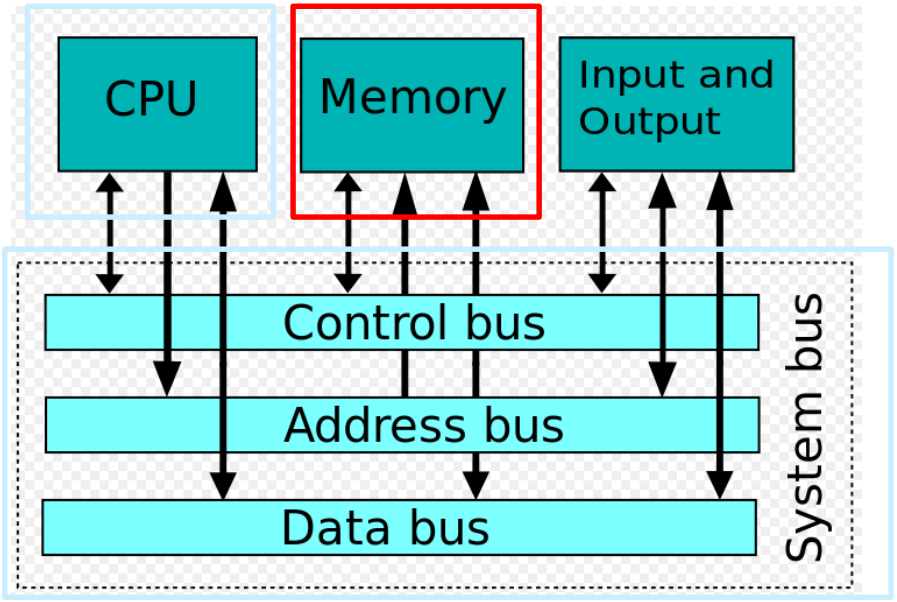
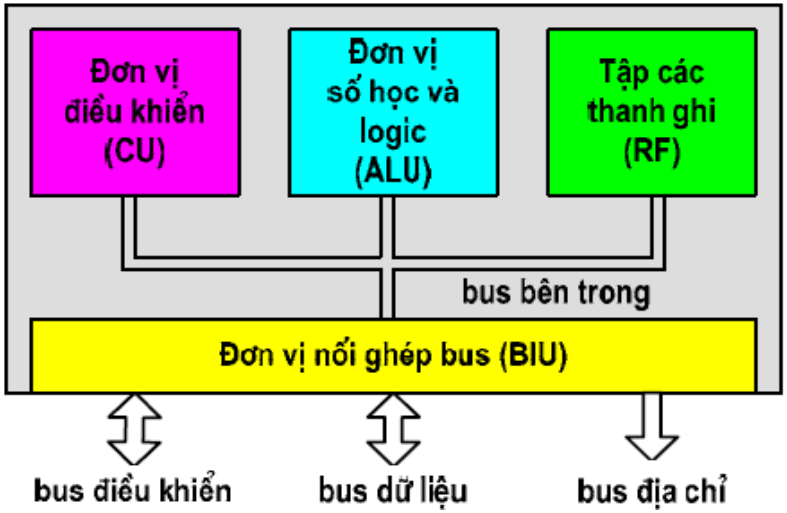
Chương 5

Hệ thống nhớ máy tính

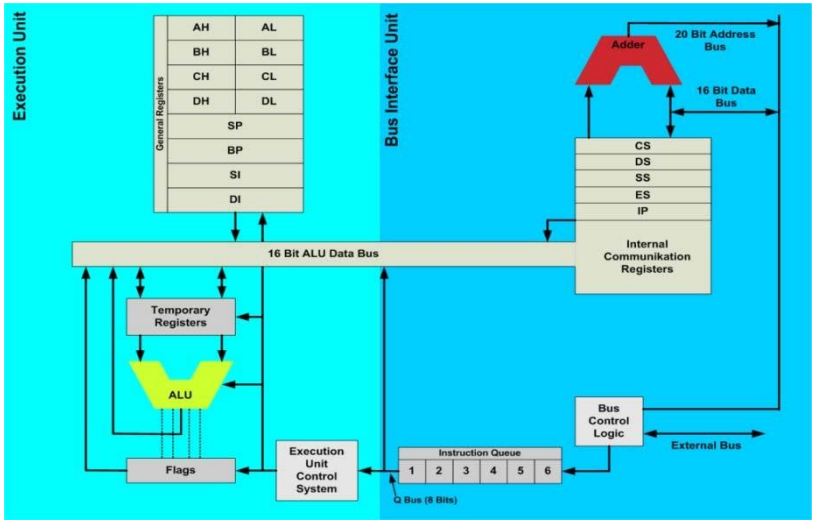


Giảng viên: ThS. Phan Như Minh

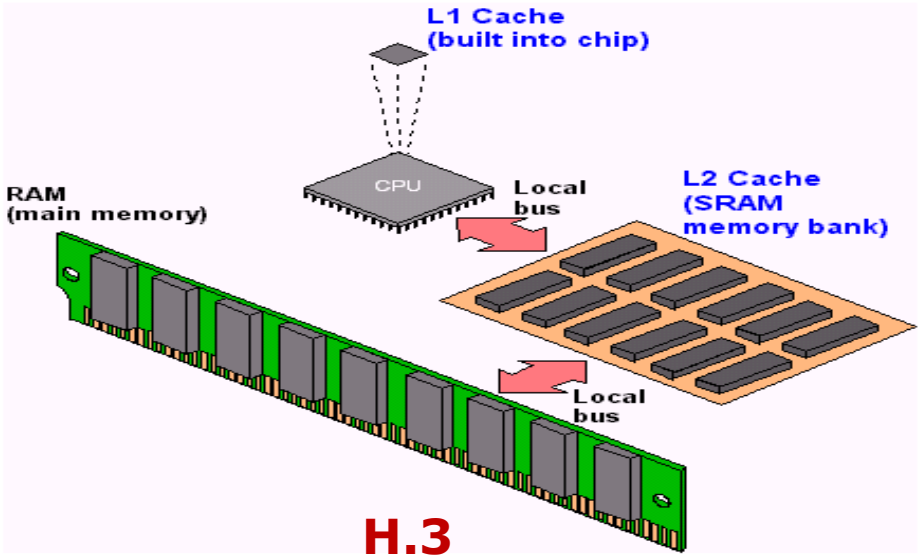
Quan sát và định vị các thành phần nhớ?



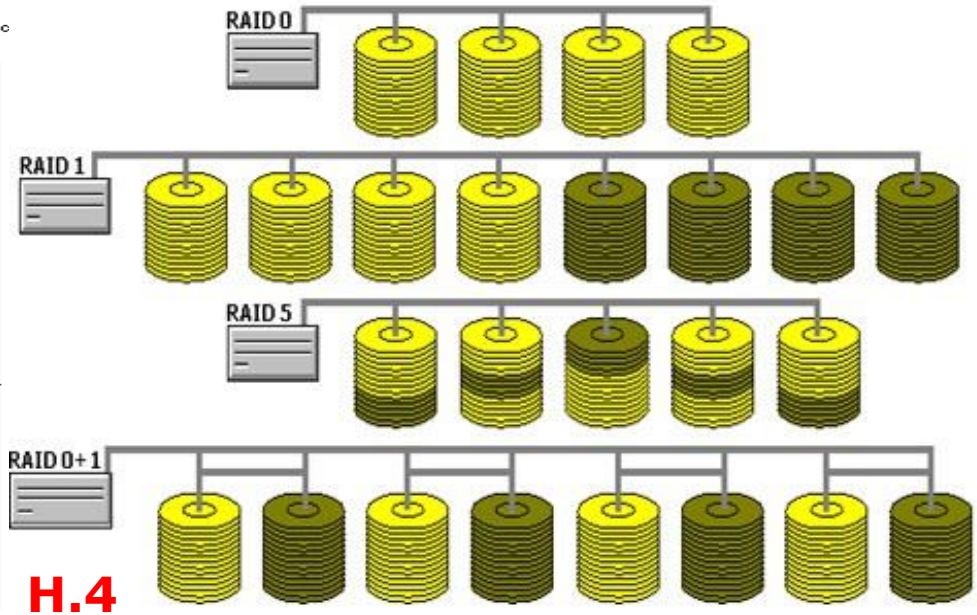
H.1



H.3



H.3



H.4

5.3 Bộ nhớ trong:

Bộ nhớ trong gồm 2 loại:

- Bộ nhớ chính (Main Memory)
- Bộ nhớ đệm nhanh (Cache Memory)

