Drake Bridgewater & Ryan Phillips
May 11, 2014
CS 472 HW 5

(9.2)

Why do computer use cache memory?

Computers use cache memory to give the appearance of speed. When cache memory is used in conjunction with system memory and primary (secondary as the book puts it) storage the system can operate at much higher rates.

(9.3)

What is the meaning of temporal locality and spatial locality?

Addresses are said to exhibit **spatial locality** because they are clustered within the same region of memory. Which other addresses are said to exhibit **temporal locality** because they are accessed over and over again within a short time span.

(9.4)

From first principles, derive an expression for the speedup ratio of memory system with cache (assume the hit ration is h and the ratio of the main storage access time to cache access time to cache access time is k, where k¡1). Assume that the system is an ideal system and that you don't have to worry about the effect of clock cycle times.

(9.5)

For the following systems, calculate the speedup ratio S in the case t_c is the access time of the cache memory, t_m is the access time of the main store, and h is the hit ratio. Using $S = \frac{1}{1-H(1-k)}$ where $k = \frac{t_c}{t_m}$

a $t_m = 70ns,\ t_c = 7ns,\ H = 0.9$: **5.263**

b $t_m = 60ns,\ t_c = 3ns,\ H = 0.9$: **6.897**

c $t_m = 60ns,\ t_c = 3ns,\ H = 0.8$: **4.167**

d $t_m = 60ns,\ t_c = 3ns,\ H = 0.97$: **12.739**

(9.6)

For the following ideal systems, calculate the hit ratio h required to achieve the stated speedup ratio S. Using this equation$S = \frac{1}{1-H(1-k)}$ where $k = \frac{t_c}{t_m}$ we can solve it for H to get $H = \frac{1}{1-k} - \frac{1}{S(1-k)}$

a $t_m = 60ns,\ t_c = 3ns,\ S = 1.1$: **0.0957**

b $t_m = 60ns,\ t_c = 3ns,\ S = 2.0$: **0.5263**

c $t_m = 60ns,\ t_c = 3ns,\ S = 5.0$: **0.8421**

d $t_m = 60ns,\ t_c = 3ns,\ S = 15.0$: **0.9824**

(9.8)

For the following system that use a clocked microprocessor, calculate the maximum speedup ratio you could expect to see as h approaches 100%.

a $t_{cyc} = 20ns,\ t_m = 75ns,\ t_c = 15$

a $t_{cyc} = 20ns,\ t_m = 75ns,\ t_c = 25$

a $t_{cyc} = 10ns,\ t_m = 75ns,\ t_c = 15$

(9.11)

In a direct mapped cache memory system, what is the meaning of the following terms: word, line and set?

In a direct-mapped cache, the lines are arranged into units called sets, where the size of a set is the same size as the cache. The lines are then divided into words. These terms are all terms to narrow down where. Essentially like a persons address city like set, street like line, and word like house number.

(9.12)

How is data in main store mapped on to each of the following: a direct-mapped cache, a fully associative cache, and a set-associative cache?

(9.17)

What is cache coherency?

> "Cache coherence is the consistency of shared resource data that ends up stored in multiple local caches. When clients in a system maintain caches of a common memory resource, problems may arise with inconsistent data. This is particularly true of CPUs in a multiprocessing system." Wikipedia says it best.

(9.22)

Why is it harder to design a data cache than an instruction cache?

(9.23)

When a CPU writes to the cache, both the item in the cache and the corresponding item in the memory must be updated. If data is not in the cache, it must be fetched from memory and loaded in the cache. If t_1 is the time taken to reloaded the cache on a miss, show that the effective average time of the memory system is given by $t_{avg} = ht_c + (1-h)t_m + (1-h)t_1$

(9.26)

A system has a level 1 cache and a level 2 cache. the hit rate of the level 1 cache is 90% and the hit rate of the level 2 cache is 80%. An access to level 1 cache requires one cycle, an access to level 2 cache requires four cycles, and an access to main memory requires 50 cycles. What is the average access time?

If we take a sample of 100 accesses to memory/cache. 90% of these call will be caught by L1 cache and will require 1 cycle to complete, leaving 10 accesses which 80% will be caught by L2 cache requiring 4 cycles, lastly if the CPU misses on L1 and L2 then it will have to fetch from main memory requiring 50 cycles.

$$100_{total\_access} = (100 * 90\%_{hits} * 1_{cycles}) + (10 * 80\%_{hits} * 4_{cycles}) + (2 * 100\%_{hits} * 50_{cycles}) \quad (1)$$

$$(90 + 22 + 100)/100 = 2.3\ Avg.\ Cycles \quad (2)$$

(9.28)

In the context of multilevel caches, what is the difference between a local miss rate and global miss rate?

(9.35)

A 64-bit processor has a 8-MB, four-way set-associative cache with 32-byte lines. How is the address arranged in terms of set, line and offset bits?

(9.41)

What are the fundamental differences between cache memory (as found in a CPU) and cache memory found in a hard disk drive?

(9.42)

What are the differences between write-back and write-through caches, and what are the implications for system performance?

(9.43)

A computer with a 32-bit address architecture has a memory management system with single-level 4KB page tables. How much memory space must be devoted to the page tables?

(9.45)

Look at the table in the book to complete. A computer runs with the characteristics in the following table and determine the average number of cycles per instruction?

(9.46)

Consider the following code that accesses three values in memory scalar integers x and s , and an integer vector$y[i]$. what is the memory latency in clock cycles for a trip round the loop (after the first iteration)? Assume that the array is not cached and each new access to the array results in a miss. the system has both L1 and L2 caches. the access time of the L1 cache is 2 cycles, L2 cache is 6 cycles, and the main memory has access time of 50 cycles. in this case all memory and cache memory accesses tale place in parallel

```
for (i=0; i<100; i++)
{
x=y[i];
s=s+x;
}
```

(9.57)

A compter with a 24-bit address bus has a main memory of size 16MB and a cache size of 64KB. The word length is two bytes. What is the address format for eh direct-mapped cache with a line size of 32 words? What is the address format for a fully associative cache with a line size of 32 words? What is the address format for a f four-way set associative cache with line size of 16 words?