CURS 12 21.05.2019

3.8. Replacement policies

a) Random

b) FIFO

c) LRU - Least Recently Used

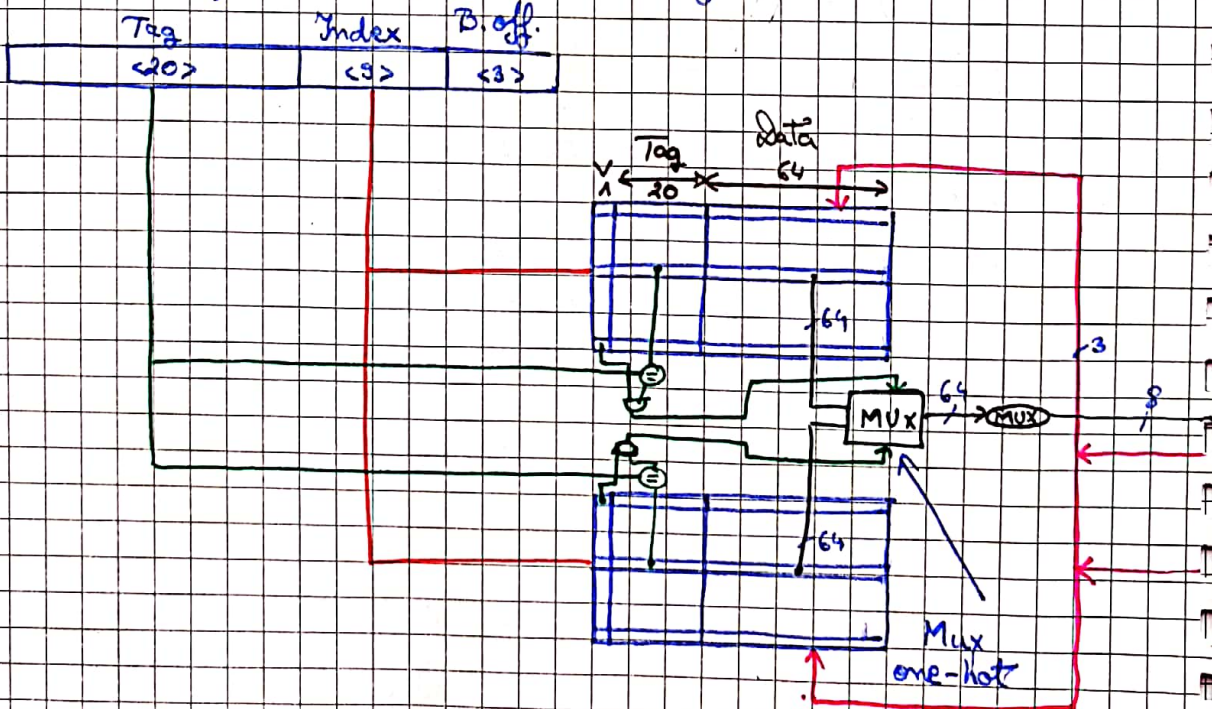
No. of "age" bits = $\log_2$ (set associativity)

Example 1: Pp. că avem un set asociativ pe 4 căi cu adrese în zecimal și pp. că toate adresele se mapează la același index

		Data 0	Data 1	Data 2	Data 3	Age 0*	Age 1*	Age 2*	Age 3*	
		0	0	0	0	0	0	0	0	
Miss →	14	14	0	0	0	0	0	0	0	
(cache M →	7	14	7	0	0	1	0	0	0	
gol) M →	6	14	7	6	0	2	1	0	0	
M →	12	14 (c.m. VECHI)	7	6	12	3	2	1	0	
M →	10	10	7	6	12	0	3	2	1	
(Hit) H →	6	10	7	6	12	1	3	0	2	← Cei cu vârstă mai mare rămân la fel, cei cu mai mică se incrementează
H →	7	10	7	6	12	2	0	1	3	
M →	14	10	7	6	14	3	1	2	0	
H →	7	10	7	6	14	3	0	2	1	
M →	3	3	7	6	14	0	1	3	2	
H →	14	3	7	6	14	1	2	3	0	

Aici de fapt e cuv. de la adresa 14, nu "14"

Example 2: Mașina Vax 11/780 + 2-way SA



2^3 words / block = 2^3 B / block = $2^3 \cdot 2^3$ bits / block = 64 bits / block

Cache Size (data) = s.a $\times 2^{index} \times$ block size = $2 \times 2^9 \times 2^3 B = 2^{13} B$ = 8 KiB

