

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG & TIN HỌC



Bài giảng

KIẾN TRÚC MÁY TÍNH

Giảng viên : Phạm Huyền Linh
Bộ môn : Toán Tin



CHƯƠNG 7

BỘ NHỚ MÁY TÍNH

Chương 7



1. Tổng quan hệ thống bộ nhớ
2. Bộ nhớ bán dẫn
3. Bộ nhớ chính
4. Bộ nhớ đệm (cache)
5. Bộ nhớ ngoài
6. Bộ nhớ ảo

1. Tổng quan hệ thống bộ nhớ



- Các đặc trưng của bộ nhớ
- Phân cấp bộ nhớ
- Phát hiện và chỉnh lỗi trong bộ nhớ

Các đặc trưng



- Chức năng

- Lưu chương trình và dữ liệu.

- Các thao tác cơ bản

- Thao tác ghi (Write)
- Thao tác đọc (Read)

- Các thành phần chính

- Bộ nhớ trong:
 - Bộ nhớ đệm (Cache)
 - Bộ nhớ chính (Main memory)
- Bộ nhớ ngoài:
 - HDD, SSD, RAID
 - Các thiết bị lưu trữ khác: CD, DVD, USB, Thẻ nhớ,...

Các đặc trưng



- Dung lượng nhớ: Tổng số byte của bộ nhớ
 - Đơn vị DL nhỏ nhất trong bộ nhớ là bit
 - 1byte = 8bit

Theo thập phân			Theo nhị phân		
Đơn vị	Viết tắt	Giá trị	Đơn vị	Viết tắt	Giá trị
kilobyte	KB	10^3	kibibyte	KiB	$2^{10} = 1024$
megabyte	MB	10^6	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	exbibyte	EiB	2^{60}

Các đặc trưng



- Phương pháp truy nhập
 - Truy nhập tuần tự (băng từ)
 - Truy nhập trực tiếp (các loại đĩa)
 - Truy nhập ngẫu nhiên (bộ nhớ bán dẫn)
 - Truy nhập liên kết (cache)
- Kiểu vật lý
 - Bộ nhớ bán dẫn
 - Bộ nhớ từ
 - Bộ nhớ quang

Các đặc trưng



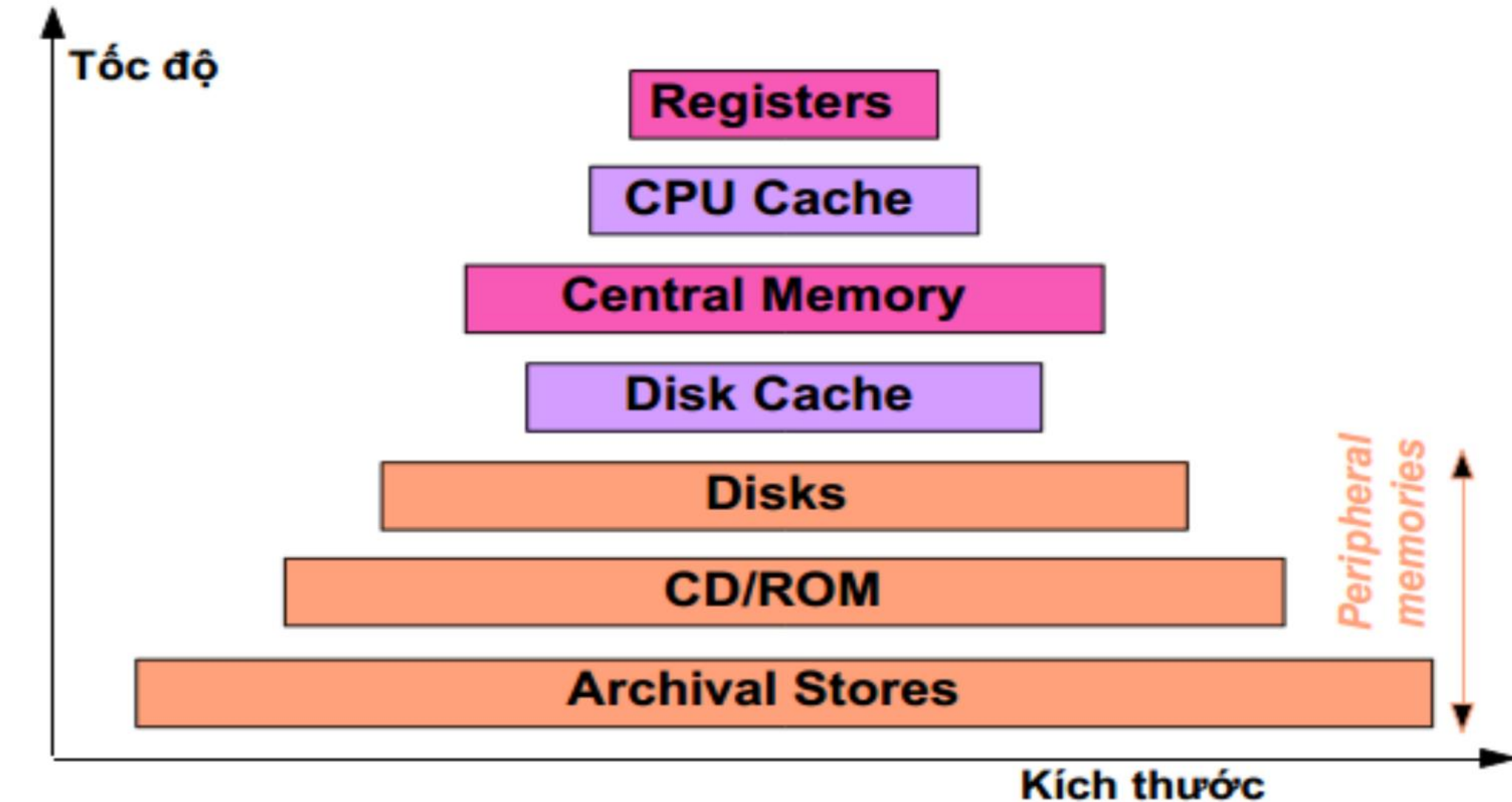
- Hiệu năng (performance)
 - Thời gian truy nhập: T/g từ lúc đưa ra địa chỉ cho đến khi đọc được ô nhớ đó
 - Chu kỳ nhớ: Thời gian giữa 2 lần liên tiếp thâm nhập bộ nhớ
- Các đặc tính vật lý
 - Khả biến / Không khả biến (volatile / nonvolatile)
 - Xoá được / không xoá được

