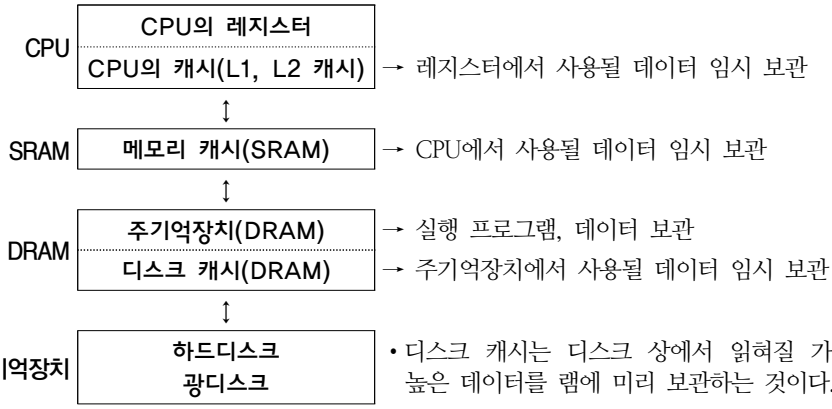
17. 캐시(cache)

캐시는 **CPU와 주기억장치의 속도 차이를 보완**하기 위해 사용되는 메모리이다.

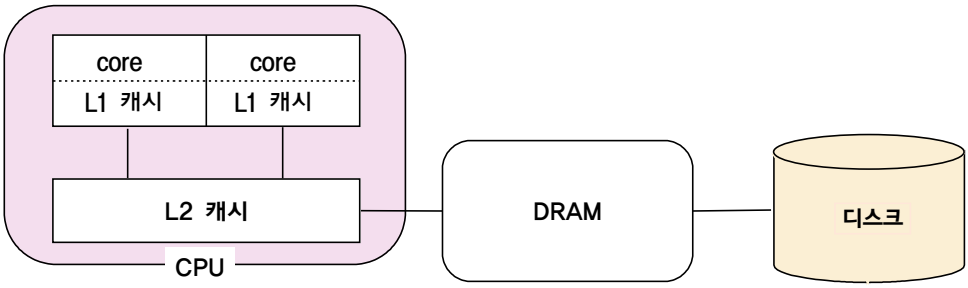
〈캐시 계층구조〉



• 현재, 캐시도 계층화되어 L1과 L2가 있고, 최근에는 L3 개념도 사용되고 있다.

다음은 듀얼코어(dual-core) 프로세서가 장착된 컴퓨터 구조를 보여준다.

〈Dual CPU core chip 컴퓨터 구조〉



• 듀얼코어(dual-core) 프로세서는 2개의 코어를 포함하고 있고
• 쿼드코어(quad-core) 프로세서는 4개의 코어를 포함하고 있고
• 옥타코어(octa-core) 프로세서는 8개의 코어를 포함하고 있고
• 데카코어(deca-core) 프로세서는 10개의 코어를 포함하고 있다.

// 펜티엄 2 컴퓨터에서 CPU 및 각 메모리의 데이터 읽기 속도

메모리 종류	데이터 읽기 속도
CPU(레지스터)	1 cycle
L1 캐시	1-4 cycle
L2 캐시	10-20 cycle
DRAM	50-300이상 cycle

- 메모리 속도에서 CPU는 레지스터를 의미한다. CPU는 레지스터로 구성되므로
- 저장장치 접근속도 : 레지스터 〉 L1 캐시 〉 L2 캐시 〉 L3 캐시 〉 DRAM 〉 보조기억장치

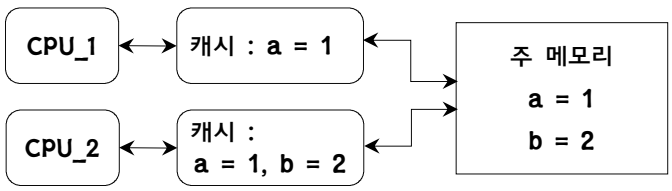
1. 캐시 정책

캐시 쓰기 정책은 캐시 일관성 유지하기 위한 것이다.