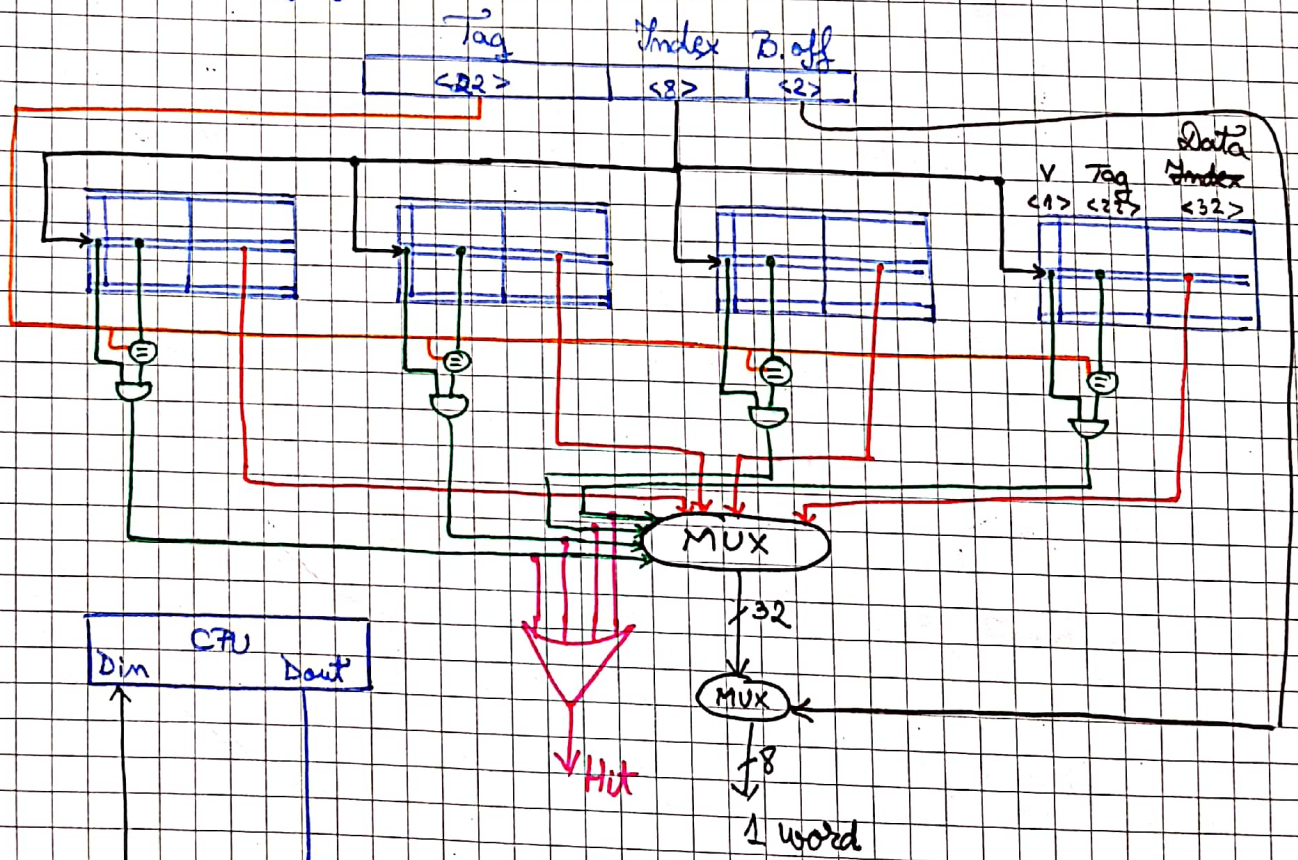
↑ set associativity

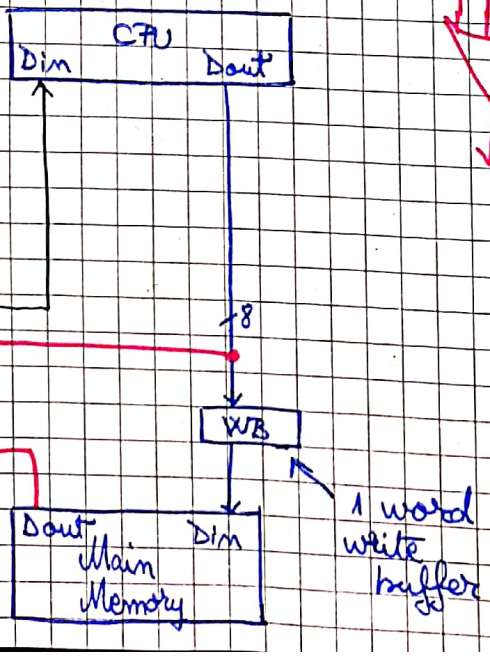
Write Policy → Write Through (Write Buffer) → No Allocate in case of a Write miss.