1. Tổng quan hệ thống bộ nhớ

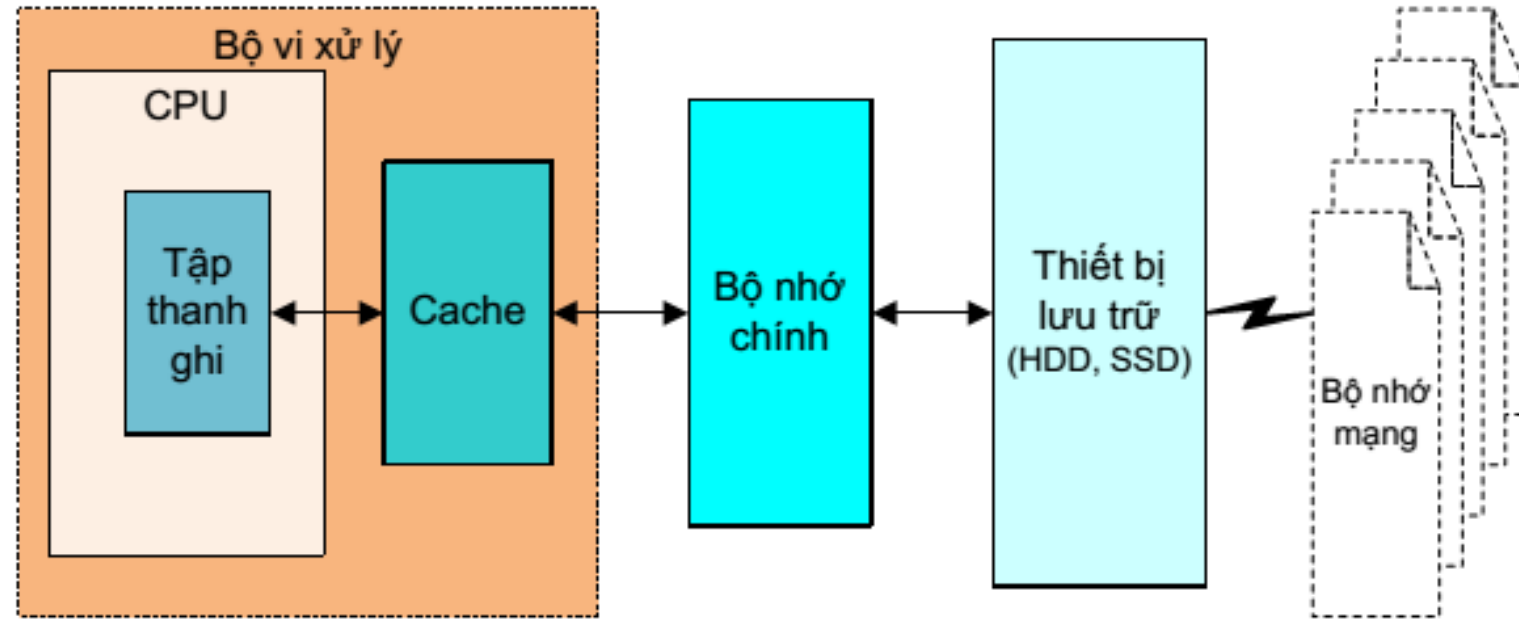


- Các đặc trưng của bộ nhớ
- Phân cấp bộ nhớ
- Phát hiện và chỉnh lỗi trong bộ nhớ

Phân cấp bộ nhớ



Phân cấp bộ nhớ



- **Từ trái sang phải:**
 - Dung lượng tăng dần
 - Tốc độ giảm dần
 - Giá thành/byte hay bit giảm dần

