ELEC 5200-001/6200-001 (Spring 2019)
Homework 5 - Solution
Assigned 03/25/19, due 04/01/19
(35 points possible)

Problem 1: Consider a 16-bit address space with the following list of memory addresses, given as word addresses.

3, 180, 43, 2, 191, 88, 190, 14, 181, 44, 186, 253

(a) For each address given, provide the full 16-bit binary memory **byte** address along with the binary tag(s) and index for a direct-mapped cache with 16 one word blocks. Assuming these addresses are accessed sequentially as listed above, determine if each memory access is a hit or miss assuming the cache is initially empty. (4 points)

Word	16-Bit Byte Address	Tag	Index	Access
3	0000 0000 0000 1100	0000 0000 00	0011	Miss
180	0000 0010 1101 0000	0000 0010 11	0100	Miss
43	0000 0000 1010 1100	0000 0000 10	1011	Miss
2	0000 0000 0000 1000	0000 0000 00	0010	Miss
191	0000 0010 1111 1100	0000 0010 11	1111	Miss
88	0001 0010 0110 0000	0001 0010 01	1000	Miss
190	0000 0010 1111 1000	0000 0010 11	1110	Miss
14	0000 0000 0011 1000	0000 0000 00	1110	Miss
181	0000 0010 1101 0100	0000 0010 11	0101	Miss
44	0000 0000 1011 0000	0000 0000 10	1100	Miss
186	0000 0010 1110 1000	0000 0010 11	1010	Miss
253	0000 0011 1111 0100	0000 0011 11	1101	Miss

(b) For each of the above memory references, provide the full 16-bit binary memory **byte** address along with the binary tag(s) and index for a direct-mapped cache with 8 two-word blocks. Assuming these addresses are again accessed sequentially, determine if each memory access is a hit or miss assuming the cache is initially empty. (4 points)

Word	16-Bit Byte Address	Tag	Index	Access
3	0000 0000 0000 1100	0000 0000 00	001	Miss
180	0000 0010 1101 0000	0000 0010 11	010	Miss
43	0000 0000 1010 1100	0000 0000 10	101	Miss
2	0000 0000 0000 1000	0000 0000 00	001	Hit
191	0000 0010 1111 1100	0000 0010 11	111	Miss
88	0001 0010 0110 0000	0001 0010 01	100	Miss
190	0000 0010 1111 1000	0000 0010 11	111	Hit
14	0000 0000 0011 1000	0000 0000 00	111	Miss
181	0000 0010 1101 0100	0000 0010 11	010	Hit
44	0000 0000 1011 0000	0000 0000 10	110	Miss
186	0000 0010 1110 1000	0000 0010 11	101	Miss
253	0000 0011 1111 0100	0000 0011 11	110	Miss

Problem 2:

(a) A one-level cache system contains a SRAM cache that is n times faster than the main memory. The cache has an access time T and its miss rate is m. Show that this system will provide an average memory access speed up of n/(1 + mn). Irrespective of how fast the cache hardware is, show that the speed up has an upper bound 1/m. (4 points)

Answer:

Average access time for a one-level cache system is given by,

$$t = T + (1 - h)\, nT = T + mnT$$

where nT and T are cycle times of the main memory and cache, respectively, and h = 1 – m is the hit rate.

Therefore,

$$\text{Speed up} = \frac{nT}{t} = \frac{nT}{T + mnT} = \frac{n}{1 + mn}$$

If we let the cache be infinitely faster than the main memory, then

$$\text{Speed up} = \lim_{n \to \infty} \frac{n}{1 + mn} = \lim_{n \to \infty} \frac{1}{(1/n) + m} = 1/m$$

(b) For a two-level cache, show that the average memory access speed up has an upper bound $1/(m_1 \times m_2)$, given that m_1 and m_2 are the miss rates of L1 and L2 caches, respectively. (4 points)

Answer:

Suppose L2 cache hardware is k times ($1 < k < n$) slower than L1 hardware whose access time is T. Then the average access time for a two-level cache system is,

$$t = T + m_1\,[kT + m_2\, n\, T]$$

Therefore,

$$\text{Speed up} = \frac{nT}{t} = \frac{nT}{T + m_1\,[kT + m_2\, n\, T]} = \frac{1}{1/n + m_1\, k/n + m_1\, m_2}$$

Substituting $n \to \infty$ in the last expression, we get speed up $\to 1/(m_1\, m_2)$.

