Chapter 6

Memory System Hierarchy: Role of Memory System (cont.)

Cache Performance:

A better measure of memory-hierarchy performance is the average memory access time (AMAT):

$$\text{AMAT (clock cycle)} = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty}$$

Where:

Hit time: is the time to hit the data requested by CPU in the cache
(*usually takes 1 clock cycle*)

Miss penalty: In case of a cache miss, the time needed to retrieve the first word of the missed MM block from the MM memory to the cache and CPU, it represents the miss penalty (sometimes is called latency) .

Example 1: A typical computer system has a MM of 1Mword and a direct mapped cache memory of 32Kword. The cache block size is 64words and the cache miss penalty is 100 clock cycle.

1) What is the size of the tag?
2) If the cache hit time is 1 clock cycle, what is the hit rate would be required to achieve an AMAT equal to 4.25 clock cycle?

Solution:

1) What is the size of the tag?

$$\text{MM size} = 1Mword = 2^{20}word \equiv 2^{n}$$

$$\therefore n = 20\ bit\ : MM\ address\ lines$$

$$MM\ block\ size = Cache\ line = 64word = 2^{6} = 2^{w}$$

$$\therefore word\ offset\ (w) = 6\ bit$$

$$cache\ size = 32Kword = 2^{15}$$

$$cache\ lines: C = \frac{Cache\ size}{block\ size} = \frac{32Kword}{64\ word} = 512\ lines \equiv 2^{c}$$

$$No.of\ bits\ needed\ to\ address\ cahe\ lines: c = 9bit$$

$$cache\ tag: t = n - c - w = 20 - 9 - 6 = 5bit$$

2) If the cache hit time is 1 clock cycle, what is the hit rate would be required to achieve an AMAT equal to 4.25 clock cycle?

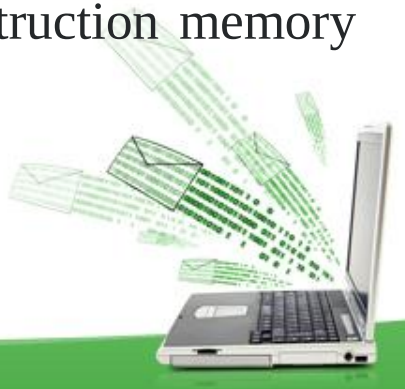
$$AMAT = hit\ time + miss\ rate\ * miss\ penalety$$
$$4.25\ clock\ cycle = 1 + \text{miss rate} * 100$$
$$miss\ rate = \frac{4.25 - 1}{100} = 0.0325$$

$$hit\ rate = 1 - miss\ rate = 1 - 0.0325 = 0.9675$$

Split and Unified Caches

A **split cache** is a cache that consists of two *physically separate* parts, where one part:

- Instruction cache (I- cache), is dedicated for holding instructions and,
- Data cache (D- cache), is dedicated for holding data (i.e., instruction memory operands).



Both of the I- cache and D- cache are ***logically*** **considered to be a single cache**, described as a split cache, because both are hardware-managed caches for the same physical address space at the same level of the memory hierarchy.

- Instruction fetch requests are handled only by the instruction cache and
- memory operand read and write requests are handled only by the data cache.

- A cache that is not split is called a "unified cache".

E X A M P L E 2:

Which has the lower miss rate: a 16-KB instruction cache with a 16-KB data cache or a 32-KB unified cache?

Use the miss rates in Figure 1 to help calculate the correct answer, assuming 47% of the instructions are data transfer instructions. Assume a hit takes 1 clock cycle and the miss penalty is 100 clock cycles. A load or store hit takes 1 extra clock cycle on a unified cache if there is only one cache port to satisfy two simultaneous requests. What is the average memory access time in each case? Assume write-through caches with a write buffer and ignore stalls due to the write buffer.

Size	Instruction cache	Data cache	Unified cache
8 KB	8.16	44.0	63.0
16 KB	3.82	40.9	51.0
32 KB	1.36	38.4	43.3
64 KB	0.61	36.9	39.4
128 KB	0.30	35.3	36.2
256 KB	0.02	32.6	32.9

Figure 1: Miss per 1000 instructions for instruction, data, and unified caches of different sizes. The percentage of instruction references is about 78%. The data are for two way associative caches with 64-byte blocks for the same computer

ANSWER

First let's convert misses per 1000 instructions into miss rates. Solving the general formula is from above, miss rate is

$$miss\ rate = \frac{misses/1000}{memory\ accesses/instruction}$$

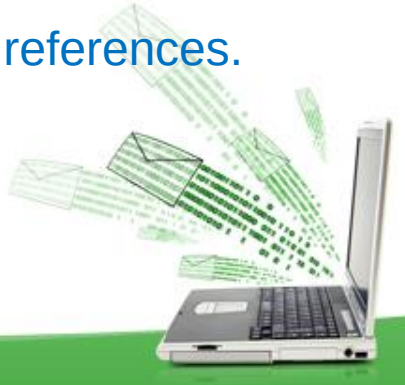
Since every instruction access has exactly 1 memory access to fetch the instruction, the instruction miss rate is:

$$miss\ rate_{16KB\ I-cache} = \frac{3.82/1000}{1} = 0.004$$

Since 47% of the instructions are data transfers, the data miss rate is:

$$miss\ rate_{16KB\ I-cache} = \frac{40.9/1000}{0.47} = 0.087$$

As stated above, about 78% of the memory accesses are instruction references. Thus, the overall miss rate for the split caches is



Miss rate – split caches = $(78\% \times 0.004) + (22\% \times 0.087) = 0.022$

The unified miss rate needs to account for instruction and data accesses:

$$miss\ rate_{32KB\ unified\ cache} = \frac{43.3/1000}{1 + 0.47} = 0.029$$

Thus, a 32-KB unified cache has a higher effective miss rate than two16- KB caches.

