BỘ NHỚ TRONG (INTERNAL STORAGE)

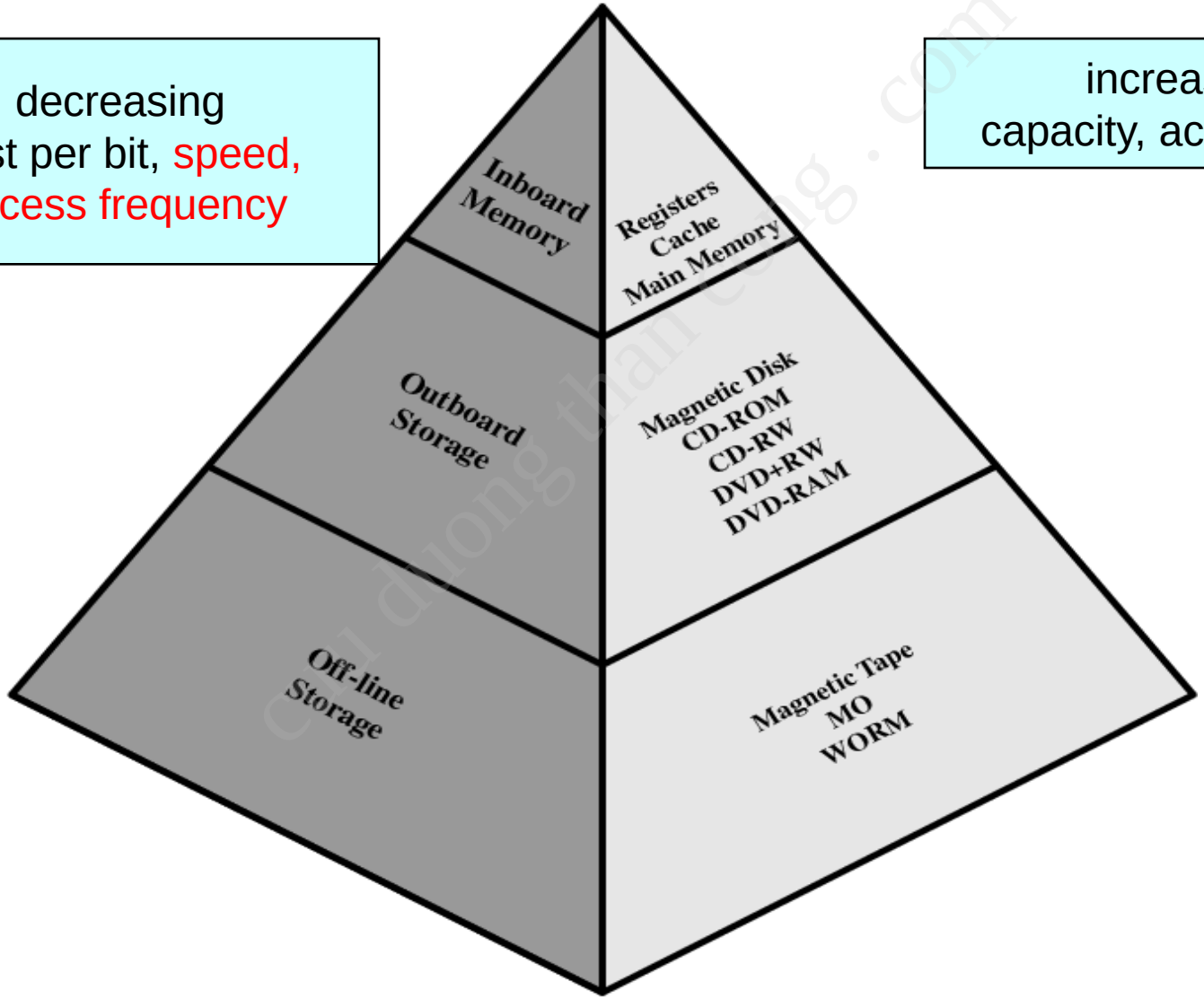
ĐẶC TRƯNG CỦA HỆ THỐNG NHỚ

- Vị trí: trong hay ngoài, trong CPU
- Dung lượng: kích thước từ nhớ, số lượng từ nhớ
- Đơn vị truyền tải: từ hay khối
- Phương pháp truy xuất: tuần tự, trực tiếp, ngẫu nhiên, liên kết
- Hiệu suất: thời gian truy xuất, tốc độ truyền, chu kỳ
- Dạng vật lý: bán dẫn hay băng từ
- Đặc tính vật lý: thay đổi/không thay đổi, xóa được/không xóa được
- Tổ chức bộ nhớ: sắp xếp vật lý các bit để hình thành một từ

