	4	16	132	232	160	1024	30	140	3100	180	2180

d) For each reference, list (1) its tag, index, and offset, (2) whether it is a hit or a miss, and (3) which bytes were replaced (if any).

e) What is the hit ratio?

f) List the final state of the cache, with each valid entry represented as a record of `<index, tag, data>`.

Problem 5.10

In this exercise, we will look at the different ways capacity affects overall performance. In general, cache access time is proportional to capacity. Assume that main memory accesses take 70ns 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`	2KB	8.0%	0.66ns
`P2`	4KB	6.0%	0.90ns

a) Assuming that the L1 hit time determines the cycle times for `P1` and `P2`, what are their respective clock rates?

b) What is the Average Memory Access Time for `P1` and `P2`?

c) Assuming a base CPI of `1.0` without any memory stalls, what is the total CPI for `P1` and `P2`? Which processor is faster? (When we say a "base CPI of `1.0`", we mean that instructions complete in one cycle, unless either the instruction access or the data access causes a cache miss.)

For the next three problems, we will consider the addition of an L2 cache to `P1` to presumably make up for its limited L1 cache capacity. Use the L1 cache capacities and hit times from the previous table when solving these problems. The L2 miss rate indicated is its local miss rate.

L2 Size	L2 Miss Rate	L2 Hit Time
1MB	95%	5.62ns

d) What is the AMAT for `P1` with the addition of an L2 cache? Is the AMAT better or worse with the L2 cache?

e) Assuming a base CPI of `1.0` without any memory stalls, what is the total CPI for `P1` with the addition of an L2 cache?

f) What would the L2 miss rate need to be in order for `P1`| with an L2 cache to be faster than `P1` without an L2 cache?

g) What would the L2 miss rate need to be in order for `P1` with an L2 cache to be faster than `P2` without an L2 cache?

Problem 5.12

Multilevel caching is an important technique to overcome the limited amount of space that a first level cache can provide while still maintaining its speed. Consider a processor with the following parameters:

Base CPI, No Memory Stalls	1.5
Processor Speed	2GHz
Main Memory Access Time	100ns
First Level Cache Miss Rate per Instruction	7%
Second Level Cache, Direct-Mapped Speed	12 cycles
Global Miss Rate with Second Level Cache, Direct-Mapped	3.5%
Second Level Cache, Eight-Way Set Associative Speed	28 cycles
Global Miss Rate with Second Level Cache, Eight-Way Set Associative	1.5%

a) Calculate the CPI for the processor in the table using: 1) only a first level cache, 2) a second level direct-mapped cache, and 3) a second level eight-way set associative cache. How do these numbers change if main memory access time is doubled? (Give each change as both an absolute CPI and a percent change.) Notice the extent to which an L2 cache can hide the effects of a slow memory.

b) It is possible to have an even greater cache hierarchy than two levels. Given the processor above with a second level, direct-mapped cache, a designer wants to add a third level cache that takes 50 cycles to access and will have a 13% miss rate. Would this provide better performance? In general, what are the advantages and disadvantages of adding a third level cache?

c) In older processors such as the Intel Pentium or Alpha 21264, the second level of cache was external (located on a different chip) from the main processor and the first level cache. While this allowed for large second level caches, the latency to access the cache was much higher, and the bandwidth was typically lower because the second level cache ran at a lower frequency. Assume a 512 KB off-chip second level cache has a global miss rate of 4%. If each additional 512 KB of cache lowered global miss rates by 0.7%, and the cache had a total access time of 50 cycles, how big would the cache have to be to match the performance of the second level direct-mapped cache listed above?