Example 3:

- 4 way - SA ; 1 word = 1 B ; 1 block = 4 words = 2^2 words
- Cache size = 2^{10} blocks = 1 KiBlocks
- MM size = 4 GiB

$$\frac{\text{No. of blocks}}{\text{s.a}} = \frac{2^{10}}{2} = 2^8 \text{ blocks / bank}$$





Example 4: Se dau 2 memorii cache diferite ale căror 2 sisteme de calcul de ~~ost~~ altfel identice.

Instruction miss rate ← ~~Index~~ ; Data miss rate

cache A: 2 way s.a ; $I_{mr} = 2\%$; $D_{mr} = 1,7\%$

cache B: 4 way s.a. ; $I_{mr} = 1,6\%$; $D_{mr} = 1,4\%$

A și B sunt mașini de calcul load/store, 20% din instrucțiuni fiind load/store.

Absolut miss penalty = 200 ns

$CCT_A = 20$ ns

La B, pt. că s.a e mai mare, mux e mai complicat, timpul e mai mare, deci CCT_B se mărește cu 10% (s-a degradat)

$CPI_{ideal} = 3$ clock cycle-uri

Care mașină e mai bună?

$$CPU_{time} = IC \times \left(CPI_{ideal} + \underbrace{\text{Memory accesses per instr.} \times \text{Miss Rate}}_{\text{Misses per instr.}} \times \text{Miss Penalty}\right) \times CCT$$

Misses per instr. A = $1 \times I_{mr} + 0,2 \times D_{mr} = 0,02 + 0,2 \times 0,017 = 0,0234$ (0,2 ↑ 20%)

Misses per instr. B = $0,016 + 0,2 \times 0,014 = 0,0188$

Miss Penalty A = $\left[\frac{200 \text{ ns}}{20 \text{ ns}}\right] = 10$ clock cycles ($20 \text{ ns} \uparrow CCT_A$)

Miss Penalty B = $\left[\frac{200 \text{ ns}}{22 \text{ ns}}\right] = 10$ clock cycles (trb. să fie număr întreg) ($22 \text{ ns} \uparrow CCT_B = CCT_A + 20\% \cdot CCT_A$)

$CPU_{time\ A} = IC \times (3 + 0,0234 \times 10) \times 20 \text{ ns} =$

$= IC \times 3,234 \times 20 = IC \times \underbrace{64,64 \text{ ns}}_{\text{timpul mediu pt. o instrucțiune, luând în considerare imperfecțiunea cache-ului}}$

$CPU_{time\ B} = IC \times (3 + 0,0188 \times 10) \times 22 \text{ ns} = IC \times 70,136 \text{ ns}$

Example: Admitem un cache de 64 kBytes:

CCT = 20 ns

Media nr. de accese la instrucțiune: Memory access per instr. = 1,3

$CPI_{ideal} = 1,5$

Pt. acest cache, dacă se adoptă set asociativ, det. o degradare a frecv. clock-ului cu 8,5 %. În acest context, se analizează cache-ul în felul urm.: mapare directă vs. 2 way S.A.

Statistic: Miss rate $_{DM} = 3,9\%$

Miss rate $_{2\text{ way S.A}} = 3\%$

Care soluție e mai ok știind că în ambele cazuri, pt. Miss Penalty = 200 ns (expresie temporală).

În mod normal, Miss Penalty e exprimat în clockuri

$$AMAT = HCCT + Miss\ Rate \times \frac{Miss\ Penalty}{Time}$$

$$CPU_{time} = IC \times \left(CPI_{ideal} \times CCT + Mem.\ acc\ per\ instr. \times Miss\ Rate \times \frac{Miss\ Penalty}{Time}\right)$$