Ví dụ



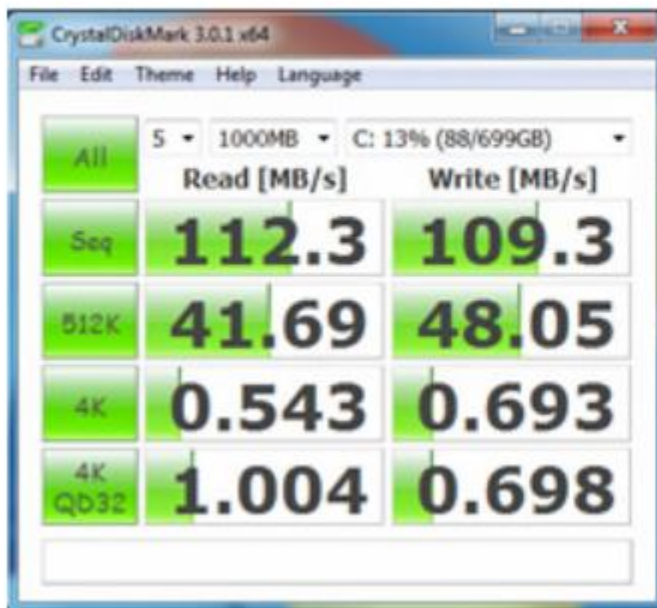
Công nghệ bộ nhớ	Thời gian truy nhập	Giá thành/GiB (2012)
SRAM	0,5 – 2,5 ns	\$500 – \$1000
DRAM	50 – 70 ns	\$10 – \$20
Flash memory	5.000 – 50.000 ns	\$0,75 – \$1
HDD	5 – 20 ms	\$0,05 – \$0,1

- Bộ nhớ lý tưởng
 - Thời gian truy nhập nhanh như SRAM
 - Dung lượng và giá thành rẻ như ổ đĩa cứng

Ví dụ 2: SS tốc độ RAM, SSD, HDD



Hard Drive



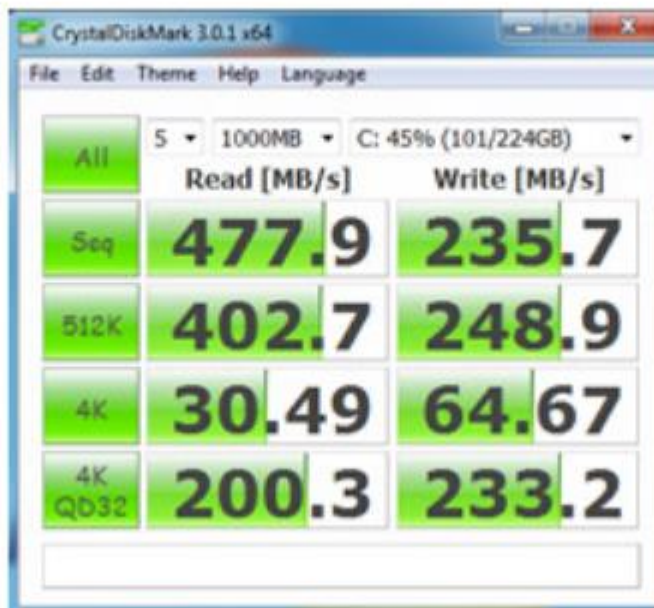
CrystalDiskMark 3.0.1 v64

File Edit Theme Help Language

5 1000MB C: 13% (88/699GB)

	Read [MB/s]	Write [MB/s]
All	112.3	109.3
Seq	112.3	109.3
512K	41.69	48.05
4K	0.543	0.693
4K Qb32	1.004	0.698

SSD



CrystalDiskMark 3.0.1 v64

File Edit Theme Help Language

5 1000MB C: 45% (101/224GB)

	Read [MB/s]	Write [MB/s]
All	477.9	235.7
Seq	477.9	235.7
512K	402.7	248.9
4K	30.49	64.67
4K Qb32	200.3	233.2

RAM Disk



CrystalDiskMark 3.0.1 v64

File Edit Theme Help Language

5 1000MB R: 1% (48/4089MB)

	Read [MB/s]	Write [MB/s]
All	5766	7760
Seq	5766	7760
512K	5649	7172
4K	657.0	554.8
4K Qb32	631.9	544.7

- Sự khác biệt tốc độ là do độ rộng bus
- Do HDD được điều khiển bởi hệ thống cơ khí, bộ nhớ RAM bằng điện tử