Write-through (즉시쓰기)	• 쓰기 동작이 발생할 때마다 **캐시와 주기억장치의 내용을 한꺼번에 갱신**한다. • 처리는 간단하지만, 동시 갱신으로 시간지연 발생(성능저하) • 성능향상을 위해 **쓰기 버퍼** 사용**(버퍼링된 Write-through)**
Write-back (나중쓰기)	• 쓰기 동작 발생 시 **캐시만 갱신**한다.(처리는 복잡, 성능은 우수) • 나중에 메모리를 갱신할 수 있도록 위치를 표시해 둔다. • **Write-back**에서도 **쓰기 버퍼**를 사용한다.(블록이 교체될 때 갱신된 것 쓰기 수행)

◈ 캐시 일관성(cache coherence)

- 캐시 일관성은 공유 메모리 시스템에서 각 CPU가 가진 로컬 캐시 사이의 일관성을 의미한다.
- 캐시 일관성 유지는 데이터 불일치 현상을 없애는 것을 의미한다.



좌측 그림에서 변수 a, b의 값은 공유하는 값이다. 일관성을 유지해야 한다.

2. 캐시 적중률

$$\text{적중률(hit ratio)} = \frac{\text{적중횟수}}{\text{적중횟수} + \text{실패횟수}} \times 100$$

- CPU가 데이터를 처리할 때 먼저 캐시에 있는지를 찾아본다.
- 만약 원하는 데이터가 있으면(적중), 메인 메모리(주기억장치)로 갈 필요가 없게 된다.
- RAM의 데이터를 캐시로 전송하는 것을 매핑(mapping) 프로세스라 한다.
- 캐시 적중률은 프로그램 실행이 가지는 **지역성** 특성에 의존한다.(시간지역성, 공간지역성)

다음은 **캐시 적중률**과 관련된 문제이다.

[문제 1] 순돌이는 공무원 시험을 준비하기 위해 매일 아침에 집에서 도서관으로 간다. 집에서 가까운 '도서관 1'에 가면 자리가 없을 수도 있다. 그러면, 멀리 있는 '도서관 2'로 간다. 다음과 같은 조건일 때 순돌이가 도서관까지 가는데 걸리는 평균시간 얼마인가?

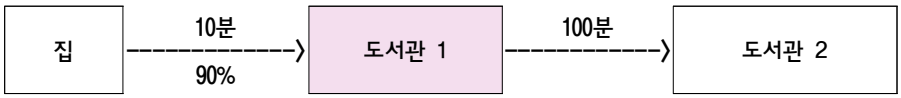
• 도서관 1까지 가는데 걸리는 시간은 10분이다.
• 도서관 1에서 자리를 확보할 가능성은 90%이다.
• 도서관 1에 자리가 없으면 도서관 2로 간다.
• 도서관 1에서 도서관 2까지 걸리는 시간은 100분이다.
• 도서관 2에 가면 반드시 자리가 있다.

---

① 10분 ② 15분
③ 20분 ④ 25분

♤ **풀이 – 평균시간**

• 순돌이가 도서관까지 가는데 소요되는 시간을 그림으로 그리면



• 평균시간 = (도서관 1까지 가는데 걸리는 시간 × 자리를 확보할 확률)
+ (도서관 2까지 가는데 걸리는 시간 × 자리를 확보할 확률)

• 평균시간 = 10 × 0.9 + (10 + 100) × 0.1 = 9 + 11 = **20(분)**

정답 : ③

// 주어진 문제의 응용

• 주어진 문제는 캐시메모리 등에서 원하는 데이터에 실제 접근하는 시간을 구하는 모든 유형의 문제에 응용할 수 있다.
• 캐시메모리 응용에서 도서관 1은 캐시메모리, 도서관 2는 주기억장치가 될 수 있다.

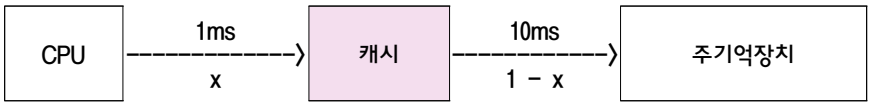
[문제 2] 주기억장치와 캐시 기억장치만으로 구성된 시스템에서 다음과 같이 기억장치 접근시간이 주어질 때 캐시 적중률(hit ratio)은? [2012년 계리직]