$$\frac{Miss\ Penalty}{Time} = Miss\ Penalty\ clocks \times CCT$$

$$AMAT_{DM} = 20\ ns + 0,039 \times 200\ ns = 20\ ns + 7,8\ ns = 27,8\ ns$$

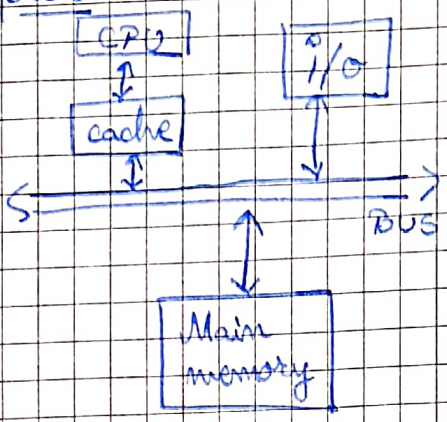
Better $AMAT_{SA2} = \underbrace{21,7}_{=20+8,5\%\cdot 20} + 0,03 \times 200\ ns = 27,7\ ns$

Better $CPU_{time\ DM} = IC \times (1,5 \times 20\ ns + 1,3 \times 0,039 \times 200) \cong IC \times 40,14\ ns$

$CPU_{time\ 2\ SA} = IC\ (1,5 \times 21,7\ ns + 1,3 \times 0,03 \times 200) \cong IC \times 40,35\ ns$

CPU_{time} ia în considerare mai mult ca AMAT, deci DM e mai bun.

Example 6:



Cât la sută din BUS Bandwidth este „mâncat" de imperfecțiunea cachelui?

Se dă un sist. de calcul cu:

BUS Bandwidth = 10^9 words/sec

Pp. că: Miss Rate = 10%

1 block = 4 words

CPU reference frequency = 10^8 words/sec (reads & writes)

BUS width = 2 words

Pp. că în (*) mom., 35% din blocurile cachelui au fost modificate (blockuri dirty). Cache are o politică write-allocate, frecvența scrierilor e de 30%.

Alternative de design, știind că Bandwidth e mâncat:

a) write back (WB)

b) write true (WT)

a) Nb. de accese la Bus datorită imperfecțiunii cache = $10^8 \times 0,1 \times \left(\%\text{Reads} \times \text{Read Miss Penalty} + \%\text{Writes} \times \text{Write Miss Penalty}\right)$

$= 10^8 \times 0,1 \times 2,7$

Read Miss Penalty $_{WB} = \left(\frac{\text{Block size}}{\text{Bus width}}\right) \times \text{Bus writes} \times 0,35 +$ ↳ blockurile dirty

↳ nr. de accese pe bus — Update

$+ \left(\frac{\text{Block size}}{\text{Bus width}}\right) \cdot \text{Bus reads}$ — Allocate

$= 2 \cdot 0,35 + 2 = 2,7 =$ Write miss penalty

% Bus Used $_{WB} = \frac{10^7 \cdot 2,7}{10^9} = \frac{2,7}{100} = 2,7\%$

b) Temă: WT

Nr. acc... = $10^8 \times 0,1 + (\%$ Reads $\times$ Read Miss Penalty + % Writes $\times$ $\times$ Write Miss Penalty) + $10^8 \times 0,9 \times \underbrace{0,3}_{\%\text{Writes}} \times \underbrace{\text{Write Hit Penalty}}_{=1}$

Read Miss Penalty = Write Miss Penalty =

$= \underbrace{2 \text{ BUS Reads}}_{\text{Allocate}} + 0,35 \cdot \underbrace{2 \text{ BUS Writes}}_{\text{Update for Main Memory}}$

CVRS 13

28.05.2019

b) Nr. acces = $\underbrace{10^8 \text{ words/second}}_{\text{Frequency of memory references}} \times \underbrace{0,1}_{\text{Miss Rate}} \times (0,3 \text{ Write Miss Penalty} +$

$+ 0,7 \text{ Read Miss Penalty}) + 10^8 \times 0,9 \times 0,3 \text{ Write Hit Penalty}$

Read Miss Penalty = 2 BUS Reads (nu mai updateaza)

Write Miss Penalty = 2 BUS Reads + 1 BUS Write =

= 3 BUS Accesses

Write Hit Penalty = 1 BUS Write = 1 BUS Access

Asadar: Nr. acces: $10^8 \cdot 0,1 \, (0,3 \cdot 3 + 0,7 \cdot 2) + 10^8 \cdot 0,9 \cdot 0,3 \cdot \underline{1} =$

$= 10^7 \cdot 2,3 + 2,7 \cdot 10^7 = 5 \cdot 10^7$

% Bus Used$_{WT}$ $= \frac{10^7 \cdot 5}{10^9} = \frac{5}{100} = 5\%$

3.9. Reducing miss rate with multi-level caches