1. Tổng quan hệ thống bộ nhớ



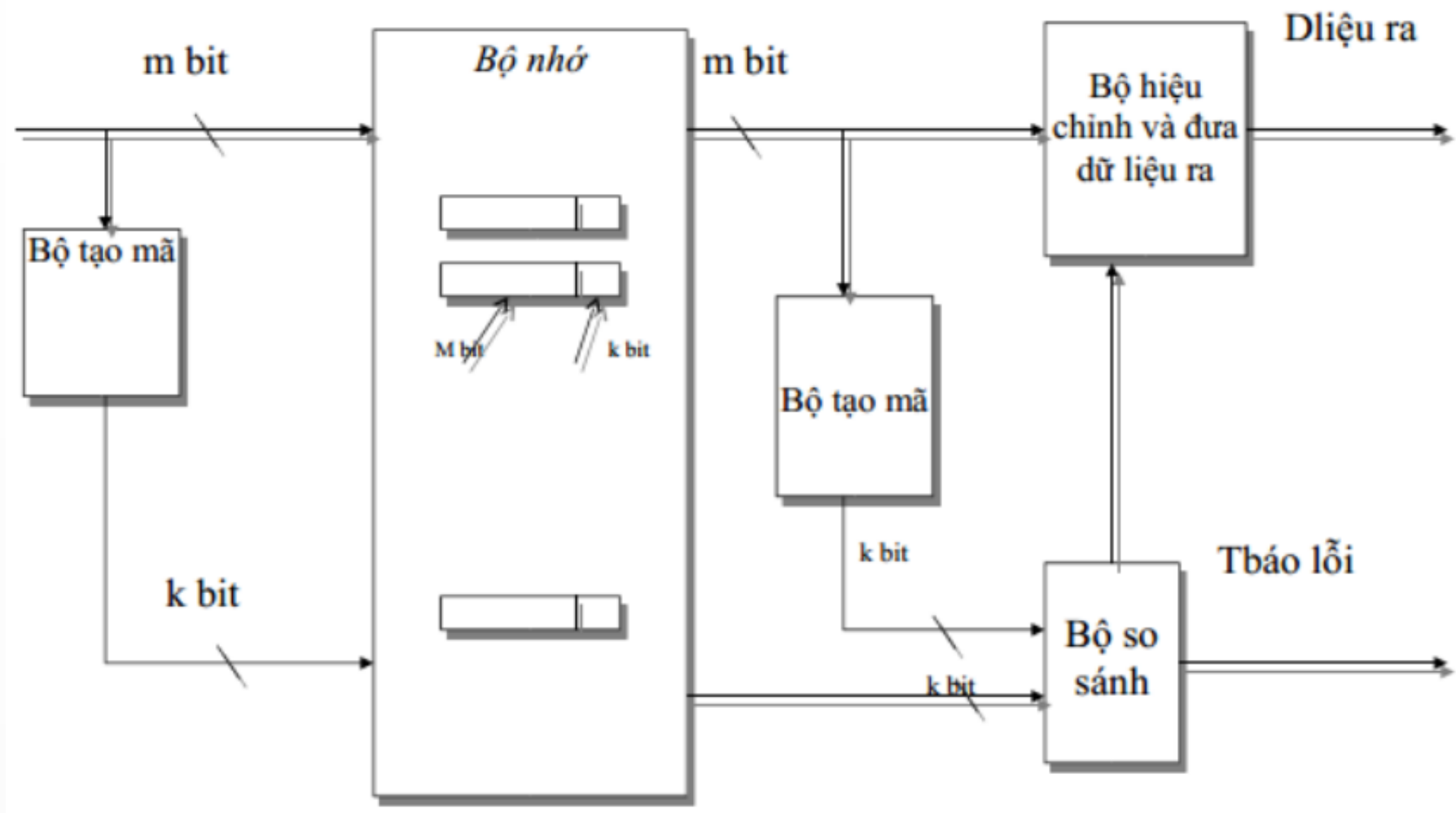
- Các đặc trưng của bộ nhớ
- Phân cấp bộ nhớ
- Phát hiện và chỉnh lỗi trong bộ nhớ

Phát hiện và chỉnh lỗi trong bộ nhớ



- **Nguyên tắc:** Tạo và lưu trữ thêm thông tin thừa
 - Từ dữ liệu cần ghi: m bit
 - Tạo và lưu trữ thêm: k bit
 - > Lưu trữ $m+k$ bit ứng với mỗi từ nhớ
- **Khi đọc:**
 - Không phát hiện thấy lỗi
 - Phát hiện lỗi và hiệu chỉnh được
 - Phát hiện lỗi và không hiệu chỉnh được -> Phát tín hiệu thông báo lỗi

Sơ đồ phát hiện và hiệu chỉnh lỗi



Chương 7



1. Tổng quan hệ thống bộ nhớ
2. Bộ nhớ bán dẫn
3. Bộ nhớ chính
4. Bộ nhớ đệm (cache)
5. Bộ nhớ ngoài
6. Bộ nhớ ảo

2. Bộ nhớ bán dẫn



- Phân loại
- Thiết kế modul nhớ bán dẫn

Phân loại



Kiểu bộ nhớ	Tiêu chuẩn	Khả năng xóa	Cơ chế ghi	Tính khả biến
Read Only Memory (ROM)	Bộ nhớ chỉ đọc	Không xóa được	Mặt nạ	Không khả biến
Programmable ROM (PROM)			Bảng điện	
Erasable PROM (EPROM)	bằng tia cực tím, cả chip			
Electrically Erasable PROM (EEPROM)	bằng điện, mức từng byte			
Flash memory	bằng điện, từng khối			
Random Access Memory (RAM)	Bộ nhớ đọc-ghi	bằng điện, mức từng byte	Bảng điện	Khả biến