- 평균 기억장치 접근시간 : Ta = 1.9ms
- 주기억장치 접근시간 : Tm = 10ms
- 캐시 기억장치 접근시간 : Tc = 1ms

① 80% ② 85%
③ 90% ④ 95%

♤ **캐시 적중률**

// 캐시 적중률을 x라 하면



- 평균 기억장치 접근시간 = (캐시 접근시간 × x)
  + (캐시 접근시간 + 주기억장치 접근시간) × (1 − x)

- 평균 기억장치 접근시간 = (1 × x) + (1 + 10) × (1 − x)

↓

1.9 = (1 × x) + (1 + 10) × (1 − x)
1.9 = x + 11 × (1 − x)
1.9 = x + 11 − 11x
1.9 = 11 − 11x + x
1.9 = 11 − 10x
10x = 11 − 1.9
x = (11 − 1.9) / 10
x = 9.1 / 10 = 0.91 = **91%**

↓

그런데, 정답에는 **91%가 없음**
가장 가까운 값인 **90%**를 정답으로 함**(출제를 잘못한 부분)**

정답 : ③

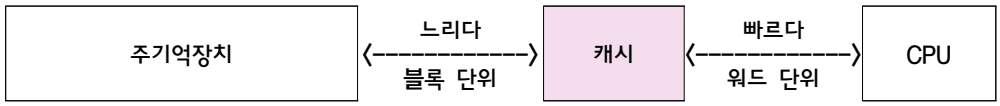
3. 캐시메모리와 사상(mapping)

주기억장치와 캐시메모리 사이의 정보 이동을 사상(mapping)이라 한다.

직접사상 : 주기억장치의 각 블록이 캐시의 정해진 **특정 슬롯(라인)**에만 적재 가능
연관사상 : 주기억장치의 각 블록이 캐시의 **임의의 어떤 슬롯(라인)**에 적재 가능
집합-연관사상 : 주기억장치의 각 블록이 캐시의 특정 **슬롯(라인) 집합 내**에 적재 가능

• 집합-연관 사상은 직접사상 방식과 연관사상 방식을 **혼합**한 방식이다.

주기억장치와 캐시메모리의 관계는 다음과 같다. 용어를 잘 정리하자!



물리주소	데이터
00 000	abcd
00 001	pass
00 010	raid
00 011	cost
:	:
10 010	book
:	:
11 111	disk

주기억장치

태그	데이터	라인
00	abcd	000
00	pass	001
10	book	010
:	:	:
11	disk	111

캐시메모리

• 사상 방식에 따라 물리주소를 다르게 해석
• 물리주소는 상황에 따라 태그, 라인, 집합으로 분리
• 여기서, 상황은 캐시메모리 크기에 의존
↓왼쪽 그림에서
• 태그 : 물리주소의 좌측 2비트
• 라인 : 물리주소의 우측 3비트

• **라인**(line) : 주기억장치의 한 블록이 저장되는 장소, **슬롯**(slot) 또는 **인덱스**라고도 함
• **태그**(tag) : 라인에 적재되어 있는 블록을 **구별**하기 위한 정보
• **블록**(block) : 주기억장치에서 한꺼번에 인출되는 데이터 그룹