❖ Bộ nhớ chính: Các đặc trưng cơ bản

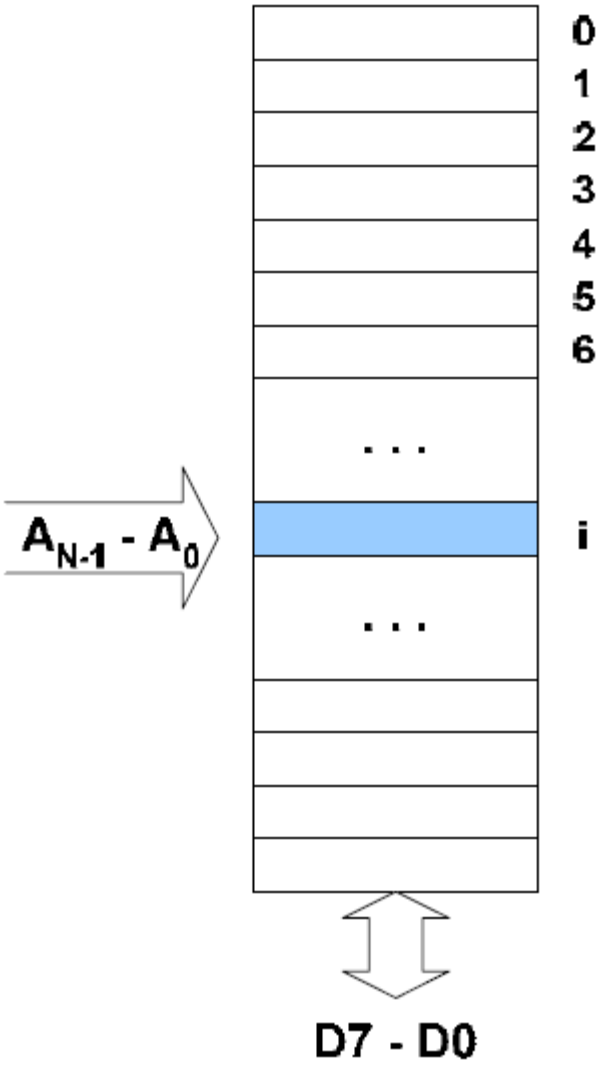
- Chứa các chương trình đang thực hiện và các dữ liệu đang được sử dụng
- Tồn tại trên mọi hệ thống máy tính
- Bao gồm các ngăn nhớ được đánh địa chỉ trực tiếp bởi CPU
- Dung lượng của bộ nhớ chính nhỏ hơn không gian địa chỉ bộ nhớ mà CPU quản lý.
- Việc quản lý logic bộ nhớ chính tuỳ thuộc vào hệ điều hành

Tổ chức bộ nhớ đan xen (interleaved memory)

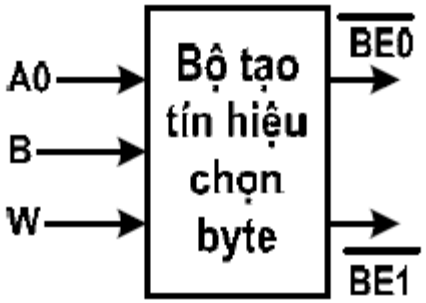
- Độ rộng của bus dữ liệu để trao đổi với bộ nhớ: m = 8, 16, 32, 64,128 ... bit
- Các ngăn nhớ được tổ chức theo byte