ROM (Read Only Memory)

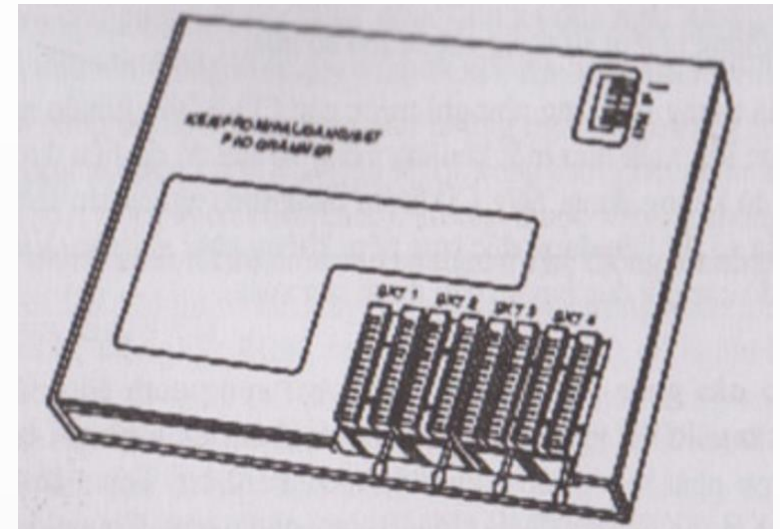


- Bộ nhớ không khả biến (nonvolatile)
- Lưu trữ các thông tin sau:
 - Thư viện các chương trình con
 - Các chương trình điều khiển hệ thống (BIOS)
 - Vi chương trình

Các kiểu ROM



- **ROM mặt nạ (Mask ROM):**
 - Bộ nhớ chỉ đọc
 - Thông tin được ghi khi sản xuất (thực chất là làm khuôn sẵn, nếu thay đổi 1bit phải làm lại khuôn)
 - Tốc độ truy nhập nhanh
- **PROM (Programmable ROM)**
 - Bộ nhớ chỉ đọc
 - Là trống khi sản xuất, có thể lập trình theo y/c sử dụng
 - Chỉ ghi được một lần
 - Cần thiết bị chuyên dụng để ghi (Máy lập trình, hay máy đốt)
 - Tốc độ truy nhập nhanh



Máy lập trình ROM

Các kiểu ROM

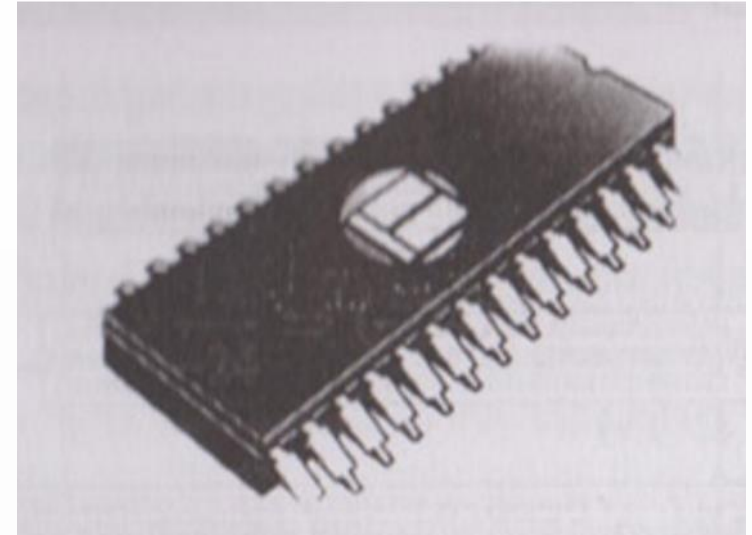


■ EPROM (Erasable PROM)

- Xóa được bằng tia tử ngoại (Phơi sang dưới các UV mạnh) , xóa toàn bộ
- Ghi lại được nhiều lần
- Đòi hỏi thiết bị chuyên dụng để ghi
- Tốc độ truy cập nhanh

■ EEPROM (Electrically Erasable PROM)

- Có thể xóa bằng điện theo từng byte
- Đọc nhanh, nhưng xóa và ghi chậm



EPROM có cửa sổ thạch anh để
xóa bằng tia tử ngoại

Bộ nhớ Flash



- Không mất DL khi mất điện (nonvolatile)
- Đọc, ghi bằng điện theo từ máy hay theo khối nhỏ.
- Xóa bằng điện, xóa theo khối lớn trước khi ghi
- Mỗi bit nhớ được lập từ 2 dạng cổng logic NAND, NOR
- Tốc độ đọc nhanh
- Tốc độ ghi và xóa chậm
- Thuận tiện để sử dụng do nhỏ gọn, dễ tháo rời (Card flash, USB,...)