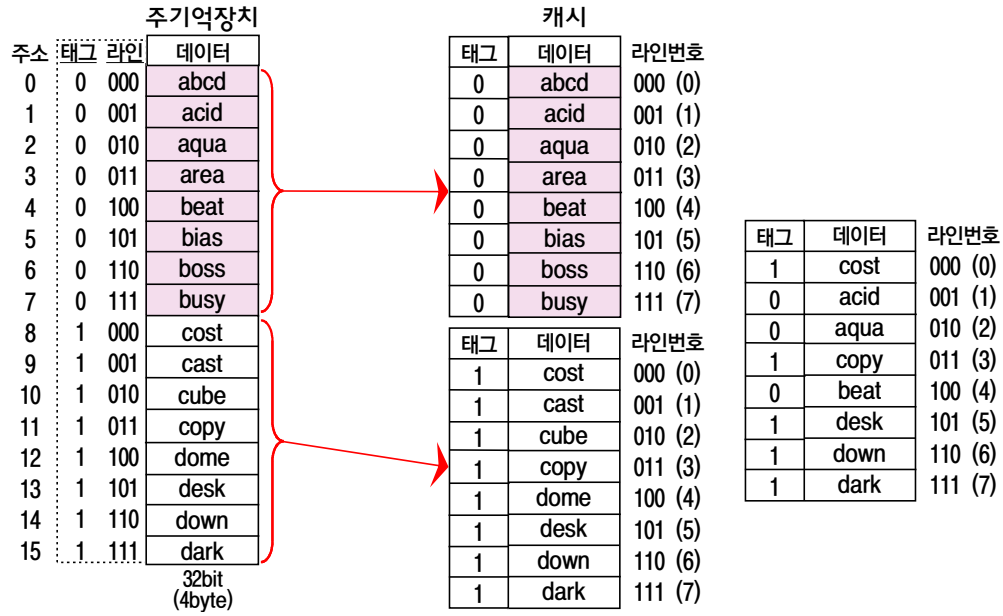
// 주기억장치와 캐시메모리 용량 비교

주기억장치	현재, 개인용 컴퓨터에 4GB~16GB 정도가 사용된다.
L1 캐시	보통 8~64KB 정도가 사용된다. (주기억장치 용량에 비해 매우 작다)
L2 캐시	보통 64KB~4MB 정도가 사용된다.

• 현재, 일부에서는 캐시 용량을 **72**MB까지 사용하고 있다.(게임 등을 위해)

(1) 직접사상(direct mapping)

다음은 직접사상에서 주기억장치의 데이터가 캐시에 적재될 수 있는 다양한 모습이다.



- 직접사상에서는 주기억장치의 **물리주소**를 캐시의 **태그**와 **라인**으로 구분한다.

블록 크기	4byte (주기억장치 하나의 주소가 4byte인 경우)
주기억장치 용량	16 × 4byte = 64byte
캐시 실제 용량	8 × 4byte = 32byte

- 주기억장치 주소 형식 :

태그 필드	라인 필드	워드(오프셋) 필드

↓ CPU가 생성하는 주소에 따른 데이터
↓ 하나의 블록 크기가 4바이트이고, 하나의 워드 크기가 1바이트인 경우

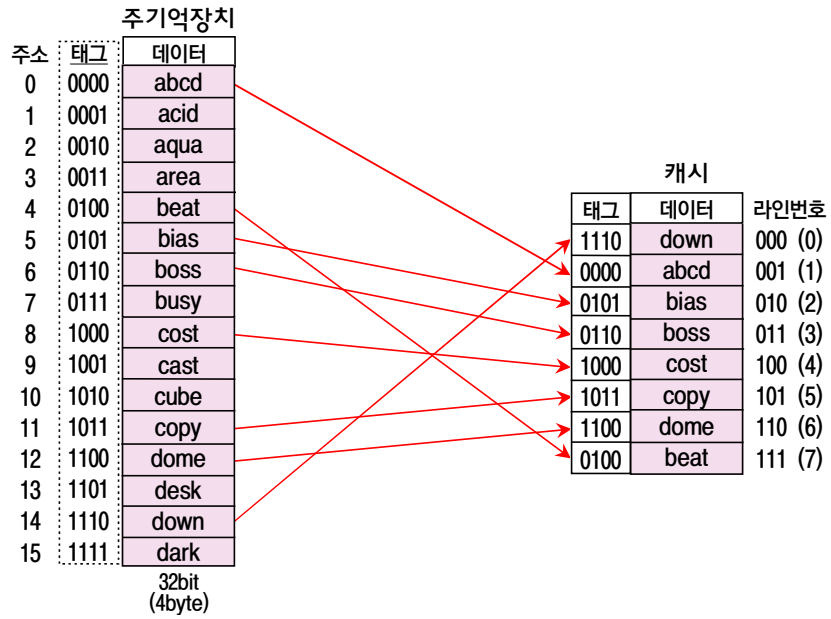
태그 번호	라인(인덱스) 번호	데이터	워드(오프셋) 번호	데이터
0	000	abcd	00	a
0	001	acid	01	c
1	000	cost	10	s
1	011	copy	11	y

주어진 수는 2진수이다.