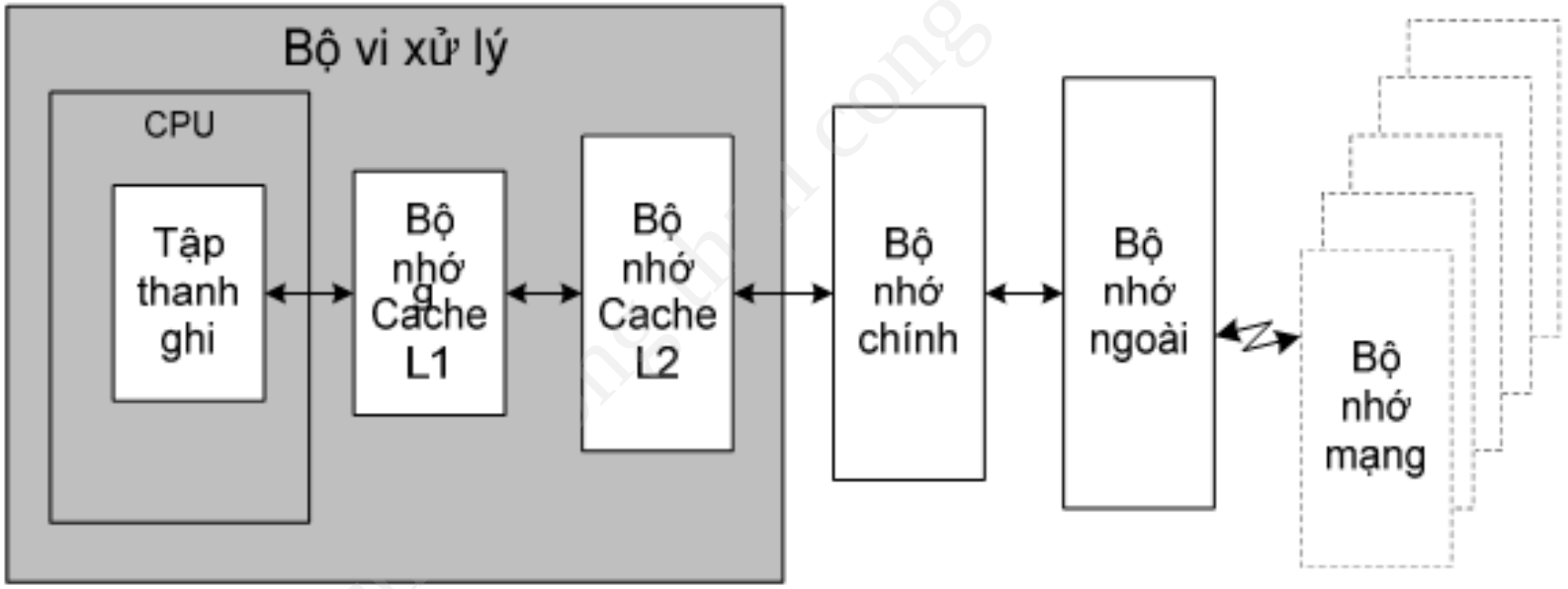
PHÂN CẤP HỆ THỐNG NHỚ

decreasing
cost per bit, speed,
access frequency

increasing
capacity, access time



PHÂN CẤP HỆ THỐNG NHỚ



BỘ NHỚ CHÍNH
ĐẶC TRƯNG

- Chứa các chương trình đang được 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 tùy thuộc vào hệ điều hành

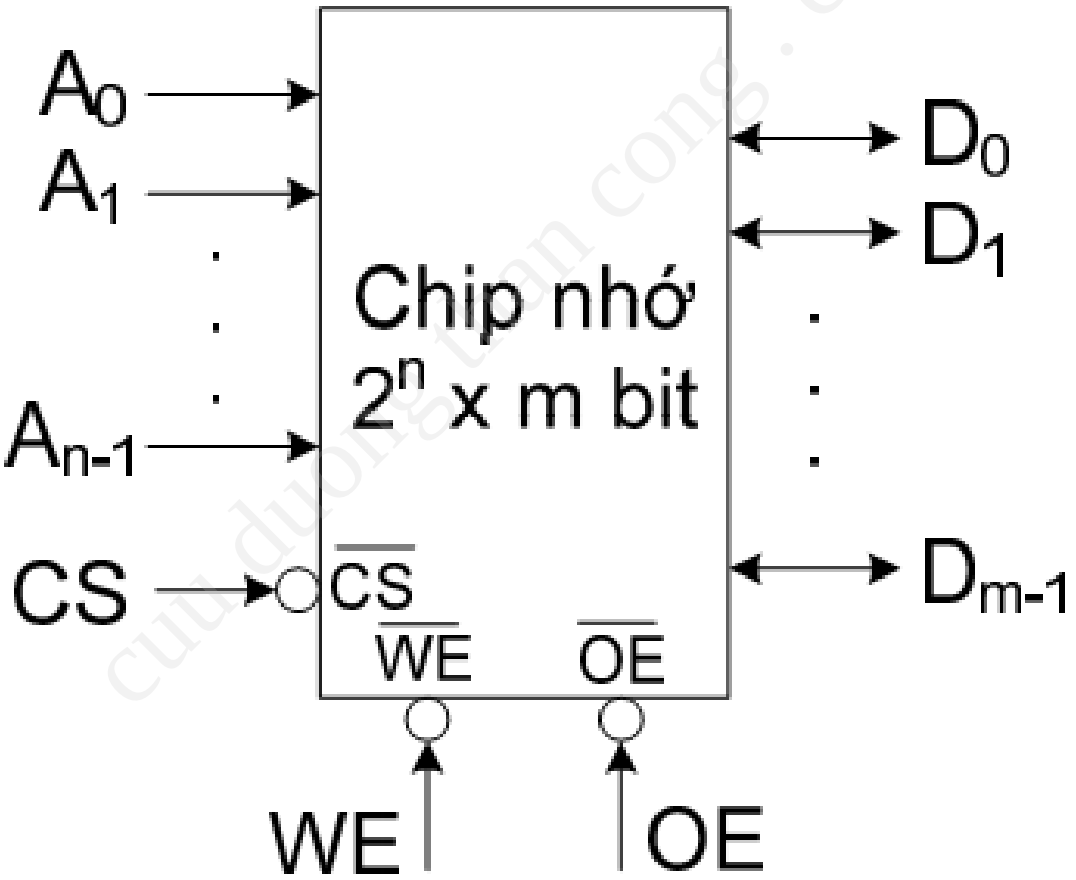
BỘ NHỚ CHÍNH
CÁC LOẠI BỘ NHỚ BÁN DẪN

- RAM (Random Access Memory)
  - RAM động: làm từ tụ điện, cần làm tươi, mật độ cao
  - RAM tĩnh: làm bằng các flip-flop, nhanh
- ROM (Read Only Memory)
  - PROM (Programmable ROM)
  - EPROM (Erasable PROM)
  - EEPROM (Electrically EPROM)
  - Flash Memory: lập trình lại rất nhanh, mật độ cao, xóa bằng điện và chỉ cần vài giây