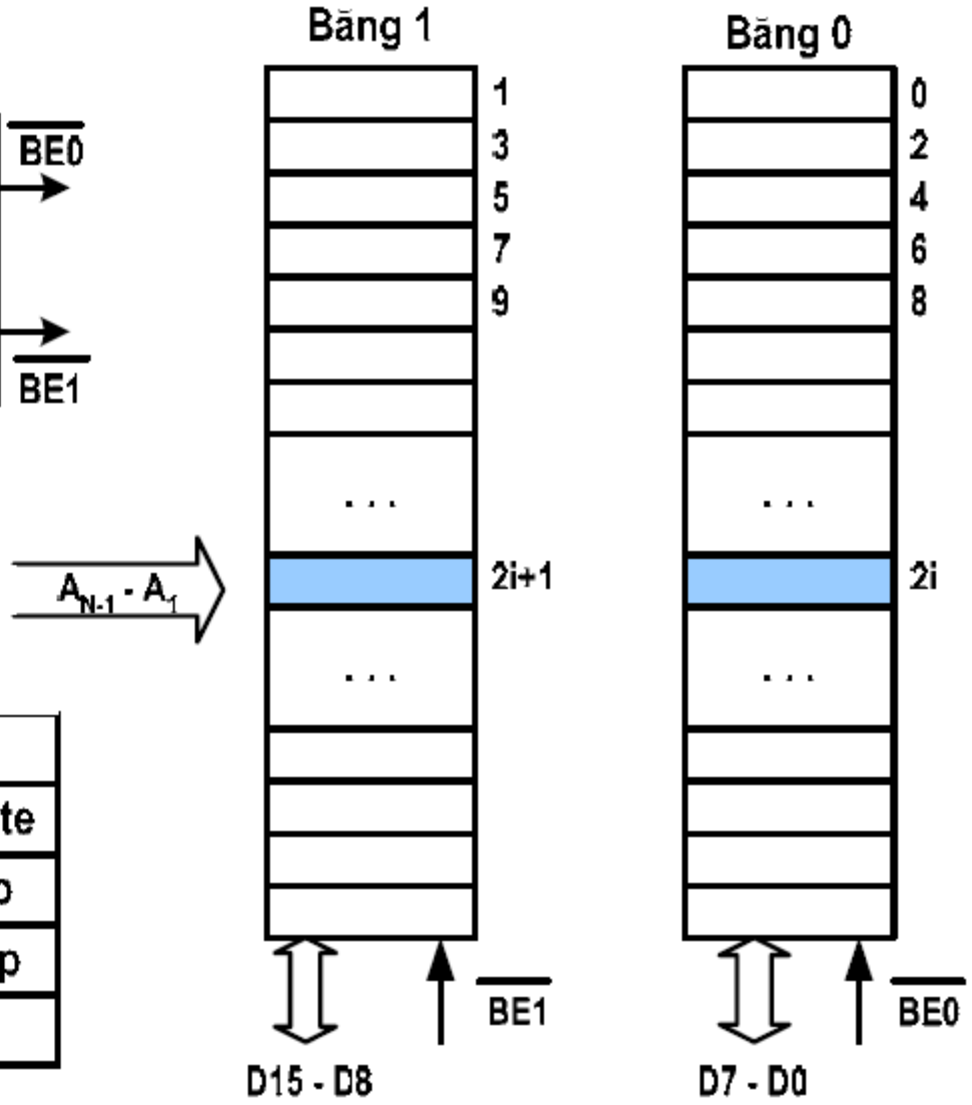
=> tổ chức bộ nhớ vật lý khác nhau

m=8bit => một băng nhớ tuyến tính



m = 16bit=>hai băng nhớ đan xen

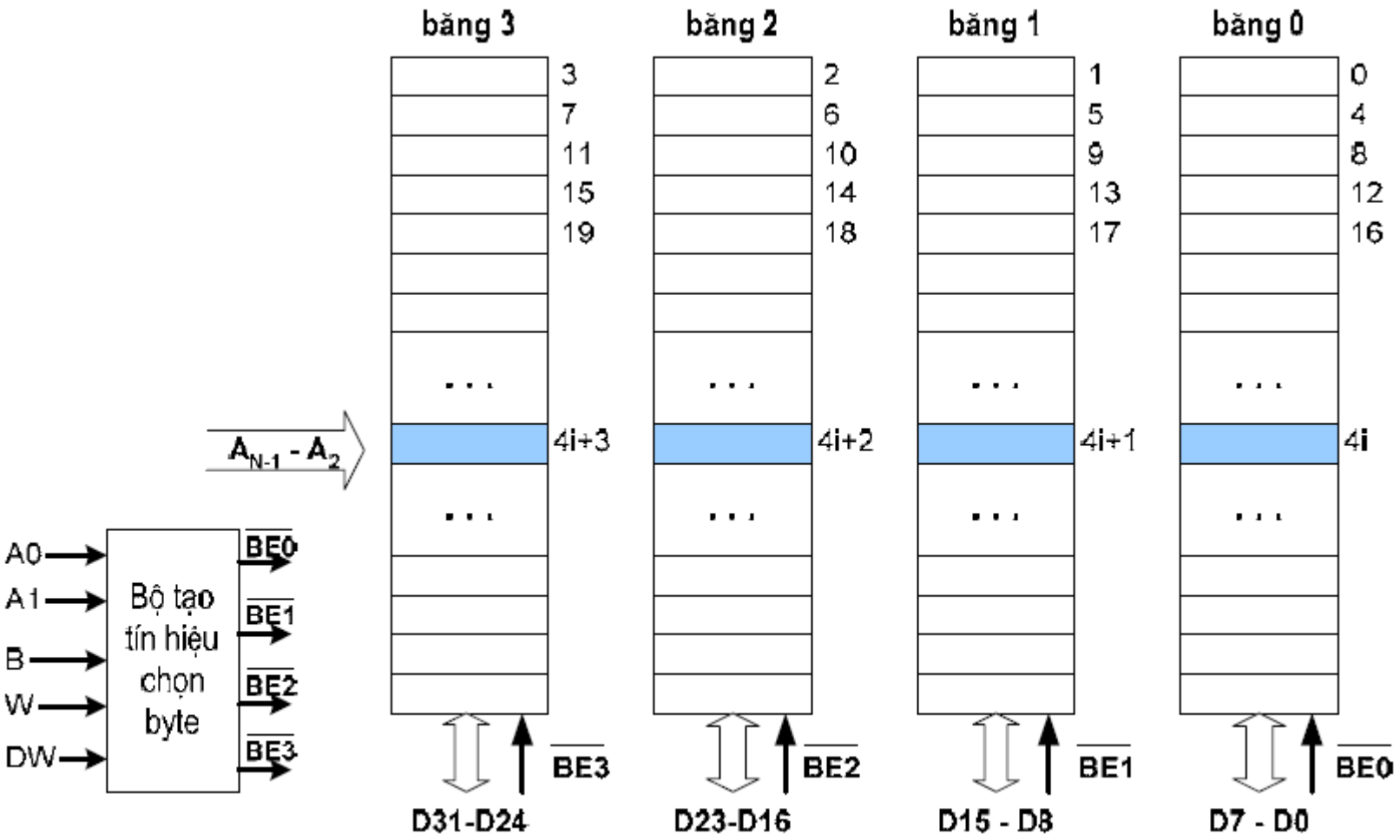




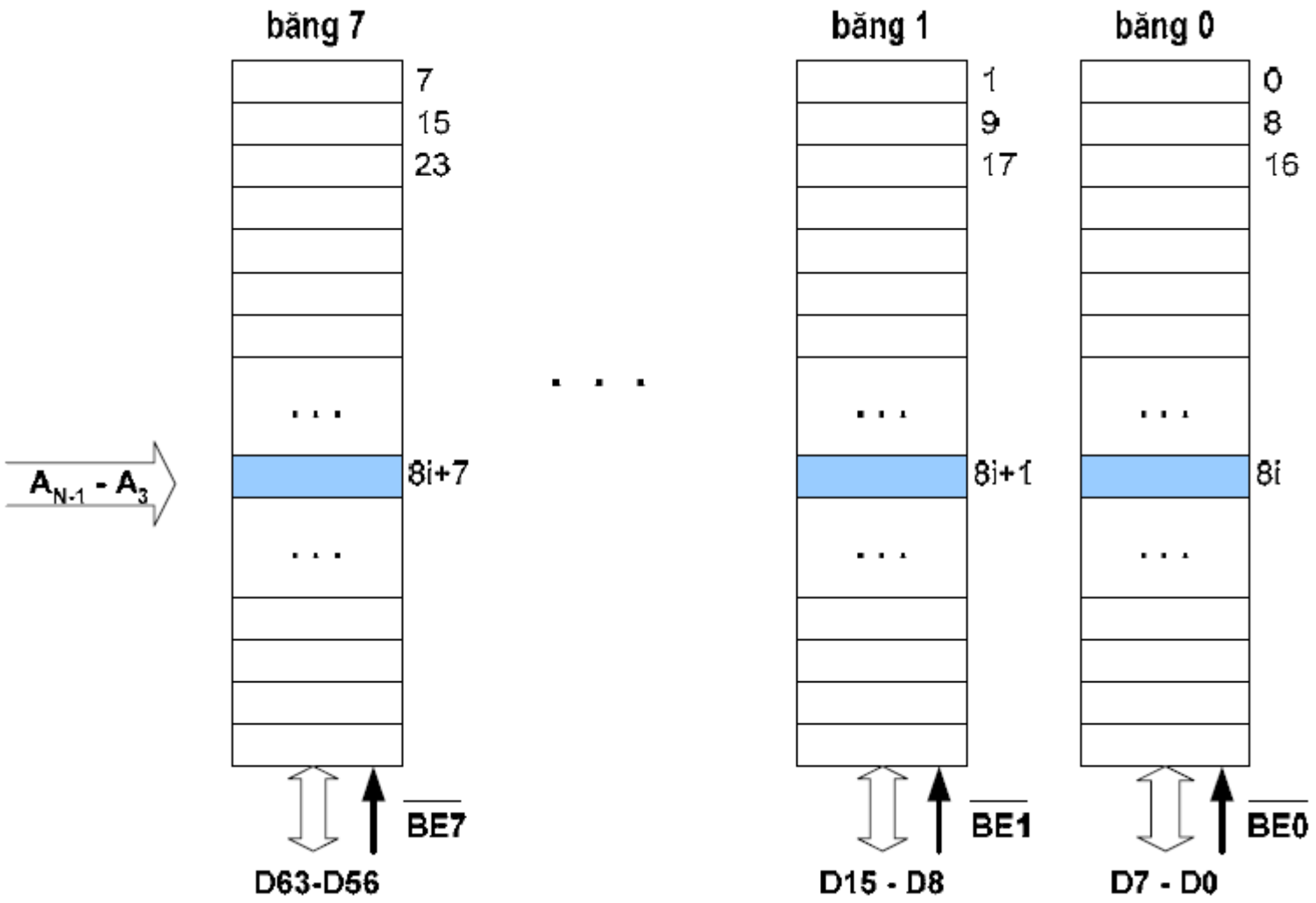
Các tín hiệu chọn byte

$\overline{BE1}$	$\overline{BE0}$	Chọn byte
0	0	Chọn cả hai byte
0	1	Chọn byte cao
1	0	Chọn byte thấp
1	1	không chọn

m = 32bit => bốn băng nhớ đan xen



m = 64bit => tám băng nhớ đan xen



❖ Bộ nhớ đệm nhanh (cache memory)

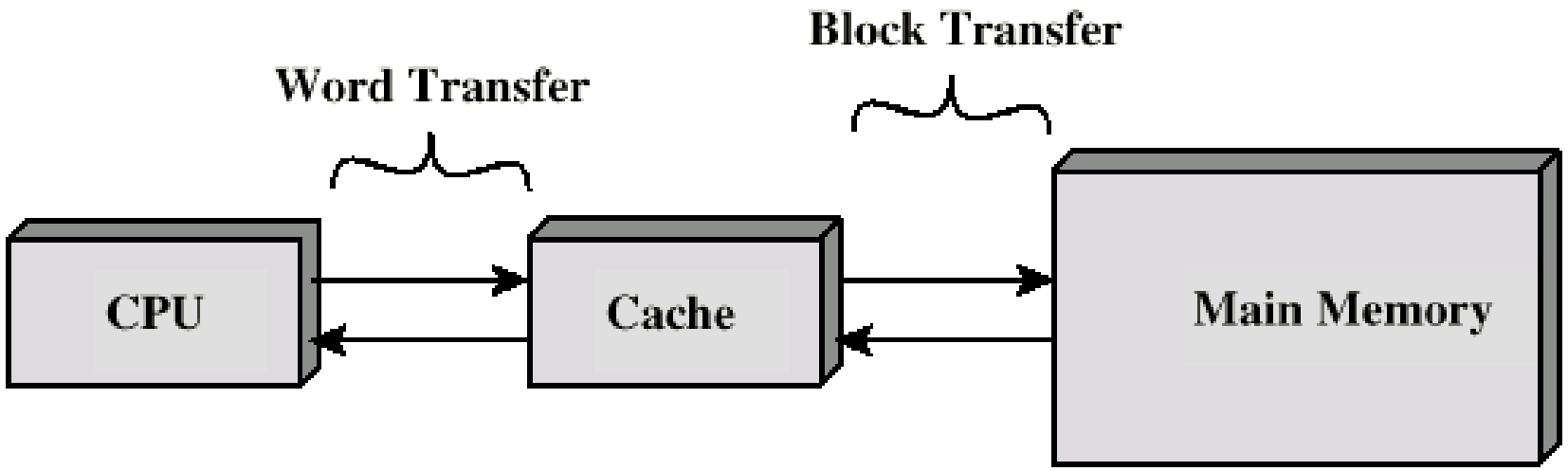
- ❖ Nguyên tắc chung của cache
- ❖ Các phương pháp ánh xạ
- ❖ Thuật toán thay thế
- ❖ Phương pháp ghi dữ liệu cache
- ❖ Một số loại cache