- 직접사상은 주기억장치 블록이 **적재될 수 있는 캐시의 라인은 하나로 고정**되어 있다.
- 직접사상은 회로 구현은 간단하지만 캐시 **적중률이 떨어**진다.
- 직접사상은 **페이지 교체** 알고리즘이 필요 없다.(정해진 라인에만 적재되므로)

(2) 연관사상(associative mapping)

다음은 연관사상에서 주기억장치의 데이터가 캐시에 적재될 수 있는 모습이다.



- 연관사상에서는 주기억장치의 **물리주소**가 캐시의 **태그**가 된다.(**물리주소=태그**)
- 연관사상에서는 주기억장치의 **물리주소**에 캐시의 **라인(인덱스)**가 없다.

블록 크기	4byte (주기억장치 하나의 주소가 4byte인 경우)
주기억장치 용량	16 × 4byte = 64byte
캐시 실제 용량	8 × 4byte = 32byte

- 주기억장치 주소 형식 :

태그 필드	**워드(오프셋) 필드**

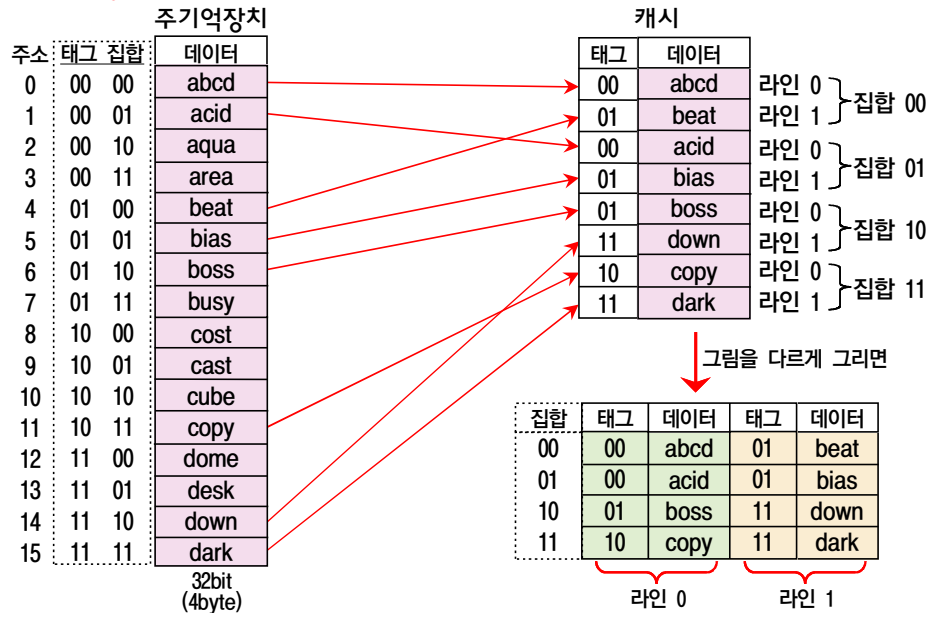
◈ **특징**

- 연관사상은 주기억장치 블록이 적재되는 캐시 라인이 정해져 있지 않다.
- 연관사상은 주기억장치 블록이 캐시에 신규 적재될 때, 캐시 라인 선택은 자유롭다.
- 연관사상은 **적중률이 향상**된다.
- 하지만, 캐시 **적중검사는 모든 라인에 대해 수행해야 하므로 검사시간 길어**진다.
- 모든 태그 번호를 고속으로 검사하기 위한 복잡한 회로 필요하다.(구현 비용 고가)
- 캐시 적중이 실패하면, **페이지 교체가 필요**하다.

(3) 집합-연관 사상(set-associative mapping)

다음은 집합-연관 사상에서 주기억장치의 데이터가 캐시에 적재될 수 있는 모습이다.

〈2-way 집합연관사상〉



- 집합-연관사상에서는 주기억장치의 **물리주소**를 캐시의 **태그**와 **집합**으로 구분한다.

블록 크기	4byte (주기억장치 하나의 주소가 4byte인 경우)
주기억장치 용량	16 × 4byte = 64byte
캐시 실제 용량	8 × 4byte = 32byte

- 각 집합이 k개의 라인으로 구성되면, **k-way 집합연관사상**이라 한다.
- 2-way 집합연관사상은 하나의 집합이 2개의 라인으로 구성되고
- 4-way 집합연관사상은 하나의 집합이 4개의 라인으로 구성된다.