The average memory access time **AMAT** formula can be divided into instruction and data accesses:

$$\text{AMAT} = \%\ \text{instructions} \times (\text{Hit time} + \text{Instruction miss rate} \times \text{Miss penalty}) + \%\ \text{data} \times (\text{Hit time} + \text{Data miss rate} \times \text{Miss penalty})$$

Therefore, the time for each organization is

$$\text{Average memory access time}_{\text{split}}$$
$$= 78\% \times (1 + 0.004 \times 100) + 22\% \times (1 + 0.087 \times 100)$$
$$= (78\% \times 1.38) + (22\% \times 9.70) = 1.078 + 2.134 = 3.21$$

$$\text{Average memory access time}_{\text{unified}}$$
$$= 78\% \times (1 + 0.029 \times 100) + 22\% \times (1 + 1 + 0.029 \times 100)$$
$$= (78\% \times 3.95) + (22\% \times 4.95) = 3.080 + 1.089 = 4.17$$

Hence, the split caches in this example—which offer two memory ports per clock cycle, have a better average memory access time (AMAT) than the unified cache.

1- Multi-Level Caches

The performance gap between processors and memory leads the architect to this question:

Should I make the cache faster to keep pace with the speed of CPUs, or make the cache larger to overcome the widening gap between the CPU and main memory?

The answer is: both. Adding another level of cache between the original cache and memory simplifies the decision.

- The first-level cache can be small enough to match the clock cycle time of the fast CPU.
- The second-level cache can be large enough to capture many accesses that would go to main memory, thereby lessening the effective miss penalty.

Two level cache

Let's start with the definition of *average memory access time AMAT* for a two-level cache. Using the subscripts L1 and L2 to refer, respectively, to a first-level and a second-level cache, the original formula is

$$\text{Average memory access time} = \text{Hit time}_{L1} + \text{Miss rate}_{L1} \times \text{Miss penalty}_{L1}$$

and

$$\text{Miss penalty}_{L1} = \text{Hit time}_{L2} + \text{Miss rate}_{L2} \times \text{Miss penalty}_{L2}$$

so

$$\text{Average memory access time} = \text{Hit time}_{L1} + \text{Miss rate}_{L1} \times (\text{Hit time}_{L2} + \text{Miss rate}_{L2} \times \text{Miss penalty}_{L2})$$

To avoid ambiguity, these terms are adopted here for a two-level cache system:

- *Local miss rate*—This rate is simply the *number of misses in a cache divided by the total number of memory accesses to this cache*. As you would expect, for the first-level cache it is equal to Miss rate$_{L1}$ and for the second-level cache it is Miss rate$_{L2}$.

- *Global miss rate*—This rate is *the number of misses in the cache divided by the total number of memory accesses generated by the CPU*. Using the terms above, the global miss rate for the first-level cache is still just Miss rate$_{L1}$ but for the second-level cache it is Miss rate$_{L1}$ × Miss rate$_{L2}$.

Note: The *local miss* rate is large for second level caches because the first-level cache skims the cream of the memory accesses.

EXAMPLE 3

Suppose that in 1000 memory references there are 40 misses in the first level cache and 20 misses in the second-level cache. What are the various miss rates? Assume the miss penalty from L2 cache to Memory is 100 clock cycles, the hit time of L2 cache is 10 clock cycles, the Hit time of L1 is 1 clock cycles, and there are 1.5 memory references per instruction. What is the average memory access time?

Solution:

L1- cache:

Local miss rate = Global miss rate = L1- cache misses/ Number of accesses to L1 cache =40/1000 or 4%.

L2- Cache:

local miss rate L2-cache = L2- cache misses/ number of accesses to L2-cache
= 20/40 or 50%.

The global miss rate of L2- cache = L2- cache misses/ total number of accesses generated by the CPU = 20/1000 or 2%.

$$\text{Average memory access time} = \text{Hit time}_{L1} + \text{Miss rate}_{L1} \times (\text{Hit time}_{L2} + \text{Miss rate}_{L2} \times \text{Miss penalty}_{L2})$$

AMAT = 1 + 0.04 × (10 + 0.5 × 100) = 1 + 0.04 × 60 = 3.4 clock cycles

Q: If a direct mapped cache has a hit rate of 95%, a hit time of 4 ns, and a miss penalty of 100 ns, what is the AMAT?

Solution:

- If a direct mapped cache has a hit rate of 95%, a hit time of 4 ns, and a miss penalty of 100 ns, what is the AMAT?

  AMAT = Hit time + Miss rate x Miss penalty = 4 + 0.05 x 100 = 9 ns

Q: If an L2 cache is added with a hit time of 20 ns and a hit rate of 50%, what is the new AMAT?

Solution:

If an L2 cache is added with a hit time of 20 ns and a hit rate of 50%, what is the new AMAT?

$$AMAT = Hit\ Time_{L1} + Miss\ Rate_{L1} \times (Hit\ Time_{L2} + Miss\ Rate_{L2} \times Miss\ Penalty_{L2})$$

$$= 4 + 0.05 \times (20 + 0.5 \times 100) = 7.5\ ns$$