1. Nguyên tắc chung của cache

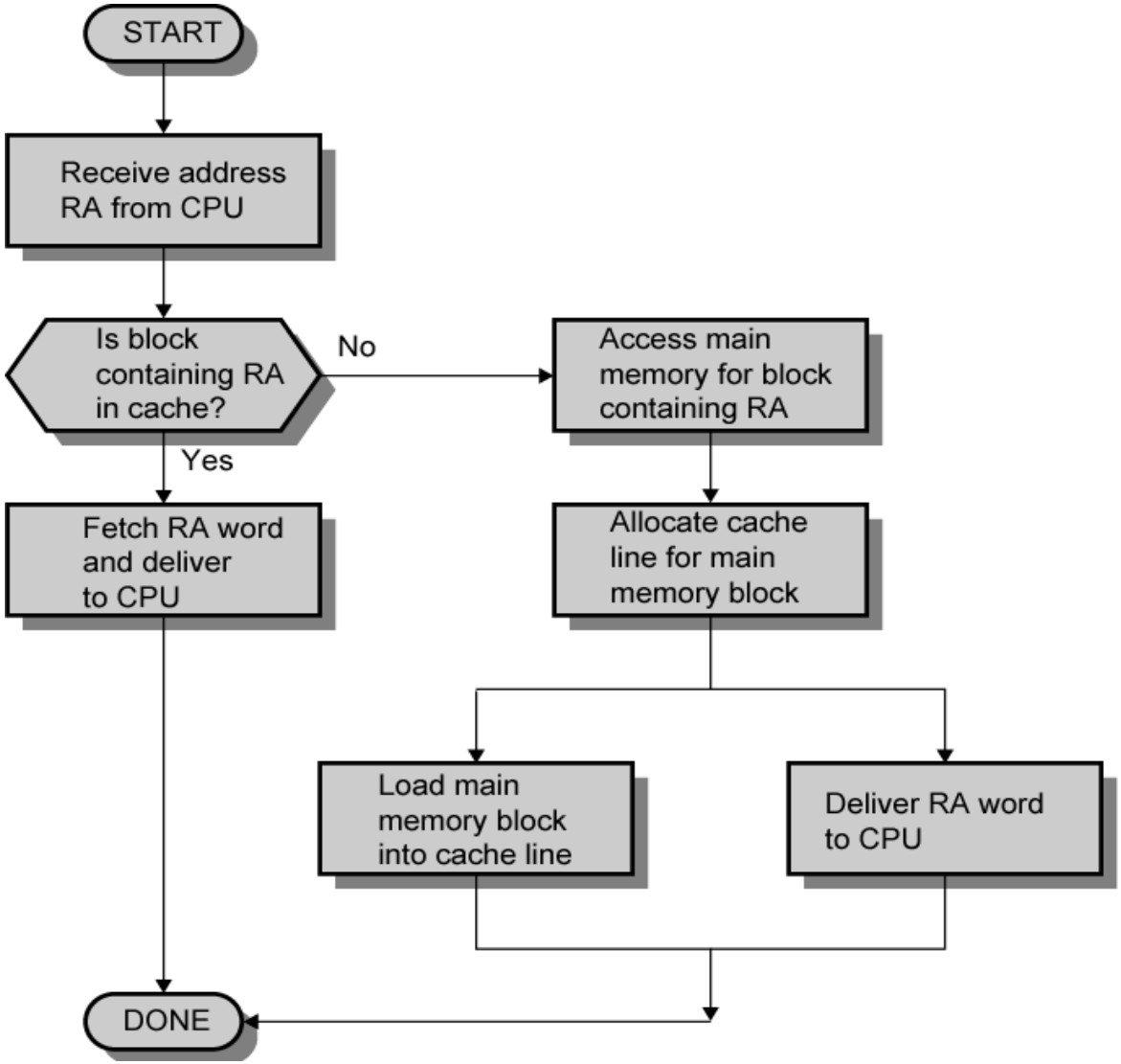
- Nguyên lý cục bộ hoá tham chiếu bộ nhớ: Trong một khoảng thời gian đủ nhỏ CPU thường chỉ tham chiếu các thông tin trong một khối nhớ cục bộ
- Ví dụ:
  - Cấu trúc chương trình tuần tự
  - Vòng lặp có thân nhỏ
  - Cấu trúc dữ liệu mảng

Nguyên tắc chung của cache (tiếp)

- Cache có tốc độ nhanh hơn bộ nhớ chính
- Cache được đặt giữa CPU và bộ nhớ chính nhằm tăng tốc độ CPU truy cập bộ nhớ
- Cache có thể được đặt trên chip CPU



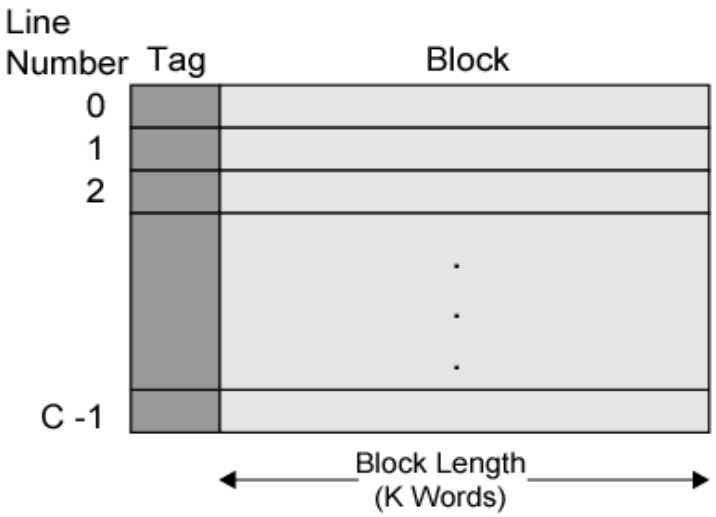
Ví dụ về thao tác của cache



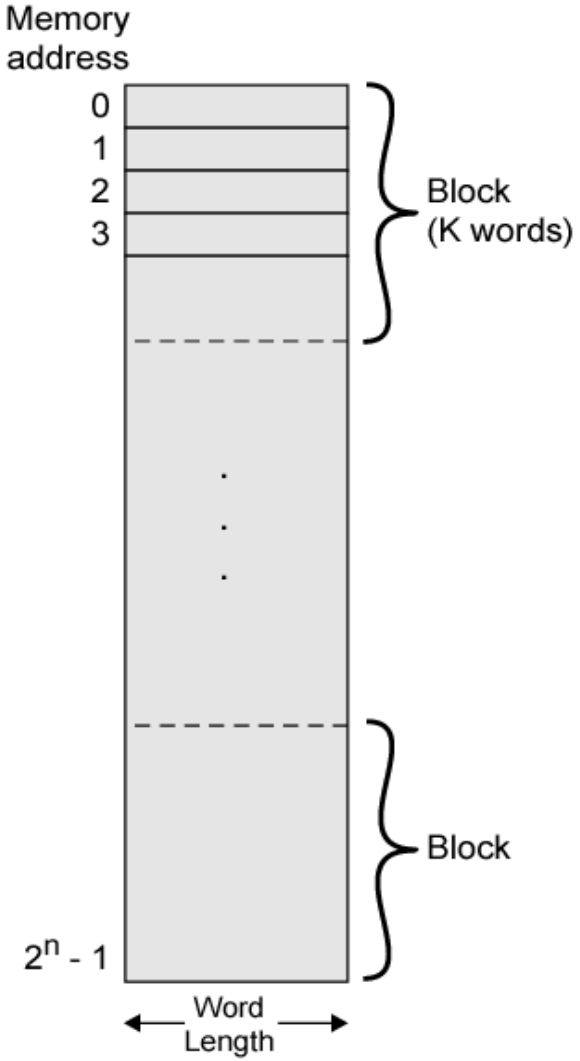
Ví dụ về thao tác của cache (tiếp)

- CPU yêu cầu nội dung của ngăn nhớ
- CPU kiểm tra trên cache với dữ liệu này
- Nếu có, CPU nhận dữ liệu từ cache (nhanh)
- Nếu không có, đọc Block nhớ chứa dữ liệu từ bộ nhớ chính vào cache
- Tiếp đó chuyển dữ liệu từ cache vào CPU

Cấu trúc chung của cache / bộ nhớ chính



(a) Cache



(b) Main memory

Cấu trúc chung của cache / bộ nhớ chính (tiếp)

- Bộ nhớ chính có 2^N byte nhớ
- Bộ nhớ chính và cache được chia thành các khối có kích thước bằng nhau
- Bộ nhớ chính: $B_0, B_1, B_2, \ldots, B_{p-1}$ (p Blocks)
- Bộ nhớ cache: $L_0, L_1, L_2, \ldots, L_{m-1}$ (m Lines)
- Kích thước của Block = 8,16,32,64,128 byte

Cấu trúc chung của cache / bộ nhớ chính (tiếp)

- Một số Block của bộ nhớ chính được nạp vào các Line của cache.
- Nội dung Tag (thẻ nhớ) cho biết Block nào của bộ nhớ chính hiện đang được chứa ở Line đó.
- Khi CPU truy nhập (đọc/ghi) một từ nhớ, có hai khả năng xảy ra:
  - Từ nhớ đó có trong cache (cache hit)
  - Từ nhớ đó không có trong cache (cache miss).