- 주기억장치 주소 형식 :

태그 필드	집합 필드	워드(오프셋) 필드

◈ **특징**

- 집합-연관 사상은 **직접사상과 연관사상 방식을 조합**한 방식이다.
- 집합-연관 사상은 하나의 집합 영역에 서로 다른 태그를 갖는 다수의 워드로 구성된다.
- 캐시 적중이 실패하면, **페이지 교체가 필요**하다.
- **예** : 2-way 집합연관사상에서 집합 필드가 4bit일 때, 캐시 라인 수 = $2 \times 2^4 = 32$

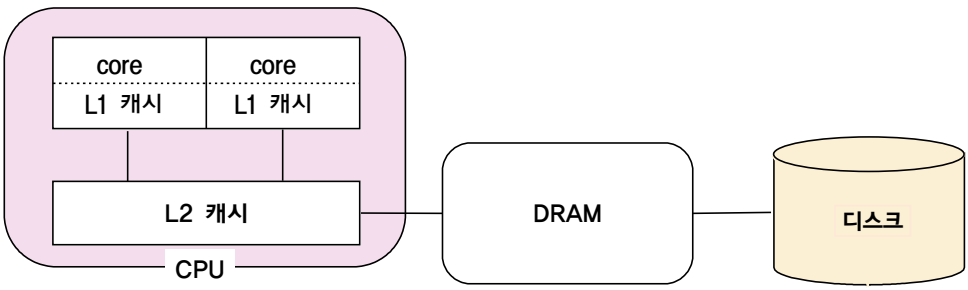
기출문제 분석

1. CPU와 메인 메모리의 속도 차이 때문에 발생하는 명령어 처리 성능 저하 현상을 방지하기 위하여, CPU와 메인 메모리 사이에 설치하는 메모리로 옳은 것은? [2021년 군무원 9급]

① 레지스터(register)
② ROM(Read Only Memory)
③ 캐시(cache)
④ I/O 버퍼(buffer)

♤ 캐시(cache)

〈Dual CPU core chip 컴퓨터 구조〉



정답 : ③

2. 열거된 메모리들을 처리속도가 빠른 순서대로 바르게 나열한 것은? [2014년 국가 9급]

ㄱ. 가상(virtual) 메모리
ㄴ. L1 캐시(Level 1 cache) 메모리
ㄷ. L2 캐시(Level 2 cache) 메모리
ㄹ. 임의 접근 메모리(RAM)

① ㄱ－ㄴ－ㄷ－ㄹ
② ㄴ－ㄷ－ㄹ－ㄱ
③ ㄷ－ㄴ－ㄱ－ㄹ
④ ㄹ－ㄱ－ㄴ－ㄹ

♤ 메모리 속도

• L1 캐시 〉 L2 캐시 〉 RAM 〉 가상메모리(디스크)

정답 : ②

3. 다음은 캐시기억장치를 사상(mapping) 방식 기준으로 분류한 것이다. 캐시 블록은 4개 이상이고 사상 방식을 제외한 모든 조건이 동일하다고 가정할 때, 평균적으로 캐시 적중률(hit ratio)이 높은 것에서 낮은 것 순으로 바르게 나열한 것은? [2015년 국가 9급]

ㄱ. 직접사상(direct-mapped)
ㄴ. 완전연관(fully-associative)
ㄷ. 2-way 집합연관(set-associative)

---

① ㄱ-ㄴ-ㄷ　② ㄴ-ㄷ-ㄱ　③ ㄷ-ㄱ-ㄴ　④ ㄱ-ㄷ-ㄴ

♤ **캐시에서 적중률**

- **적중률** : 완전연관 〉 집합연관 〉 직접사상
- **연관사상**은 주기억장치의 블록이 캐시의 어느 라인에도 적재될 수 있다.
- **연관사상**은 **적중률이 향상**된다.
- 하지만, 캐시 **적중검사는 모든 라인에 대해 수행해야 하므로 검사시간 길어**진다.
- **직접사상**은 주기억장치 블록이 **적재될 수 있는 캐시의 라인은 하나로 고정**되어 있다.
- **직접사상**은 회로 구현은 간단하지만 캐시 **적중률이 떨어**진다.