BỘ NHỚ CHÍNH
TỔ CHỨC Ô NHỚ

- Ô nhớ là phần tử cơ bản có các thuộc tính:
  - Hai trạng thái: 0 và 1
  - Có thể cài đặt trạng thái, hoạt động ghi
  - Có thể đọc trạng thái, hoạt động đọc

BỘ NHỚ CHÍNH
TỔ CHỨC CHIP NHỚ



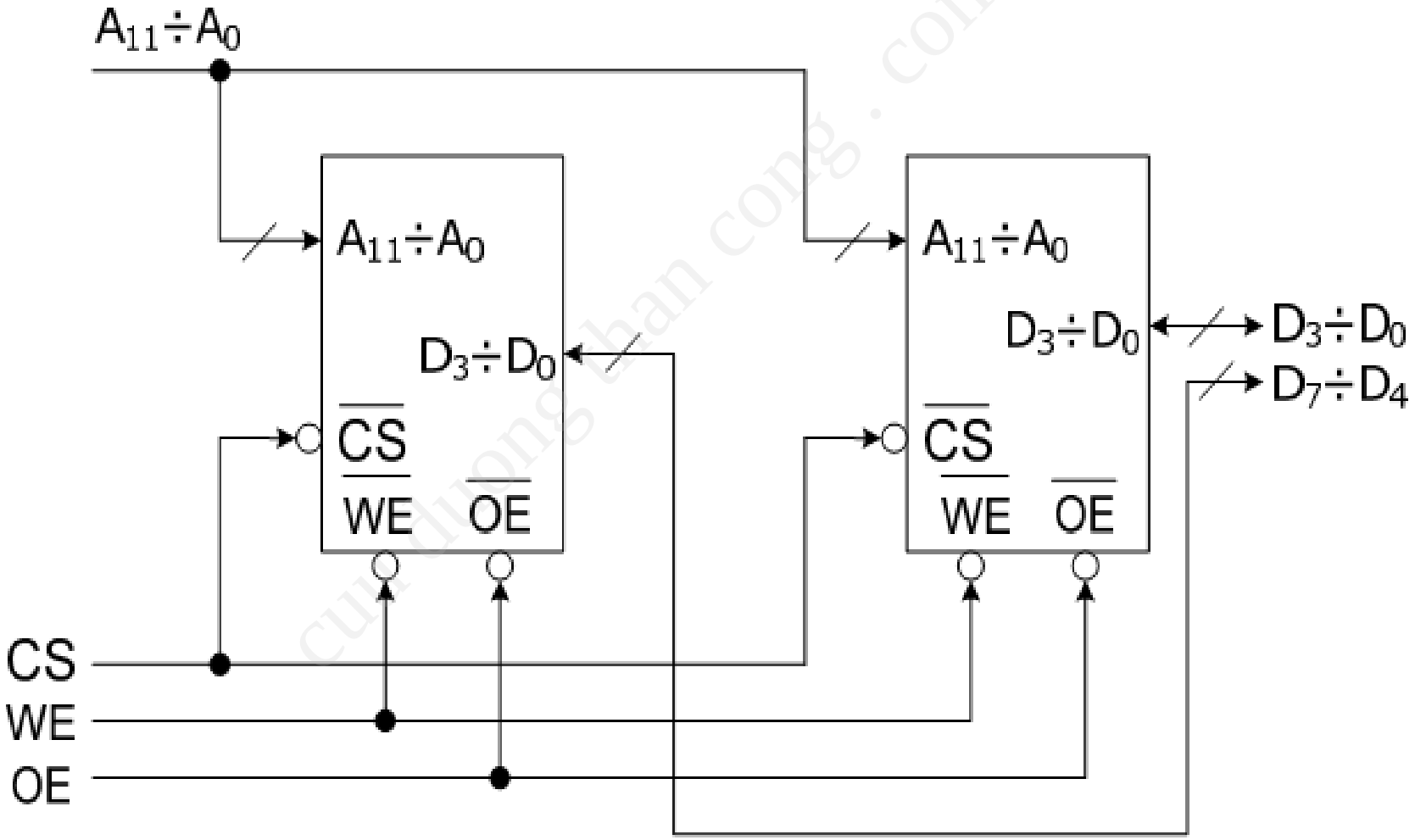
CÁC TÍN HIỆU CỦA CHIP NHỚ

- Dung lượng chip nhớ = 2^n x m bit
- Cần thiết kế để tăng dung lượng:
  - Thiết kế tăng độ dài từ nhớ
  - Thiết kế tăng số lượng từ nhớ
  - Thiết kế kết hợp

TĂNG ĐỘ DÀI TỪ NHỚ

- VD: cho chip nhớ SRAM 4K x 4 bit, thiết kế module nhớ 4K x 8 bit
- Dung lượng chip nhớ = 2^{12} x 4 bit
- Chip nhớ có: 12 chân địa chỉ, 4 chân dữ liệu
- Module nhớ cần có 12 chân địa chỉ, 8 chân dữ liệu