2. Các phương pháp ánh xạ

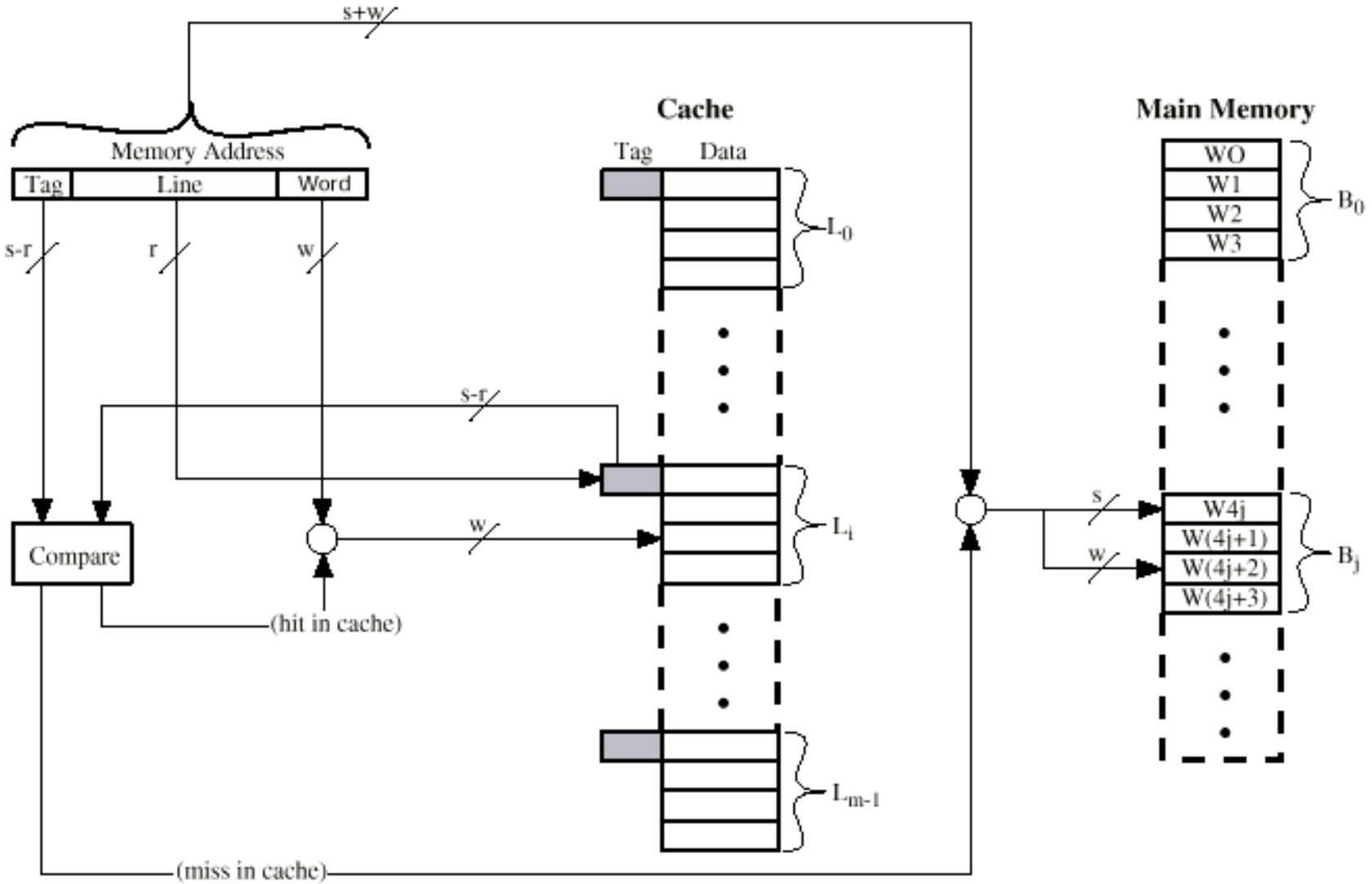
(Chính là các phương pháp tổ chức bộ nhớ cache)

- Ánh xạ trực tiếp (Direct mapping)
- Ánh xạ liên kết toàn phần (Fully associative mapping)
- Ánh xạ liên kết tập hợp (Set associative mapping)

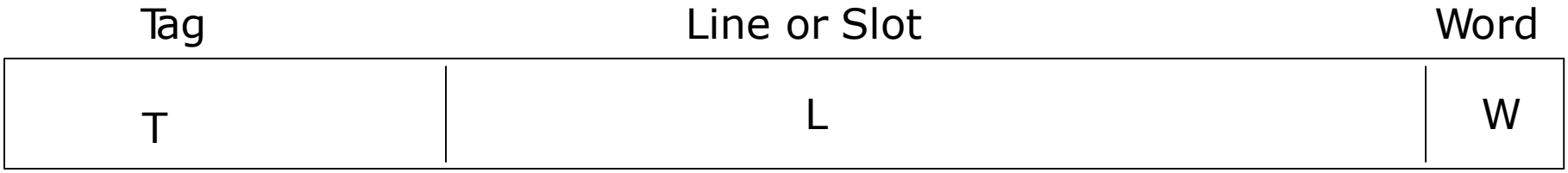
Ánh xạ trực tiếp

- Mỗi Block của bộ nhớ chính chỉ có thể được nạp vào một Line của cache:
  - $B_0 \Rightarrow L_0$
  - $B_1 \Rightarrow L_1$
  - ....
  - $B_{m-1} \Rightarrow L_{m-1}$
  - $B_m \Rightarrow L_0$
  - $B_{m+1} \Rightarrow L_1$
  - ....
- Tổng quát
  - B_j chỉ có thể nạp vào $L_{j \bmod m}$
  - m là số *Line của cache.*

Minh hoạ ánh xạ trực tiếp



Đặc điểm của ánh xạ trực tiếp



- Mỗi một địa chỉ N bit của bộ nhớ chính gồm ba trường:
  - *Trường Word* gồm W bit xác định một từ nhớ trong *Block hay Line:*

    2^W = kích thước của *Block hay Line*
  - *Trường Line* gồm L bit xác định một trong số các Line trong cache:

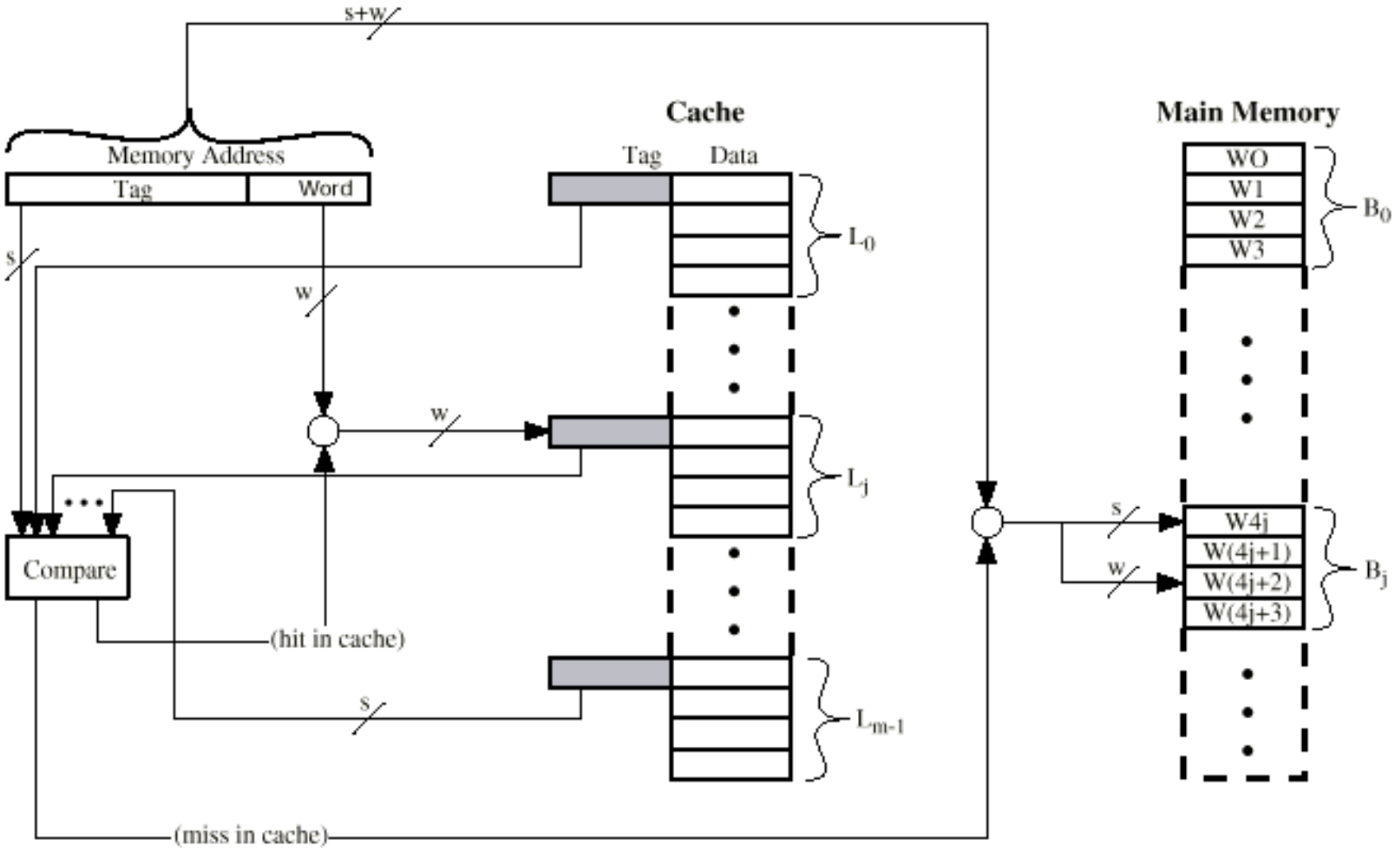
    2^L = số *Line trong cache* = *m*
  - *Trường Tag* gồm T bit:

    T = N - (W+L)
- Bộ so sánh đơn giản
- Xác suất *cache hit thấp*

Ánh xạ liên kết toàn phần

- Mỗi *Block* có thể nạp vào bất kỳ *Line* nào của *cache.*
- Địa chỉ của bộ nhớ chính bao gồm hai trường:
  - Trường Word giống như trường hợp ở trên.
  - Trường *Tag* dùng để xác định *Block* của bộ nhớ chính.
- Tag xác định Block đang nằm ở Line đó

Minh hoạ ánh xạ liên kết toàn phần



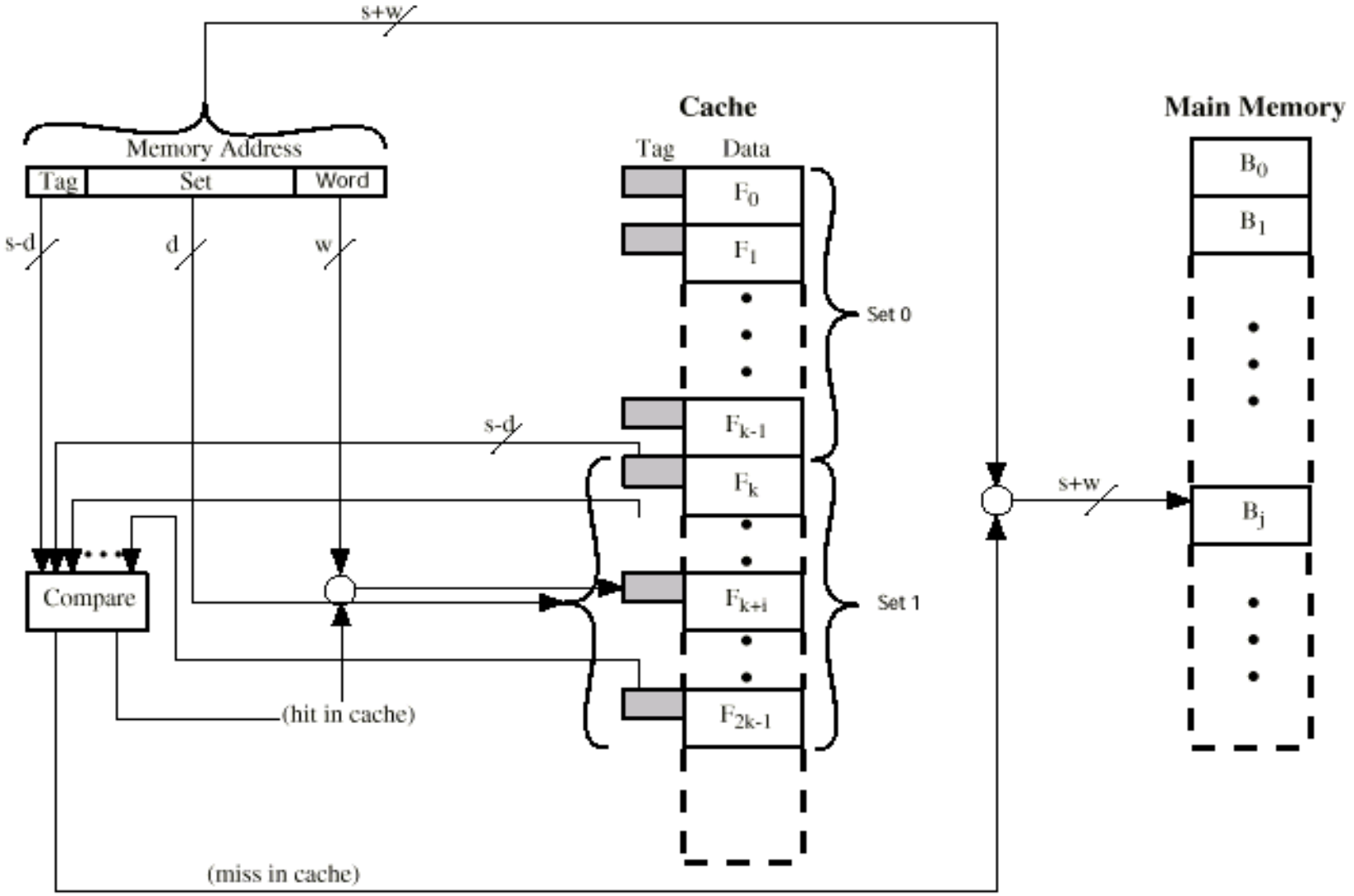
Đặc điểm của ánh xạ liên kết toàn phần

- So sánh đồng thời với tất cả các Tag => mất nhiều thời gian
- Xác suất *cache hit* cao.
- Bộ so sánh phức tạp.

Ánh xạ liên kết tập hợp

- Cache được chia thành các Tập (Set)
- Mỗi một Set chứa một số Line
- Ví dụ:
  - 4 Line/Set => 4-way associative mapping
- Ánh xạ theo nguyên tắc sau:
  - B_0 => S_0
  - B_1 => S_1
  - B_2 => S_2
  - .....

Minh hoạ ánh xạ liên kết tập hợp



Đặc điểm của ánh xạ liên kết tập hợp

- Kích thước *Block* = 2^W Word
- Trường *Set* có S bit dùng để xác định một trong số V = 2^S *Set*
- Trường *Tag có T bit:* T = N - (W+S)
- Tổng quát cho cả hai phương pháp trên
- Thông thường 2,4,8,16Lines/Set

Câu 15. Cho không gian địa chỉ bộ nhớ 32GB, dung lượng bộ nhớ Cache là 6MB, kích thước line là 32 byte. Xác định số bit của trường địa chỉ theo phương pháp:

a) Ánh xạ trực tiếp

b) Ánh xạ liên kết toàn phần

c) Ánh xạ liên kết tập hợp.

Câu 16. Cho không gian địa chỉ bộ nhớ 64GB, dung lượng bộ nhớ Cache là 8MB, kích thước line là 32 byte. Xác định số bit của trường địa chỉ theo phương pháp:

a) Ánh xạ trực tiếp

b) Ánh xạ liên kết toàn phần

c) Ánh xạ liên kết tập hợp.