정답 : ②

4. 캐시(cache)에 대한 설명으로 옳지 않은 것은? [2021년 지방 9급]

① CPU와 인접한 곳에 위치하거나 CPU 내부에 포함되기도 한다.
② CPU와 상대적으로 느린 메인(main) 메모리 사이의 속도 차이를 줄이기 위해 사용된다.
③ 다중프로세서 시스템에서는 write-through 정책을 사용하더라도 데이터 불일치 문제가 발생할 수 있다.
④ 캐시에 쓰기 동작을 수행할 때 메인 메모리에도 동시에 쓰기 동작이 이루어지는 방식을 write-back 정책이라고 한다.

♤ **캐시 정책**

Write-through (즉시쓰기)	• 쓰기 동작이 발생할 때마다 **캐시와 주기억장치의 내용을 한꺼번에 갱신**한다. • 처리는 간단하지만, 동시 갱신으로 시간지연 발생(성능저하) • 성능향상을 위해 **쓰기 버퍼** 사용(**버퍼링된 Write-through**)
Write-back (나중쓰기)	• 쓰기 동작 발생 시 **캐시만 갱신**한다.(처리는 복잡, 성능은 우수) • 나중에 메모리를 갱신할 수 있도록 위치를 표시해 둔다. • Write-back에서도 **쓰기 버퍼**를 사용한다.(블록이 교체될 때 갱신된 것 쓰기 수행)

정답 : ④

5. 2-way 집합연관사상(set-associative mapping) 방식을 사용하는 캐시기억장치를 가진 컴퓨터가 있다. 캐시기억장치 접근(access)을 위해 주기억장치 주소가 다음 3개의 필드(field)로 구분된다면, 캐시기억장치의 총 라인(line) 개수는? [2018년 지방 9급]

태그(tag) 필드	세트(set) 필드	오프셋(offset) 필드
8비트	9비트	7비트

① 128개 ② 256개
③ 512개 ④ 1,024개

♤ **집합연관사상(set-associative mapping)**

• 2-way 집합연관사상 캐시 구조

집합(set)	태그	데이터	태그	데이터
0	0000 0000	대한민국	0000 0000	프로야구
1	1000 0000	연개소문	1000 0000	프로축구
2				
:				
:				
510	0100 0000	자료구조	1111 0000	메인보드
511	0111 0000	정보보호	1111 0000	합격하자
	라인 0		라인 1	

• **세트(set) 필드**가 9비트이므로, 캐시메모리의 **집합 수** = 2^9 = **512**(개)
→ 주기억장치 주소에서 세트 필드가 9bit라는 것은 집합 수를 9bit 표현한다는 의미
• 2-way 집합연관사상이므로 **각 집합**에는 **2개의 라인**이 존재한다.
• 캐시기억장치의 **총 라인(line) 개수** = 2 × 2^9 = 2 × 512 = **1,024(개)**

• 라인(line) : 주기억장치의 각 블록이 저장되는 캐시 블록(index라고도 함)
• 태그(tag) : 라인에 적재된 블록이 주기억장치 어느 곳 블록인지 구별하기 위한 것
• 블록(block) : 주기억장치로부터 동시에 인출되는 정보 그룹
• 오프셋(offset, 변위) : 블록에서 상대적인 위치를 나타내는 값(문제 푸는데 상관없는 것!)

• 집합연관사상은 직접사상과 연관사상 방식을 조합한 방식이다.
• 하나의 주소 영역이 서로 다른 태그를 갖는 여러 개의 집합으로 이루어지는 방식이다.

정답 : ④

6. 캐시기억장치에 대한 설명으로 옳지 않은 것은? [2017년 경기 추가 9급]

① 명령어 캐시기억장치와 데이터 캐시기억장치로 분리된 구조를 가질 수 있다.
② 2개 이상의 단계(level)를 가지는 다단계 구조를 가질 수 있다.
③ 직접사상 방식을 사용할 경우, 적절한 교체(replacement) 알고리즘이 필요하다.
④ 쓰기 버퍼(write buffer)는 즉시 쓰기(write-through) 캐시기억장치에서 쓰기 동작이 오래 걸리는 문제를 개선할 수 있다.

