Sean Penney and Paul Atkinson
May 13, 2014
CS 472 HW 5

(9.2) The cache speeds up memory access by storing more frequently accessed data in the cache, rather than the slower main memory.

(9.3) Temporal locality refers to reusing specific data in a small amount of time, and spatial locality refers to reusing data that is grouped together.

(9.4) $S = t_m/(h * t_c + (1-h)t_m)$, where t_m is access time of main memory, t_c is the access time of cache memory, and h is the hit ratio.

This formula simplifies to:

$S = 1/(1 - h(1-k))$, where k is the ratio of the access time to cache memory divided by the access time to main memory.

(9.5) a. S = 526.3%

b. S = 689.6%

c. S = 416.67%

d. S = 1273.89%

(9.6) a. h = .0957

b. h = .5263

c. h = .8421

d. h = .9825

(9.8) a. $t_m/t_c = 4$

b. $t_m/t_c = 2$

c. $t_m/t_c = 4$

(9.11) a. Word is a unit of data

b. A line is made up of individual words.

c. A set is a group of lines

(9.12) a. Using direct mapped cache, the data is stored like a simple table, with columns for data and tags.

b. A fully associative cache allows you to store the data anywhere in the cache.

c. A set-associative cache is a hybrid between a fully associative and direct mapped cache. There is a general location in which the data is to be stored, and within that general location the data can be stored anywhere.

(9.17) Cache coherency refers to whether the data is consistent. Write operations should occur instantaneously, processors should see the same sequence of changes of values, etc.

(9.22) Data caches are easier to implement than instruction caches since the contents of data caches are not modified.

(9.23) The average access time of a system with a cache that's accessed in parallel with main store is $t_{ave} = ht_c + (1-h)t_m$

Let t_1 be the time to fetch a line from main store to reload the cache on a miss. We must add $(1-h)t_1$ to the average access time.

$t_{ave} = ht_c + (1-h)t_m + (1-h)t_1$

(9.26) $t_{ave} = h_1 t_{c1} + (1 - h_1) h_2 t_{c2} + (1 - h_1)(1 - h_2) t_m$

$= 2.22$ cycles

(9.28) The local miss rate is the number of misses in a specific cache divided by the total number of memory accesses to this cache.

The global miss rate is the misses in a specific cache divided by the total number of memory accesses by the CPU.

(9.35) Since there are 4 sections, 8/4 = 2 MB each. 62500 lines, and since 2 to the 16 is 65536, the set is 16 bits long. There are 5 offset bits, as 2 to the 5th = 32 bytes (the line size).

(9.41) CPU cache is a portion of memory made of high-speed static RAM.

Disk caching uses same principles as CPU caches, however, disk caching uses conventional main memory instead of SRAM.

(9.42) Write back means that a write operation to the main memory takes place only when a line in the cache is to be ejected, while write through can just do the write operation anytime.

(9.43) The amount of space that the system may be able to address is 2^{32} = 4GB.

The number of page table entries is 4 GB / 4 KB = 1 million.

There are 4 bytes (32 bits) in a page table entry.

1 million * 32 bits = 32 Mb or 4 MB

(9.45) The average cost of an instruction is 1.3 + . 5 + 5 = 6.8 cycles per instruction on average.

(9.46) To access the next element in the y[i] array, a read to main memory will be required.

x and s will be cached, presumably in the L1 cache.

Therefore, the access time is 50 cycles + 2 cycles + 2 cycles = 54 cycles for one iteration of the loop.

(9.57) a. 64 byte line size, set is 18 bits and 6 offset bits.

b. 64 byte line size, set is 24 bits and 6 offset bits.

c. 32 byte line size, set is 17 bits, and 5 offset bits