TĂNG ĐỘ DÀI TỪ NHỚ



TĂNG ĐỘ DÀI TỪ NHỚ

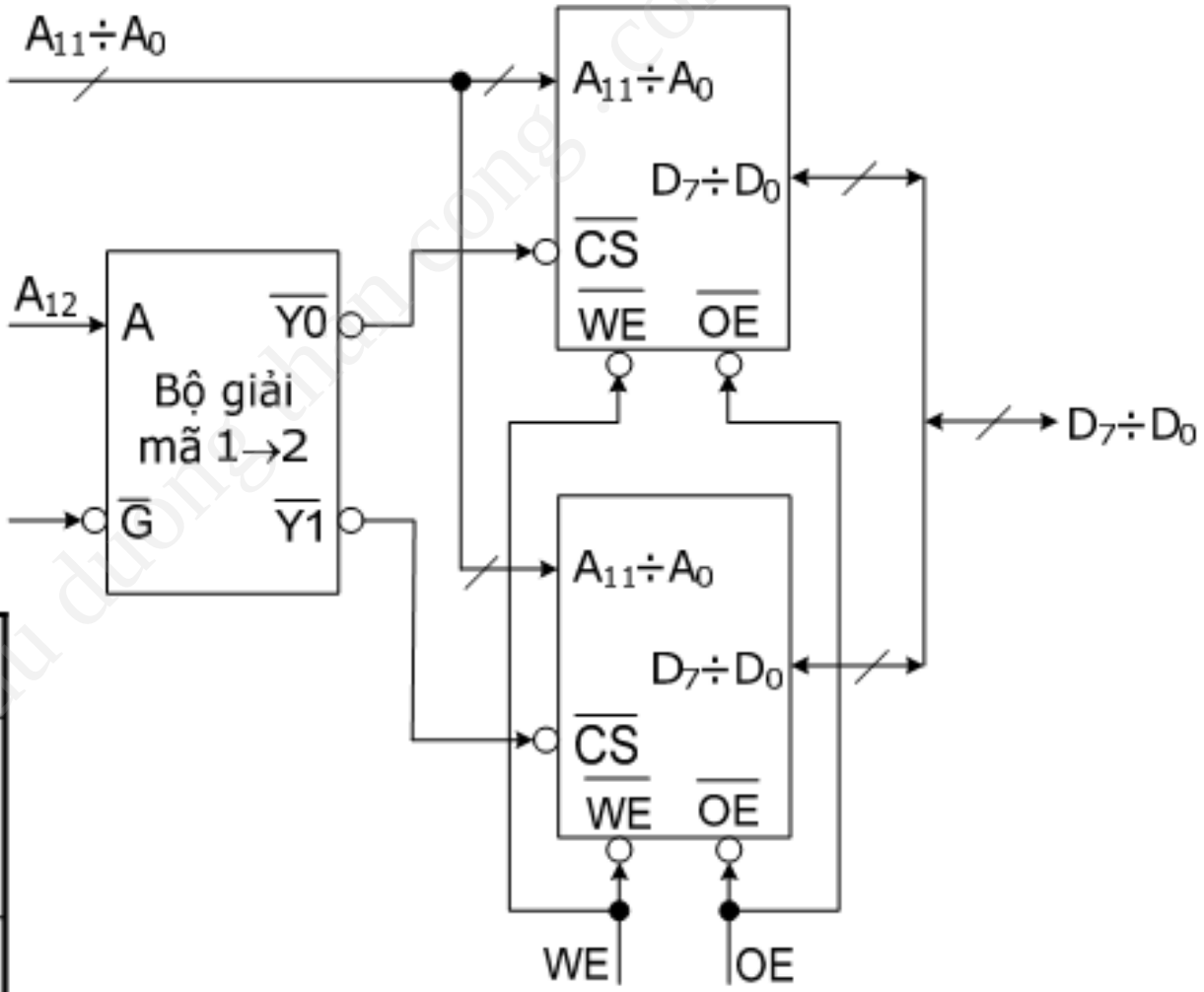
- Cho chip nhớ 2^n x m bit
- Thiết kế module nhớ 2^n x (k.m) bit
- Dùng k chip nhớ

TĂNG SỐ LƯỢNG TỪ NHỚ

- VD: cho chip nhớ SRAM 4K x 8 bit, thiết kế module nhớ 8K x 8 bit
- Dung lượng chip nhớ: 212 x 8 bit
- Chip nhớ có: 12 chân địa chỉ, 8 chân dữ liệu
- Dung lượng module nhớ 213 x 8 bit