RAM (Random Access Memory)

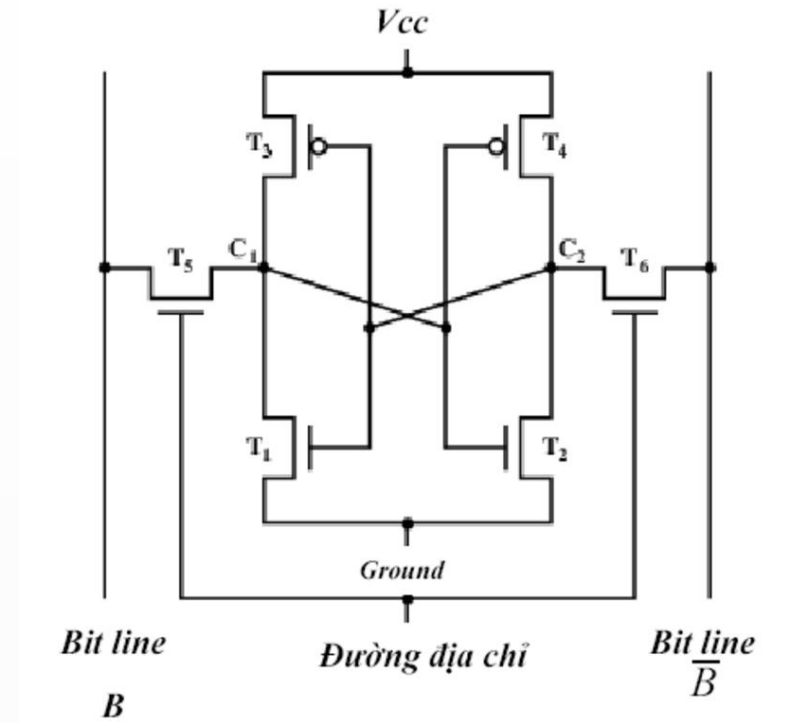


- Là bộ nhớ khả biến
- Thông thường địa chỉ được đánh theo byte, hệ thống lại có thể đọc hay ghi nhiều byte cùng 1 lúc (Phụ thuộc độ lớn bus địa chỉ)
- Thời gian truy nhập các ô nhớ là như nhau thông qua địa chỉ các ô nhớ
- Lưu trữ thông tin tạm thời, thông tin sẽ mất đi khi mất nguồn điện
- Máy tính dùng Ram lưu trữ chương trình và DL trong qt thực thi
- Có hai loại:
 - SRAM (Static RAM)
 - DRAM (Dynamic RAM)

SRAM (Static)

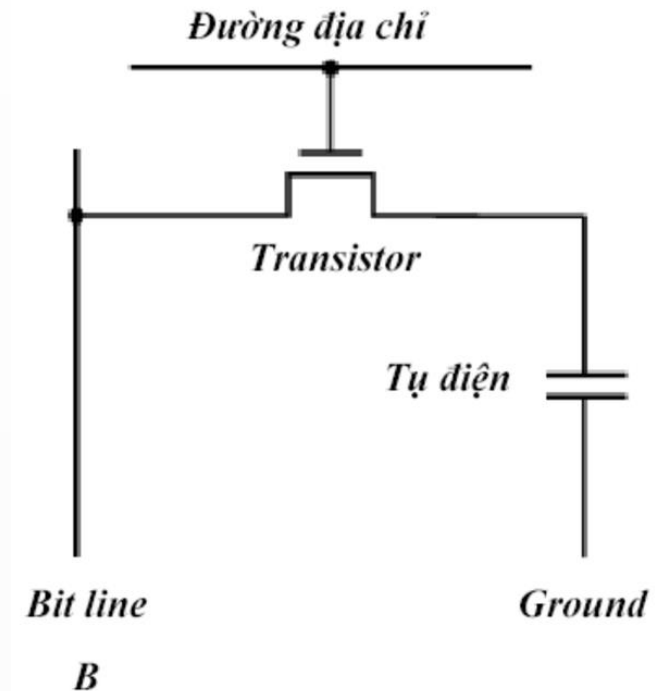


- Các bit được lưu trữ bằng các transistors
- Thông tin ổn định, việc đọc không hủy nội dung ô nhớ
- Cấu trúc phức tạp
- Dung lượng chip nhỏ
- Tốc độ nhanh
- Dùng làm bộ nhớ cache
- Từ “tĩnh” nghĩa là bộ nhớ vẫn lưu trữ DL khi có điện



DRAM (Dynamic)

- Các bit được lưu trữ bằng tụ điện.
- Điện áp tích tụ giảm dần theo tg => Phải có mạch làm tươi nên gọi là động (trong lúc này CPU không thể R/W bộ nhớ)
- Cấu trúc đơn giản
- Dung lượng lớn, hàng triệu tụ điện/1 vùng diện tích mạch nhỏ
- Tốc độ chậm hơn
- Dùng làm bộ nhớ chính
- Độ trễ lớn nên tốc độ chậm hơn CPU rất nhiều:
 - RAM còn có độ trễ: VD CL=5, Ram trễ mất 5 clocks để truyền DL ngược lại (CL: Clock Latency)



So sánh SRAM và DRAM



Tính năng	RAM động (DRAM)	RAM tĩnh (SRAM)
Mạch trữ điện	Tụ điện	Flip-flop
Tốc độ truyền tải	Thấp hơn CPU	Bằng với CPU
Độ trễ	Cao	Thấp
Mật độ	Cao	Thấp
Tiêu tốn năng lượng	Thấp	Cao
Giá thành	Rẻ	Đắt

