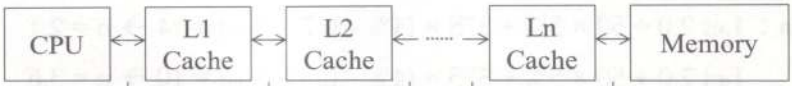
平均記憶體存取時間

➢ 在多層快取結構下，CPU存取記憶體一次所需的平均時間稱為**平均記憶體存取時間(Average Memory Access Time, AMAT)**。假設只考慮單一層快取結構則平均記憶體存取時間可計算如下：

AMAT = Time for a hit + Miss rate × Miss penalty

➢ 若考慮多層快取結構，如下圖所示，則平均記憶體存取時間可計算如下：



Hit Time	T_1	T_2		T_n
Miss rate	M_1	M_2	.....	M_n
Miss penalty	P_1	P_2	.....	P_n

$$AMAT = T_1 + M_1 \times P_1 + M_2 \times P_2 + \ldots + M_n \times P_n = T_1 + \sum_{i=1}^{n} M_i P_i$$

練習 [94東華資工]

Suppose that in 1000 memory references there are 60 misses in the first-level cache, 30 misses in the second-level cache, and 5 misses in the third-level cache. Assume the miss penalty from the L3 cache to memory is 100 clock cycles, the hit time of the L3 cache is 10 clocks, the hit time of the L2 cache is 5 clocks, the hit time of L1 is 1 clock cycle, and there are 1.5 memory references per instruction.

(a) What's the global miss rate for each level of caches?

(b) What's the local miss rate for each level of caches?

(c) What is the average memory access time?

(d) What is the average stall cycle per instruction?

Answer

(a) L1 = 60/1000 = 0.06, L2 = 30/1000 = 0.03, L3 = 5/1000 = 0.005

(b) L1 = 60/1000 = 0.06, L2 = 30/60 = 0.5, L3 = 5/30 = 0.167

(c) AMAT = 1 + 0.06 × 5 + 0.03 × 10 + 0.005 × 100 = 2.1 clock cycles

(d) (2.1 – 1) × 1.5 = 1.65 clock cycles

CPU 執行.

練習

The Average Memory Access Time equation (AMAT) has three components: hit time, miss rate, and miss penalty. For each of the following cache optimizations, indicate which component of the AMAT equation is improved.

(1) Using a second-level cache

(2) Using a direct-mapped cache

associativity ↑ => miss rate ↓ hit time ↑

(3) Using a 4-way set-associative cache

(4) Using a virtually-addressed cache p.79

(5) Performing hardware pre-fetching using stream buffers △

(6) Using a non-blocking cache

(7) Using larger blocks

Answer

(1) miss penalty

(2) hit time

(3) miss rate

(4) hit time

(5) miss rate

(6) miss penalty

(7) miss rate

(4) 不用像 physically-addressed cache 必須先把 virtual address 經過 TLB 轉換成 physical address

練習

Assume that main memory accesses take 70 ns 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	1 KB	11.4%	0.62 ns
P2	2KB	8.0%	0.66 ns

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

(2) What is the AMAT for each of P1 and P2?

(3) Assuming a base CPI of 1.0, what is the total CPI for each of P1 and P2? Which processor is faster?

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 the following problems. The L2 miss rate indicated is its local miss rate.

L2 size	L2 miss rate	L2 hit time
512 KB	98%	3.22 ns

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

(5) Assuming a base CPI of 1.0, what is the total CPI for P1 with the addition of an L2 cache?

(6) Which processor is faster, now that P1 has an L2 cache? If P1 is faster, what miss rate would P2 need in its L1 cache to match P1's

performance? If P2 is faster, what miss rate would P1 need in its L1 cache to match P2's performance?