Câu 17. Cho không gian địa chỉ bộ nhớ 128GB, dung lượng bộ nhớ Cache là 20MB, kích thước line là 32 byte. Xác định số bit của trường địa chỉ theo phương pháp:

a) Ánh xạ trực tiếp

b) Ánh xạ liên kết toàn phần

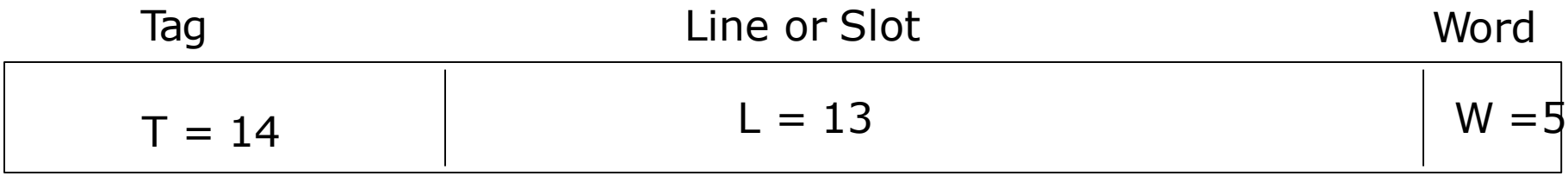
c) Ánh xạ liên kết tập hợp.

Bài toán: ánh xạ địa chỉ

- Không gian địa chỉ bộ nhớ chính = 4GB
- Dung lượng bộ nhớ *cache* là 256KB
- Kích thước *Line (Block)* = 32byte.
- Xác định số bit của các trường địa chỉ cho ba trường hợp tổ chức:
  - Ánh xạ trực tiếp
  - Ánh xạ liên kết toàn phần
  - Ánh xạ liên kết tập hợp 4 đường

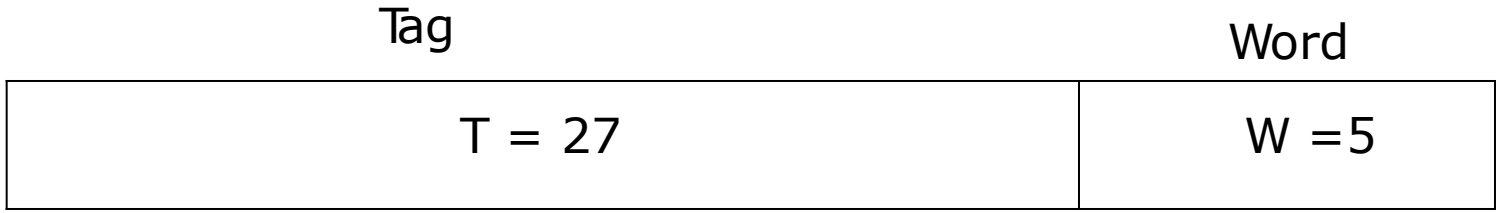
Với ánh xạ trực tiếp

- Bộ nhớ chính = 4GB = 2^{32} byte => N = 32 bit
- *Cache* = 256 KB = 2^{18} byte.
- *Line* = 32 byte = 2^5 byte => W = 5 bit
- Số *Line* trong *cache* = $2^{18}/ 2^5 = 2^{13}$ Line
- L = 13 bit
- T = 32 - (13 + 5) = 14 bit



Với ánh xạ liên kết toàn phần

- Bộ nhớ chính = 4GB = 2^{32} byte => N = 32 bit
- *Line* = 32 byte = 2^5 byte => W = 5 bit
- Số bit của trường *Tag* sẽ là: T = 32 - 5 = 27 bit

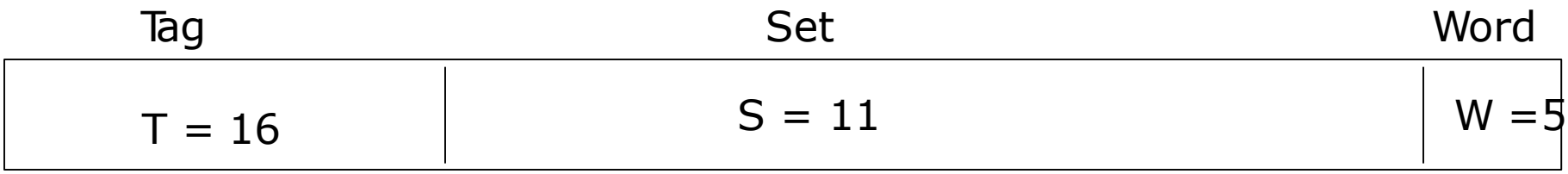


Với ánh xạ liên kết tập hợp 4 đường

- Bộ nhớ chính = 4GB = 2^{32} byte => N = 32 bit
- *Line* = 32 byte = 2^5 byte => W = 5 bit
- Số *Line* trong *cache* = $2^{18}/ 2^5 = 2^{13}$ Line
- Một *Set có 4 Line* = 2^2 Line

số *Set trong cache* = $2^{13}/ 2^2 = 2^{11}$ Set => S = 11 bit

- Số bit của trường *Tag* sẽ là: T = 32 - (11 + 5) = = 16 bit



Bài tập

- Giả thiết rằng máy tính có 128KB cache tổ chức theo kiểu ánh xạ liên kết tập hợp 4-line. Cache có tất cả là 1024 Set từ S_0 đến S_{1023}. Địa chỉ bộ nhớ chính là 32-bit và đánh địa chỉ cho từng byte.

a) Tính số bit cho các trường địa chỉ khi truy nhập cache?

b) Xác định byte nhớ có địa chỉ 003D02AF(16) được ánh xạ vào Set nào của cache?

3. Thuật giải thay thế (1): Ánh xạ trực tiếp

- Không phải lựa chọn
- Mỗi Block chỉ ánh xạ vào một Line xác định
- Thay thế Block ở Line đó

Thuật giải thay thế (2): Ánh xạ liên kết

- Được thực hiện bằng phần cứng (nhanh)
- Random: Thay thế ngẫu nhiên
- FIFO (First In First Out): Thay thế *Block* nào nằm lâu nhất ở trong *Set đó*
- LFU (Least Frequently Used): Thay thế *Block* nào trong *Set* có số lần truy nhập ít nhất trong cùng một khoảng thời gian
- LRU (Least Recently Used): Thay thế Block ở trong *Set* tương ứng có thời gian lâu nhất không được tham chiếu tới.
- Tối ưu nhất: LRU

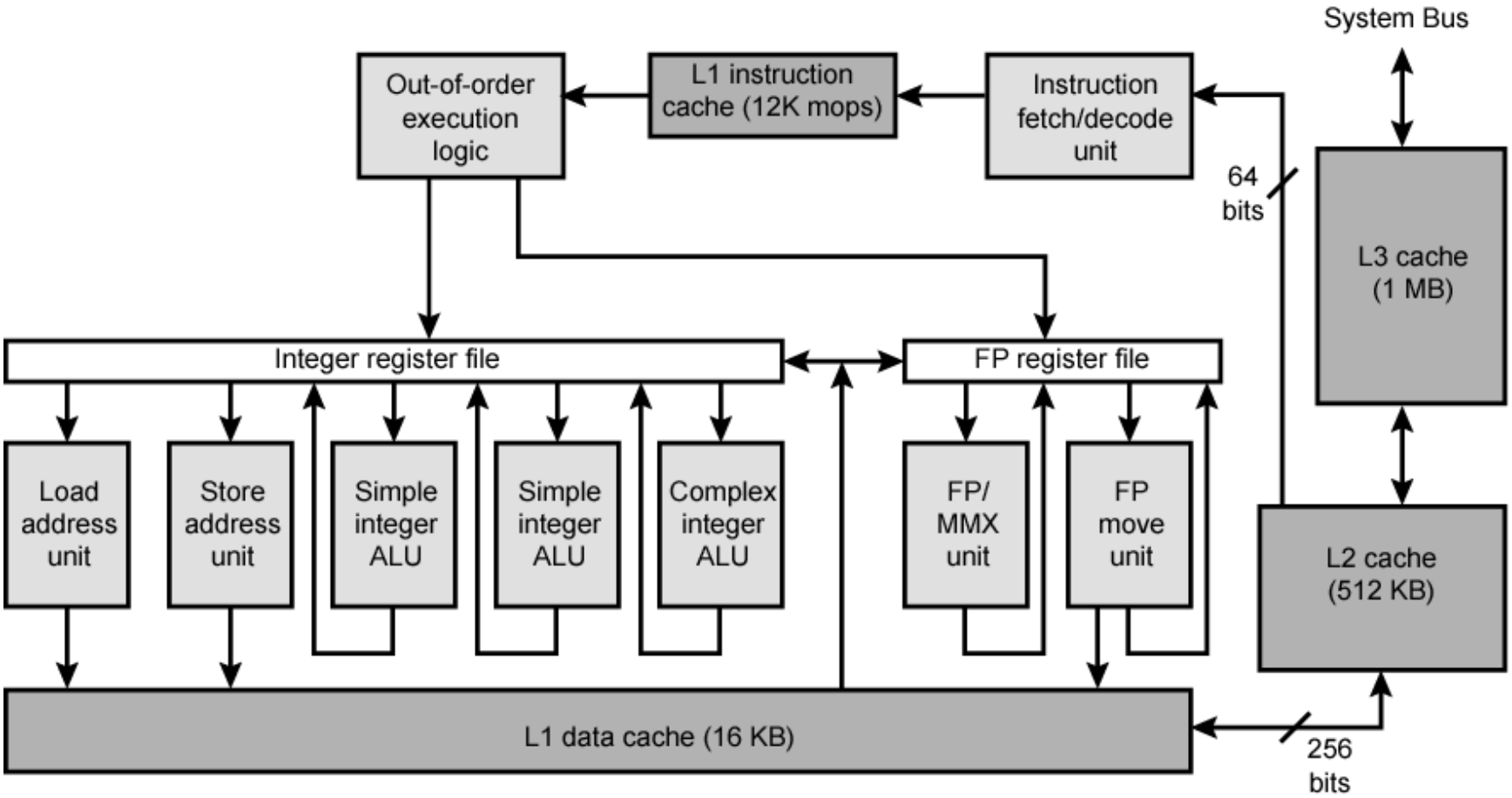
4. Phương pháp ghi dữ liệu khi cache hit