→Cần có 13 chân địa chỉ, 8 chân dữ liệu

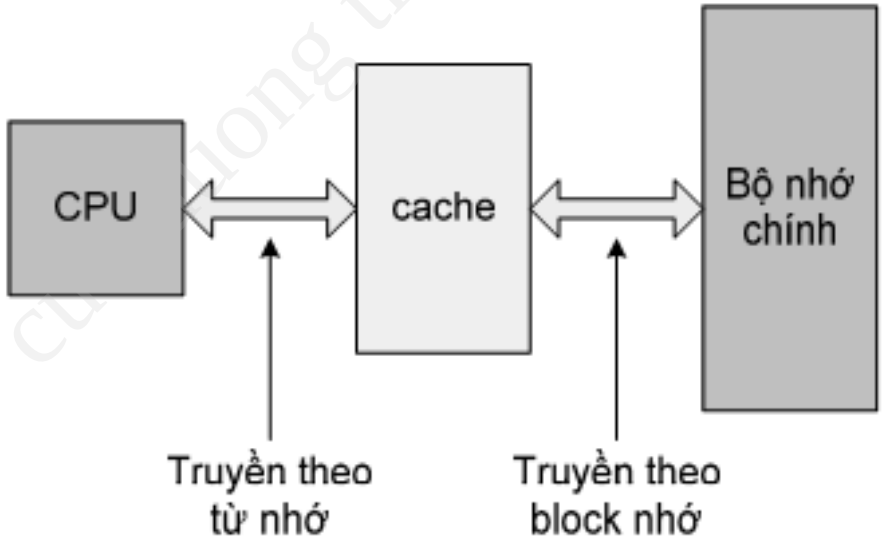
TĂNG SỐ LƯỢNG TỪ NHỚ



$\overline{G}$	A	$\overline{Y0}$	$\overline{Y1}$
0	0	0	1
0	1	1	0
1	x	1	1

BỘ NHỚ CACHE

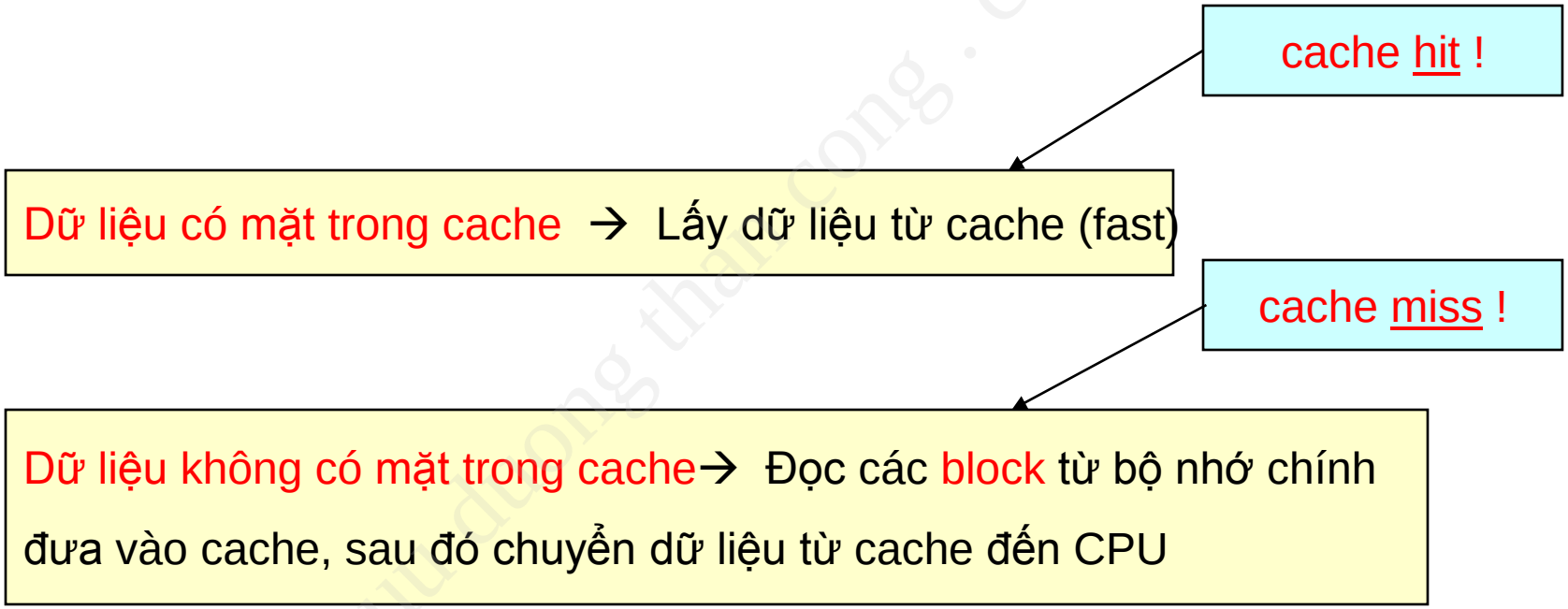
- Cache có tốc độ nhanh hơn bộ nhớ chính
- Cache được đặt giữa CPU và bộ nhớ chính nhằm tăng tốc độ truy nhập bộ nhớ của CPU
- Cache có thể được đặt trên chip của CPU