Answer

(1)		(2)		(3)	
P1	1.61 GHz	8.60 ns	13.87 cycles	18.5	P2
P2	1.52 GHz	6.26 ns	9.48 cycles	12.54	

(4)			(5)
8.81 ns	14.21 cycles	Worse	18.96

(6)
P1 with L2 cache: CPI = 18.96. P2: CPI = 12.54. P2 is still faster than P1 even with an L2 cache The miss rate for P1 in its L1 cache should be 7.83% to match P2's performance

註(3)： $CPI_{P1} = 1 + (1.36 \times 0.114 \times 70/0.62) = 18.5$

$CPI_{P2} = 1 + (1.36 \times 0.08 \times 70/0.66) = 12.54$

註(4)： L2 global miss rate = $0.114 \times 0.98 = 0.11172$

$0.62 + 0.114 \times 3.22 + 0.11172 \times 70 = 8.81$

註(5)： $1 + 1.36 \times (0.114 \times 3.22/0.62 + 0.11172 \times 70/0.62) = 18.96$

註(6)： Suppose the L1 cache miss rate is M

CPI for P1 with second level cache = $1 + 1.36 \times (M \times 3.22 / 0.62 + M \times 0.98 \times 70 / 0.62) = 1 + 157.54\ M$

P1 performance match P2 performance implies both instruction times should be the same → $(1 + 157.54\ M) \times 0.62 = 12.54 \times 0.66$

→ M = 7.83%

練習

To capture the fact that the time to access data for both hits and misses affects performance, designers often use average memory access time (AMAT) as a way to examine alternative cache designs. Average memory access time is the average time to access memory considering both hits and misses and the frequency of different accesses; it is equal to the following:

$$\text{AMAT} = \text{Time for a hit} + \text{Miss rate} \times \text{Miss penalty}$$

AMAT is useful as a figure of merit for different cache systems.

(1) Find the AMAT for a processor with a 2 ns clock, a miss penalty of 20 clock cycles, a miss rate of 0.05 misses per reference, and a cache access time (including hit detection) of 1 clock cycle. Assume that the read and write miss penalties are the same and ignore other write stalls.

(2) Suppose we can improve the miss rate to 0.03 misses per reference by doubling the cache size. This causes the cache access time to increase to 1.2 clock cycles. Using the AMAT as a metric, determine if this is a good trade-off.

(3) If the cache access time determines the processor's clock cycle time, which is often the case, AMAT may not correctly indicate whether one cache organization is better than another. If the processor's clock cycle time must be changed to match that of a cache, is this a good trade-off? Assume the processors are identical except for the clock rate and the number of cache miss cycles; assume 1.5 references per instruction and a CPI without cache misses of 2. The miss penalty is 20 cycles for both processors.

Answer

(1) AMAT = 2 ns + 0.05 × (20 × 2 ns) = 4 ns ✓

(2) AMAT = (1.2 × 2 ns) + (20 × 2 ns × 0.03) = 2.4 ns + 1.2 ns = 3.6 ns ✓

Yes, it's a good choice.

(3) Execution time$_{old}$ = 2 × IC × (2 + 1.5 × 20 × 0.05) = 7 IC

Execution time$_{new}$ = 2.4 × IC × (2 + 1.5 × 20 × 0.03) = 6.96 IC

So, it's a good choice

練習

For a data cache with a 92% hit rate and a 2-cycle hit latency, calculate the average memory access latency. Assume that latency to memory and the cache miss penalty together is 124 cycles. Note: The cache must be accessed after memory returns the data. AMAT= 2 + 0.08 ×124

Answer: AMAT = 2 + 0.08 × 124 = 11.92 cycles ✓

練習

Which of the following is generally true about a design with multiple levels of caches?

1. First-level caches are more concerned about hit time, and second-level caches are more concerned about miss rate. ✓
2. First-level caches are more concerned about miss rate, and second-level caches are more concerned about hit time.

Answer: 1

L1 cache : 跟小 hit time

L2 cache : 降低 miss rate.