- Ghi xuyên qua (Write-through):
  - ghi cả cache và cả bộ nhớ chính
  - tốc độ chậm
- Ghi trả sau (Write-back):
  - chỉ ghi ra cache
  - tốc độ nhanh
  - khi Block trong cache bị thay thế cần phải ghi trả cả Block về bộ nhớ chính

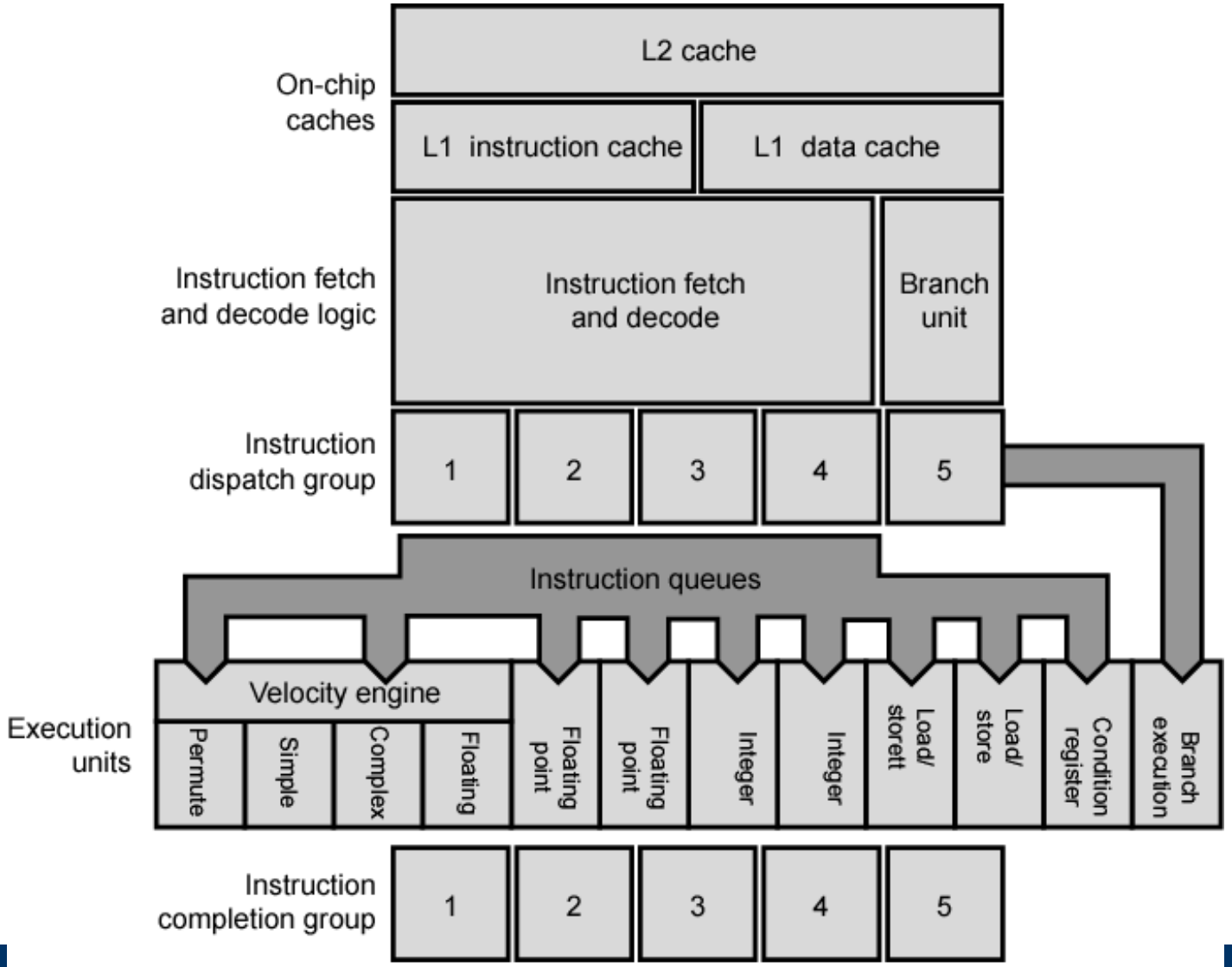
Một loại cache

- 80486: 8KB cache L1 trên chip
- Pentium: có hai cache L1 trên chip
  - Cache lệnh = 8KB
  - Cache dữ liệu = 8KB
- Pentium 4 (2000): hai mức cache L1 và L2 trên chip
  - Cache L1:
    - mỗi cache 8KB
    - Kích thước Line = 64 byte
    - ánh xạ liên kết tập hợp 4 đường
  - Cache L2
    - 256KB
    - Kích thước Line = 128 byte
    - ánh xạ liên kết tập hợp 8 đường

Sơ đồ Pentium 4



PowerPC G5 (dùng cho Power Mac)



Bài tập

- Các yếu tố ảnh hưởng tới hiệu năng bộ nhớ cache?
- Cách thức để giảm xác suất cache miss?

Câu 9. Bộ nhớ cache là gì ? Nêu vai trò của cache. Giải thích hai nguyên lý hoạt động của cache.

Câu 10. So sánh 3 phương pháp ánh xạ cache: ánh xạ trực tiếp, ánh xạ kết hợp đầy đủ và ánh xạ tập kết hợp? Phương pháp ánh xạ nào trong các phương pháp trên được sử dụng nhiều nhất trong thực tế? Tại sao?

Câu 11: Thành phần nào trong cấu trúc phân cấp bộ nhớ giúp làm tăng hiệu năng hệ thống và làm giảm giá thành sản xuất của máy tính? Tại sao?

Câu 12: Trình bày phương pháp ánh xạ trực tiếp trong các phương pháp ánh xạ bộ nhớ cache.

Câu 13: Trình bày phương pháp ánh xạ liên kết trong các phương pháp ánh xạ bộ nhớ cache.

Câu 14: Trình bày phương pháp ánh xạ liên kết tập hợp trong các phương pháp ánh xạ bộ nhớ cache.

Câu 15. Nêu các phương pháp đọc ghi và các chính sách thay thế dòng cache. Tại sao thay thế dòng cache sử dụng phương pháp LRU có khả năng cho hệ số đoán trúng (hit) cao nhất?

Câu 16: Trình bày các phương thức truy cập bộ nhớ. So sánh ưu nhược điểm các phương thức truy cập bộ nhớ đó.

Với ánh xạ trực tiếp

Bộ nhớ chính =8GB=2^33byte ➔ N=33 bit

Cache =512kb = 2^19 byte

Line =32byte = 2^5 byte ➔ W =5 bit

Số line trong cache = 2^19 / 2^5= 2^14 line

L=14 bit

T=33-(14+5) =14 bit

Tag	Line	Word
T=14	L=14	W=5

Với ánh xạ liên kết toàn phần

Bộ nhớ chính =8GB=2^33byte ➔ N=33 bit

Line =32byte = 2^5 byte ➔ W =5 bit

T = 33-5= 28 bit

Tag	Word
T=27	W=5

Với ánh xạ liên kết tập hợp 8 đường

Bộ nhớ chính =8GB=2^33byte ➔ N=33 bit

Cache =512kb = 2^19 byte

Line =32byte = 2^5 byte ➔ W =5 bit

Số line trong cache = 2^19 / 2^5= 2^14 line

Một set có 8 line = 2^3 line

Số Set trong cache = 2^14/2^3=2^11 Set ➔ S=11 bit

Số bit của trường Tag sẽ là T=33-(11+5)= 17 bit

Tag	Set	Word
T=17	S=11	W=5