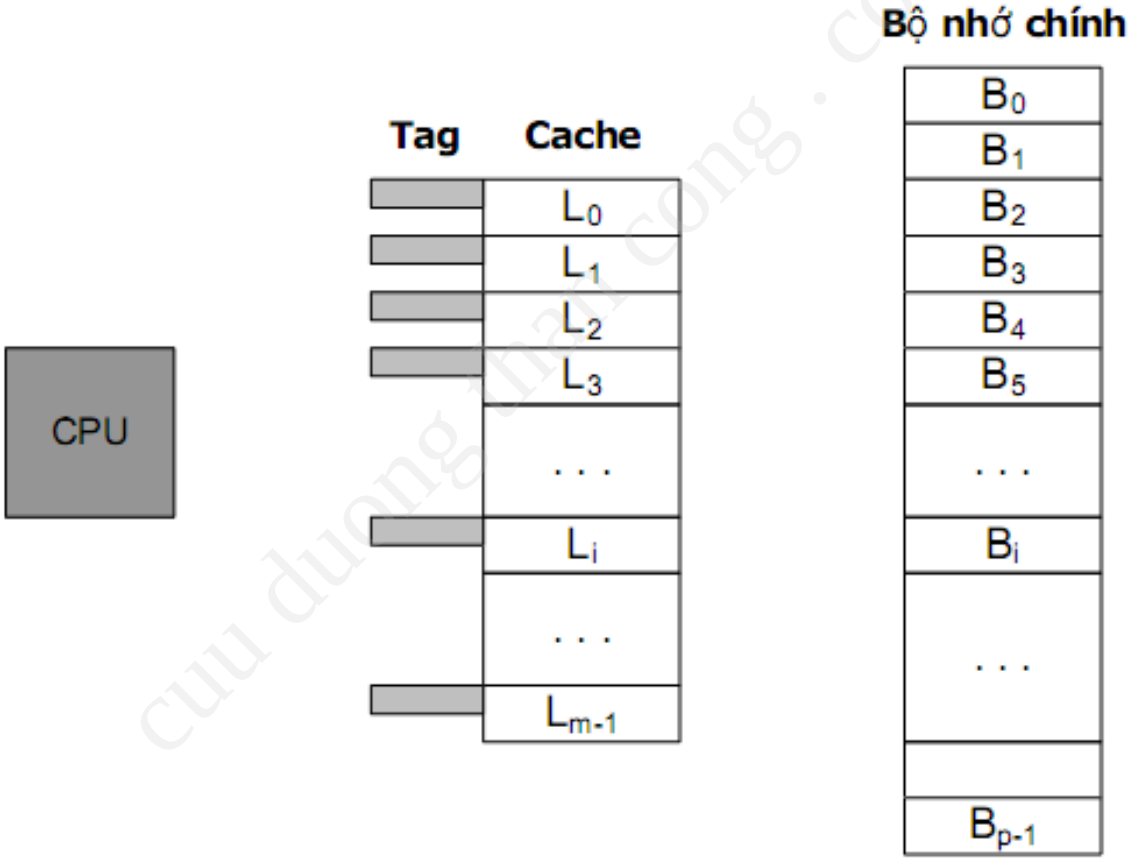
VÍ DỤ THAO TÁC CỦA CACHE

- 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

HOẠT ĐỘNG ĐỌC CACHE



CACHE VÀ BỘ NHỚ CHÍNH



CACHE VÀ BỘ NHỚ CHÍNH

- 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 2 khả năng xảy ra: cache hit hoặc cache miss
- Vì số line của cache ít hơn số block của bộ nhớ chính, cần có một **thuật giải ánh xạ thông tin trong bộ nhớ chính vào cache**.

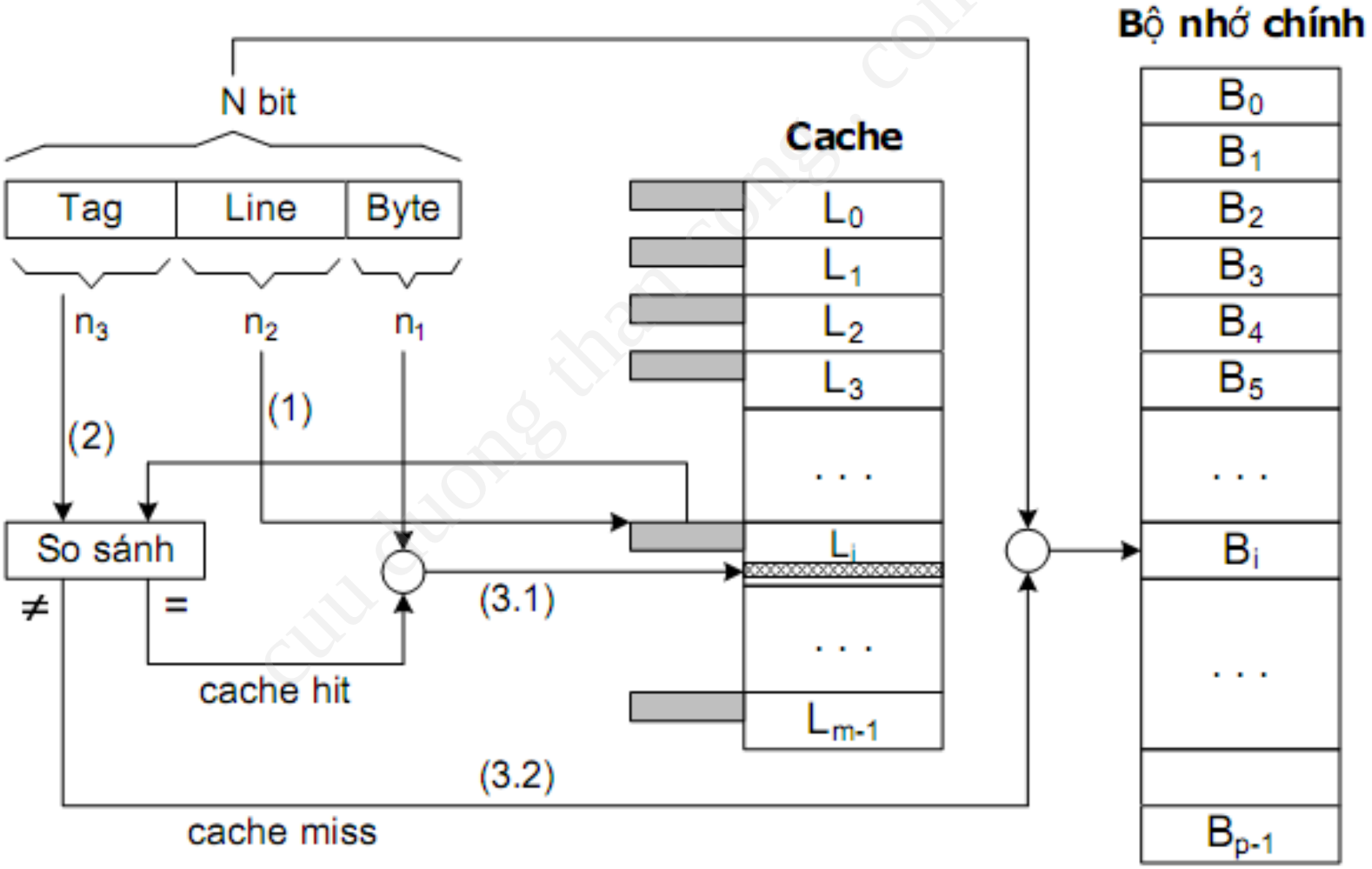
CÁC PHƯƠNG PHÁP ÁNH XẠ ĐỊA CHỈ

- Ánh xạ trực tiếp
- Ánh xạ liên kết toàn phần
- Ánh xạ liên kết tập hợp

ÁNH XẠ TRỰC TIẾP

- Mỗi block của bộ nhớ chính chỉ có thể được nạp vào 1 line duy nhất của cache.
- Quy ước nạp:
  - B0 → L0
  - B1 → L1
  - …
  - Bm → L0
  - Bm+1 → L1
  - …