Problem 3: For a direct mapped cache design with a 32-bit byte address, the following bits are used to access the cache:

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

(a) What is the cache block size (in words)? (2 points)

Answer: 2 bits for the byte offset leaves 3 bits for the block offset resulting in 2^3 or 8 words per block.

(b) How many entries does the cache have? (2 points)

Answer: 5 index bits results in a cache with $2^5 = 32$ entries

(c) Assuming a 32 bit data word, what is the ratio of total bits required to implement the cache to the number of data storage bits? (2 points)

Answer:

Our formula for bits required to implement a cache is
#blocks in cache × *(data bits/block + tag size + valid bits)* which yields
32 * (8 * 32 + 22 + 1) = 8928 bits,
therefore the ratio is:

8928 / (32 * 8 * 32) = 8928/8192 = 1.0898

Problem 4: To satisfy the architectural requirement of a computer system with a one-level cache the hit rate must be 95%. An early prototype for cost reasons uses a SRAM of limited capacity as a one-level cache that only provides a 70% hit rate. The main memory is 70 times slower than the SRAM cache.

(a) Find the average data access time for the one-level cache with 70% hit rate, expressed in terms of the cycle time T1 for the cache. Show that the data access will become almost five times faster if the hit rate could be raised to 95%. (4 points)

Answer:

Given, L1 cache SRAM has cycle time T1 and the main memory cycle time is Tm = 70T1. Then,

Average data access time, $T(h) = T1 + (1 - h)Tm$
For h = 0.70, $T(0.70) = T1 + 0.30Tm = T1(1 + 0.30\times70) = 22T1$
For h = 0.95, $T(0.95) = T1 + 0.05Tm = T1(1 + 0.05\times70) = 4.5T1$

Therefore, access time ratio, T(0.70)/T(0.95) = 22/4.5 = 4.89 ≈ 5

(b) Suppose a level-2 cache brings the data access time to the required value when the cycle time of the L2 cache is 1.5T1. Determine the minimum hit rate for the L2 cache. (4 points)

Answer:

We have, T1 = T2/1.5 = Tm/70. We design a two-level cache in which an SRAM L1 cache has a hit rate h1 = 0.70, and an L2 cache has hit rate h2, such that the total access time is not higher than the access time of 4.5T1 calculated above. Thus,

Two-level cache access time	=		T1 + (1 – h1) [T2 + (1 – h2) Tm]
	T1 + 0.05Tm	≥	T1 + 0.30 [T2 + (1 – h2) Tm]
Or	3.5T1	≥	0.45T1 + 21(1 – h2) T1
Or	3.05/21	≥	(1 – h2)
Or	h2	≥	1 – 0.145 = 0.855

Hit rate of L2 cache should be at least 85.5%.

Problem 5: In a two-level cache system, cycle times of L1 and L2 caches and main memory are 1, 8 and 64 clock cycles, respectively. The miss rate of L2 cache is twice that of L1. What should the two miss rates be so that the average data access time of this cache system is 2 cycles? (3 points)

Answer: Suppose miss rates of L1 and L2 caches are m_1 and m_2, then $m_2 = 2m_1$. The average access time for the two-level cache system is given by,

	$1 + 8\,m_1 + 64\,m_1\,m_2$	=	2	
Or	$– 1 + 8m_1 + 128m_1^2$	=	0	
Therefore,	m_1	=	$[– 8 \pm (64 + 512)^{1/2}]/256$	
Or	m_1	=	16/256	= **0.0625**
And	m_2	=	$2m_1$	= **0.125**

Problem 6: A cache must accommodate extra bits besides the data bits. Each block contains tag bits and one valid bit. Show that the overhead for a one-level direct mapped cache is,

$$\text{Cache overhead} = \frac{\text{Total bits in cache} - \text{Data bits in cache}}{\text{Data bits in cache}} = \frac{1}{Bb}\left(1 + \log_2 \frac{W}{w}\right)$$

Where W = number of words in the main memory
w = number of data words in cache
B = block size in words
b = word size in bits

(2 points)

Answer:

Data bits in cache = w * b

Total bits in cache = data bits + #blocks in cache □ (1 valid bit + #tag bits)
= w * b + (w/B) □ (1 + $\log_2$(W/w))

Cache overhead = [w * b + (w/B) □ (1 + $\log_2$(W/w)) – w * b] / (w * b)

= (1 + $\log_2$(W/w)) / (B * b)