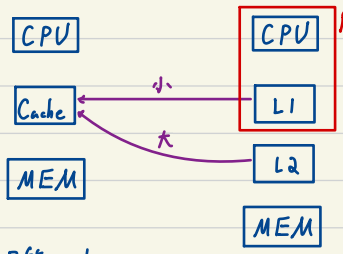
Multi-Level Cache：减少 miss penalty

1-level cache　　2-level cache



- L1 採用 split cache ⇒ 增加頻寬，減少 hit time
  L2 採用 combined cache ⇒ 減少 miss rate
- L1 採用 write through / write back 皆可
- L2 採用 write back

Cache 只能 medium size

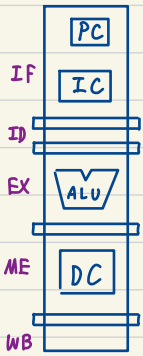
∵ 太大 ⇒ hit time 長　　L1 越小越好 ⇒ 減少 hit time, clock cycle time

太小 ⇒ miss rate 高　　L2 越大越好 ⇒ ∵ L1 miss 可到 L2 找

減少 miss penalty

Multi-Level Cache 效能分析

Split Cache：



$$\text{IC Extra CPI} = \left(\#\text{ of I-C access per instruction}\right) \times \sum_{i=1}^{k} \text{Miss Rate}_i \times \text{Miss Penalty}_i$$

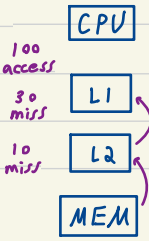
$$\text{DC Extra CPI} = \left(\#\text{ of D-C access per instruction}\right) \times \sum_{i=1}^{h} \text{Miss Rate}_i \times \text{Miss Penalty}_i$$

Combined Cache：

	MR	MP
$1 + lw/sw\%$	MR_1	MP_1
	MR_2	MP_2
	⋮	⋮
	MR_n	MP_n

# of access per instruction

Example：lw/sw = 20%



Total stall cycles：$30 \times 20 + 10 \times 50$

Stall cycles per access：$\frac{30}{100} \times 20 + \frac{10}{100} \times 50 = \sum_{i=1}^{n} MR_i \times MP_i$

Stall cycles per instruction：$1.2 \cdot \left[\frac{30}{100} \times 20 + \frac{10}{100} \times 50\right] = (1 + lw/sw\%)\left(\sum_{i=1}^{n} MR_i \times MP_i\right)$

$= \left[(1.2 \times \frac{30}{100}) \times 20 + (1.2 \times \frac{10}{100}) \times 50\right]$

$= \sum_{i=1}^{n} \text{miss rate per instruction}_i \times \text{miss penalty}_i$

<Ex P.48> base CPI: 1.0

clock rate: 4GHz

memory access time: 100 ns = 400 clock cycles

L1 miss rate per instr: 2%

if L2 exists: L1 miss penalty: 5 ns = 20 clock cycles

Original effective CPI: $1.0 + 0.02 \times 400 = 9$

增加 L2 : $1.0 + 0.02 \times 20 + 0.005 \times 400$

$= 3.4$ 快.

Original:

CPU	L1	MEM
MRP_i	0.02	X
MP_i	400	X

1次 memory access time 作 penalty time

增加 L2:

CPU	L1	L2	MEM
MRP_i	0.02	0.005	X
MP_i	20	400	X

Note:
1. 沒給定義 CPI 設為 1
2. 給 miss rate, 沒給 lw/sw ratio, 設為 0
3. penalty 時間給定絕對時間, 沒給 clock rate & clock cycle time
   以 L1 cache hit time 作 clock cycle time

Global & Local Miss Rate

1000 access	CPU	GMR	LMR
50 miss	L1	$\frac{50}{1000}$	$\frac{50}{1000}$
20 miss	L2	$\frac{20}{1000}$	$\frac{20}{50}$
5 miss	L3	$\frac{5}{1000}$	$\frac{5}{20}$
	MEM		

Global Miss Rate: 以 CPU 為觀點

Local Miss Rate: 以上一層儲存体為觀點

$MRP_i = (1 + lw/sw\%) \times GMR_i$

$= (1 + lw/sw\%) \times \prod_{k=1}^{i} LMR_i$

Average memory access time (AMAT)

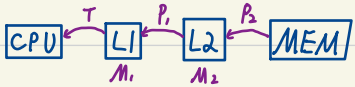
CPU	cache (T, P)	MEM
hit time	T	
miss rate	M	
miss penalty	P	

$AMAT = T \times (1 - M) + (T + P) \times M$

$= T - T \times M + T \times M + M \times P$

$= T + M \times P$

2-level



$AMAT = T \times (1 - M_1) + (T + P_1) \times (M_1 - M_2) + (T + P_1 + P_2) \times M_2$

$= T + M_1 \times P_1 + M_2 \times P_2$

n-level cache:

General, $T_{L_1} + \sum_{i=1}^{n} MR_i \times MP_i$

Note: 增加 associativity 不一定會降低 AMAT

∵ AMAT = hit time + MR × MP

降低 MR，但 hit time 上升

<Ex P.54>:

1000 access　60 miss　30 miss　5 miss

CPU	L1	L2	L3	MEM
GMR	6%	3%	0.5%	
LMR	6%	50%	$\frac{5}{30}$	
hit time	1	5	10	
MP	5	10	100	

AMAT = 1 + 0.06×5 + 0.03×10 + 0.005×100

= 2.1 clock cycles

Extra CPI = (1 + lw/sw ratio) × $\sum MR_i \times MP_i$

= 1.5 × 1.1 = 1.65 #

Note: AMAT per instruction

= (1 + lw/sw%) × AMAT

= (1 + lw/sw%) × (hit time + MR × MP)

= (1 + lw/sw%) × hit time + MRP × MP