ÁNH XẠ TRỰC TIẾP



ÁNH XẠ TRỰC TIẾP

- Trường Byte (có n1 bit) để xác định byte nhớ trong Line (Block)

$$2^{n1} = \text{kích thước 1 Line}$$

- Trường Line (có n2 bit) để xác định Line trong Cache

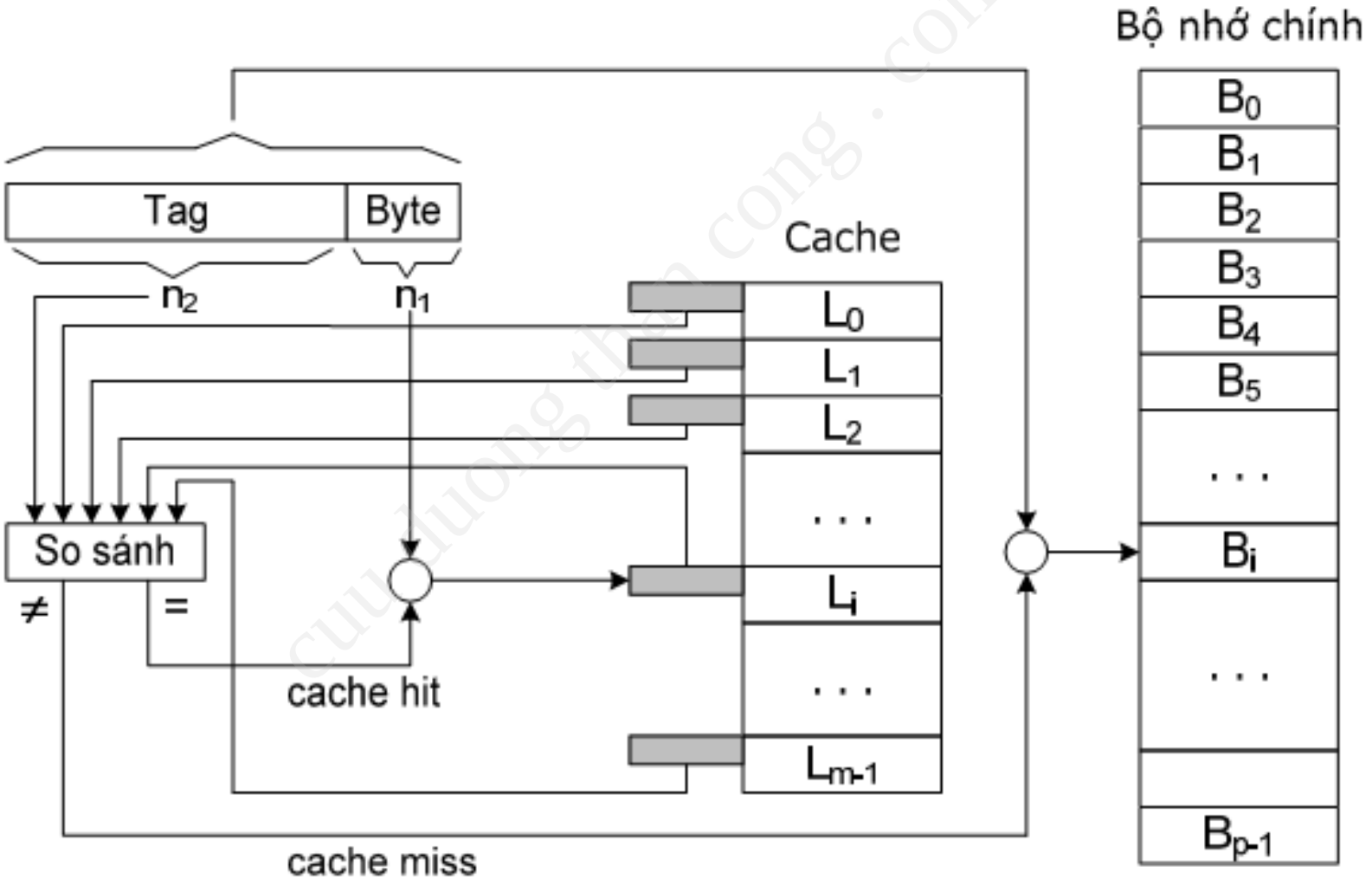
$$2^{n2} = \text{số Line trong Cache}$$

- Dung lượng Cache = $2^{n1} * 2^{n2} = 2^{n1+n2}$
- Trường Tag (có n3 bit): số bit còn lại n3 = N - (n1 + n2) > 0 vì $2^{N} >> 2^{n1+n2}$

ÁNH XẠ LIÊN KẾT TOÀN PHẦN

- Mỗi block có thể được nạp vào bất kỳ line nào của cache.
- Địa chỉ bộ nhớ do CPU phát ra được chia thành 2 phần: tag và byte.
- Để kiểm tra xem một block có trong cache hay không, phải đồng thời kiểm tra tất cả tag của các line trong cache.
- Cần các mạch phức tạp để kiểm tra.