♤ **캐시메모리에서 교체 알고리즘**

◉ **직접사상 방식에서 교체 알고리즘**

- 주기억장치의 각 블록이 캐시의 정해진 어느 **특정 슬롯(라인)**에만 적재될 수 있는 기법이다.
- 주기억장치 블록이 적재될 수 있는 캐시의 슬롯은 고정되어 있다.
- 주기억장치의 데이터가 캐시의 동일 슬롯에 저장되므로 **교체 알고리즘이 필요 없다.**

◉ **연관사상 및 집합연관사상 방식에서 교체 알고리즘**

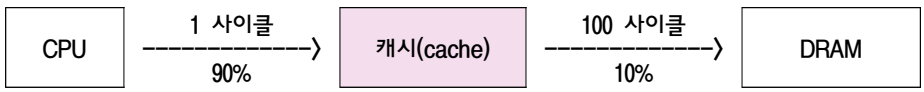
- 교체 알고리즘이 필요하다.
- 주기억장치의 데이터가 캐시의 다른 슬롯에 저장될 수 있으므로 **교체 알고리즘이 필요**하다.

정답 : ③

7. CPU와 DRAM 사이에 캐시(cache)가 있는 구조에서, CPU가 캐시와 DRAM을 접근하는데 각각 1 사이클과 100 사이클이 소요된다고 가정하자. 캐시 적중률(hit ratio)이 90%라고 할 때 평균 메모리 접근시간은? [2015년 국회 9급]

① 1.1 사이클 ② 1.9 사이클 ③ 10.1 사이클
④ 10.9 사이클 ⑤ 11 사이클

♤ **평균 메모리 접근시간**



- 평균시간 = (캐시 접근 사이클 × 확률) + (DRAM 접근 사이클 × 확률)
  = 1 × 0.9 + (1 + 100) × 0.1
  = 0.9 + 10.1
  = **11**(사이클)

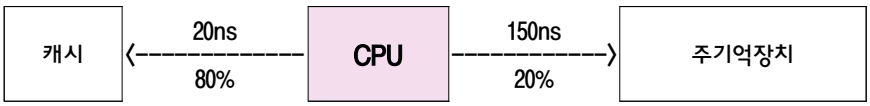
정답 : ⑤

8. 캐시기억장치 접근시간이 20ns, 주기억장치 접근시간이 150ns, 캐시기억장치 적중률이 80%인 경우에 평균 기억장치 접근시간은? (단, 기억장치는 캐시와 주기억장치로만 구성된다) [2020년 지방 9급]

① 32ns ② 46ns
③ 124ns ④ 170ns

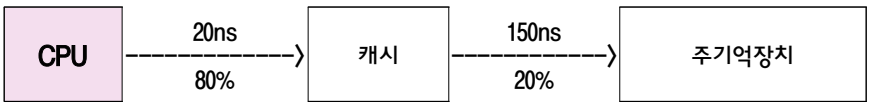
♤ **평균 기억장치 접근시간**

//풀이 1 : 캐시 적중률이 80%일 때



평균 기억장치 접근시간 = (캐시 접근시간 × 캐시 적중률)
+ (주기억장치 접근시간) × (1 − 캐시 적중률)
평균 기억장치 접근시간 = (20 × 0.8) + (150) × (1 − 0.8)
평균 기억장치 접근시간 = (20 × 0.8) + (150 × 0.2) = 16 + 30 = **46**ns

//풀이 2 : 캐시 적중률이 80%일 때



평균 기억장치 접근시간 = (캐시 접근시간 × 캐시 적중률)
+ (캐시 접근시간 + 주기억장치 접근시간) × (1 − 캐시 적중률)
평균 기억장치 접근시간 = (20 × 0.8) + (20 + 150) × (1 − 0.8)
평균 기억장치 접근시간 = (20 × 0.8) + (170 × 0.2) = 16 + 34 = **50**ns

//문제 분석----------

- 먼저, 정답은 **②번으로 발표**되었다. 출제자 풀이 방법이다.
- 그런데, CPU가 캐시를 거치지 않고 바로 주기억장치에 접근하는 것은 아니다.
- 해서, **풀이 1**은 논리적으로 맞지 않는 것이다.**(비정상적인 풀이가 정답으로 발표되는 현실)**
- 만약, **풀이 1**처럼 풀려면 주기억장치 접근시간에 캐시 접근시간이 포함된 단서가 있어야 한다.

정답 : ②