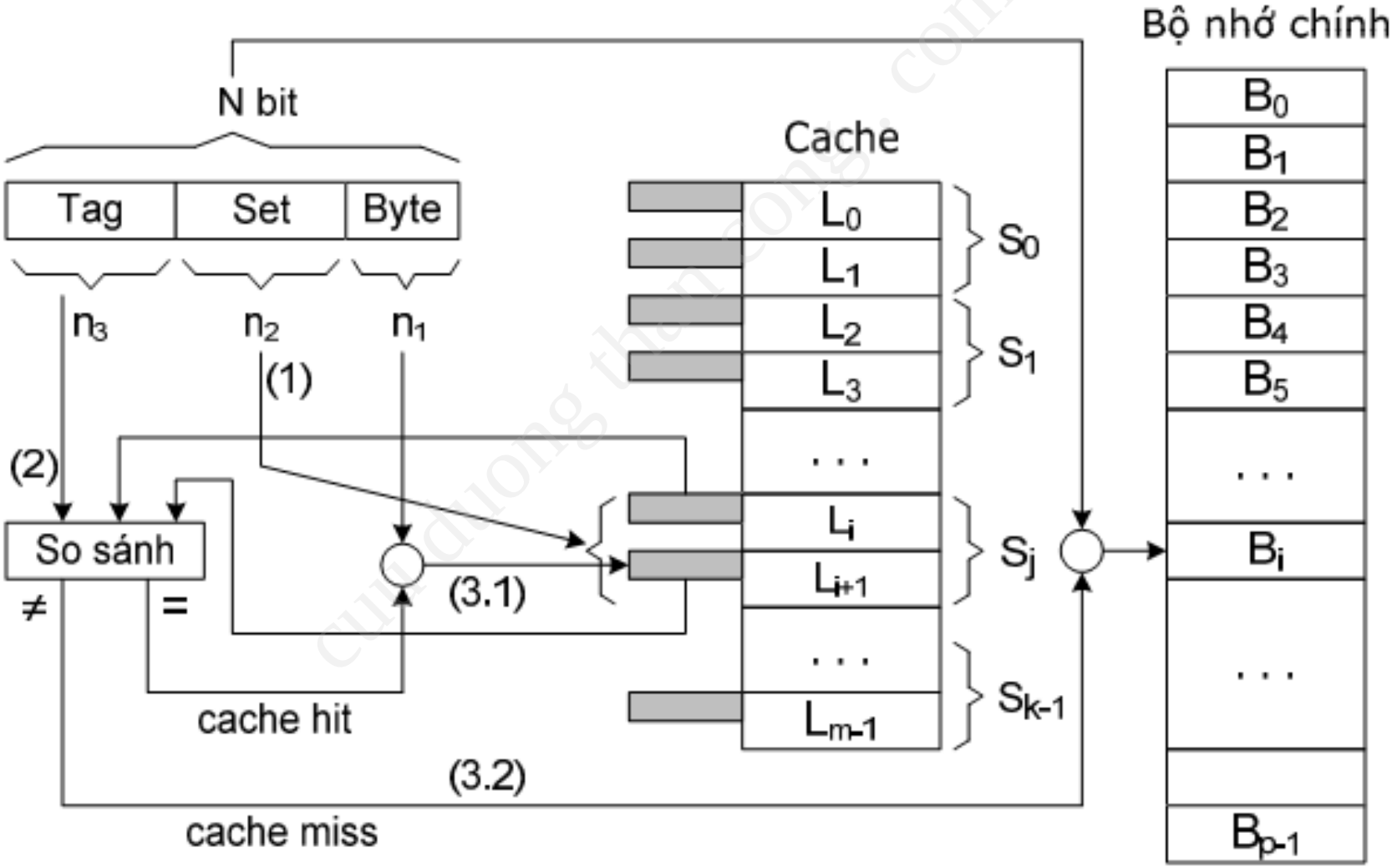
ÁNH XẠ LIÊN KẾT TOÀN PHẦN



ÁNH XẠ LIÊN KẾT TẬP HỢP

- Là phương pháp dung hòa của 2 phương pháp trên
- Chia cache thành các tập: S0, S1, S2 ...
- Mỗi Set có một số Line (2, 4, 8, 16 Line)
- Ví dụ mỗi Set có 2 line: 2-way Set Associative Mapping
- Mỗi block được nạp vào 1 line nào đó trong Set nhất định:
- Địa chỉ do CPU phát ra có 3 trường: Tag, Set, Byte

ÁNH XẠ LIÊN KẾT TẬP HỢP



VÍ DỤ

- Hệ thống có:
  - Bộ nhớ chính: 256 MB
  - Cache = 128 KB
  - Line = 16 byte
- Xác định số bit của trường địa chỉ khi
  - Ánh xạ trực tiếp
  - Ánh xạ liên kết tập hợp 4 Line/Set

VÍ DỤ

1)$2^N = 256.2^{20} = 2^{28}$ -> N = 28 bit

- Tính cho trường Byte:
- Kích thước line = 16 = 2^4 Byte -> n1 = 4 bit
- Tính cho trường Line:
- Số line trong Cache: $128.2^{10}/16 = 2^{13}$
- -> n2 = 13 bit
- Tính cho trường Tag:
- n3 = N - (n1 + n2) = 28 - (4 + 13) = 11 bit

2)Trường Byte: n1 = 4 bit

Trường Set:

Số Set = Số line/4 = $2^{13}/4 = 2^{11}$ -> n2 = 11 bit

Trường Tag: n3 = N - (n1 + n2) = 28 - (4 + 11) = 13 bit

GIẢI THUẬT
THAY THẾ BLOCK TRONG CACHE

- Khi CPU truy nhập một thông tin mà không có trong cache (cache miss) thì nạp block chứa thông tin đó vào trong cache để thay thế block cũ trong cache.
- Ánh xạ trực tiếp → chỉ có 1 cách nạp → không cần thuật giải để nạp.
- 2 phương pháp ánh xạ liên kết → cần có thuật giải để lựa chọn thay thế.

GIẢI THUẬT
THAY THẾ BLOCK TRONG CACHE

- Random: thay block một cách ngẫu nhiên.
- FIFO (First In, First Out): thay thế block đã tồn tại lâu nhất trong toàn cache đối với ánh xạ liên kết toàn phần, trong set đối với ánh xạ liên kết tập hợp.
- LFU (Least Frequently Used): thay block có số lần truy nhập ít nhất.
- LRU (Least Recently Used): thay block có khoảng thời gian dài nhất không được truy nhập.

GHI DỮ LIỆU KHI CACHE HIT

- Ghi xuyên qua (Write throught):
  - Ghi cả cache và bộ nhớ chính
  - Tốc độ chậm
- Ghi trả sau (Write back):
  - Chỉ ghi ra cache
  - Tốc độ nhanh