2. Bộ nhớ bán dẫn

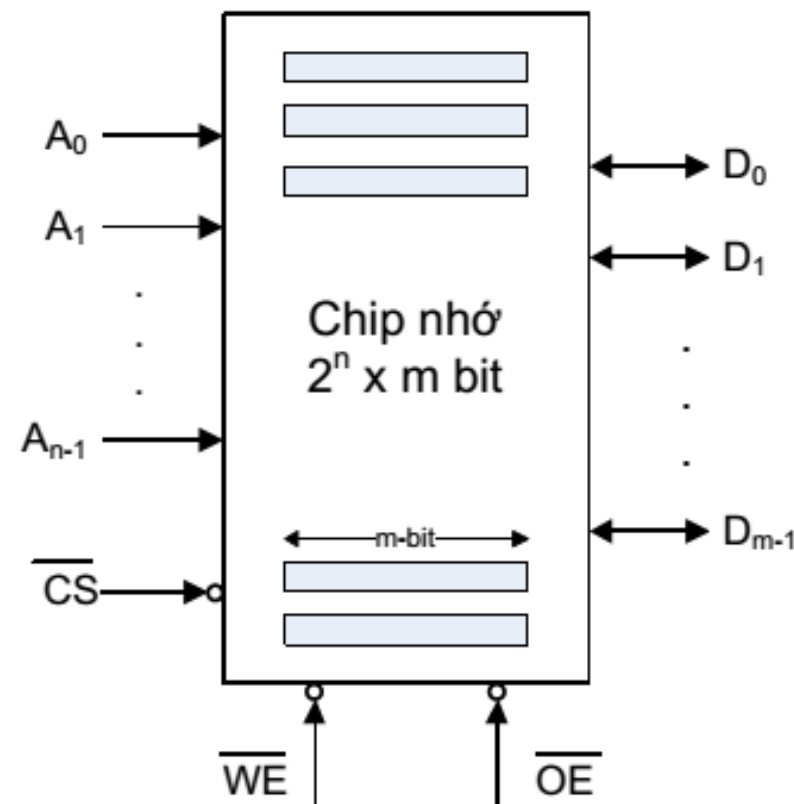


- Phân loại
- Thiết kế modul nhớ bán dẫn

Tổ chức của chip nhớ



- Các đường địa chỉ: $A_{n-1} \div A_0 \Rightarrow$ có 2^n từ nhớ
- Các đường dữ liệu: $D_{m-1} \div D_0 \Rightarrow$ độ dài từ nhớ là m bit
- Dung lượng chip nhớ = $2^n \times m$ (bit)
- Các đường điều khiển:
 - Tín hiệu chọn chip CS (Chip Select)
 - Tín hiệu điều khiển đọc OE (Output Enable)
 - Tín hiệu điều khiển ghi WE (Write Enable)
 - Các tín hiệu thường tích cực ở mức 0



Thiết kế modul nhớ bán dẫn



- Dung lượng chip nhớ $2^n \times m$ bit
- 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



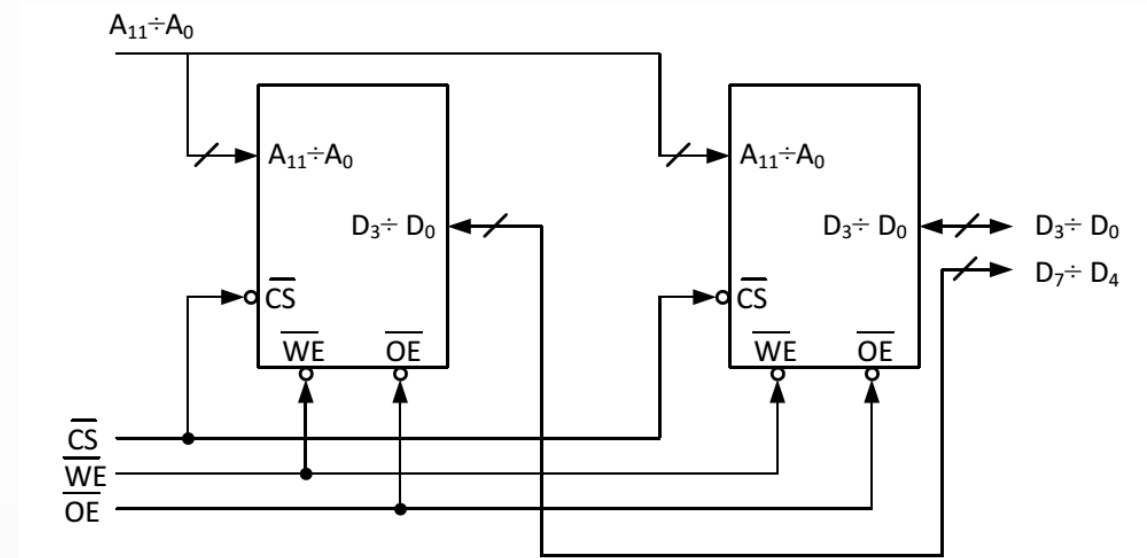
VD:

- Cho chip nhớ 4K x 4 bit
- Thiết kế modul nhớ 4K x 8 bit

Giải:

- Dung lượng chip nhớ: $2^{12} \times 4$ bit
- Chip nhớ:
 - 12 chân địa chỉ
 - 4 chân dữ liệu
- Modul nhớ:
 - 12 chân địa chỉ
 - 8 chân dữ liệu

=> Dùng 2 chip nhớ



Tăng độ dài từ nhớ



- Cho chip nhớ $2^n \times m$ bit
 - Thiết kế modul nhớ $2^n \times (k \times m)$ bit
- \Rightarrow Dùng k chip nhớ*

VD:

Cho chip nhớ $16K \times 4$ bit

Thiết kế modul nhớ $16K \times 16$ bit

Tăng số lượng từ nhớ



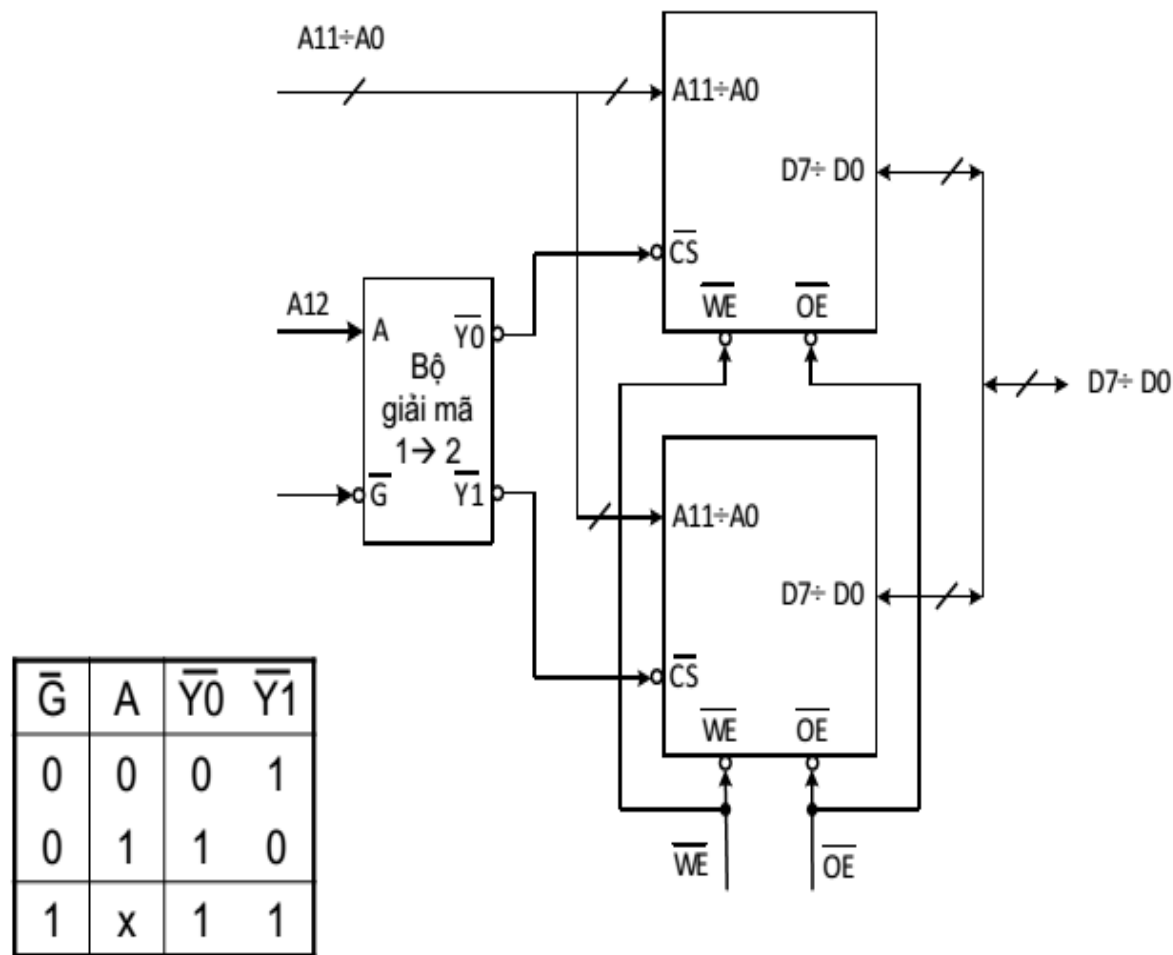
VD:

- Cho chip nhớ 4K x 8 bit
- Thiết kế modul nhớ 8K x 8 bit

Giải:

- Dung lượng chip nhớ: $2^{12} \times 8$ bit
- Chip nhớ:
 - 12 chân địa chỉ
 - 8 chân dữ liệu
- Modul nhớ
 - 13 chân địa chỉ
 - 8 chân dữ liệu

⇒ Dùng 2 chip nhớ,
1 bộ giải mã 1- \rightarrow 2



Tăng số lượng từ nhớ



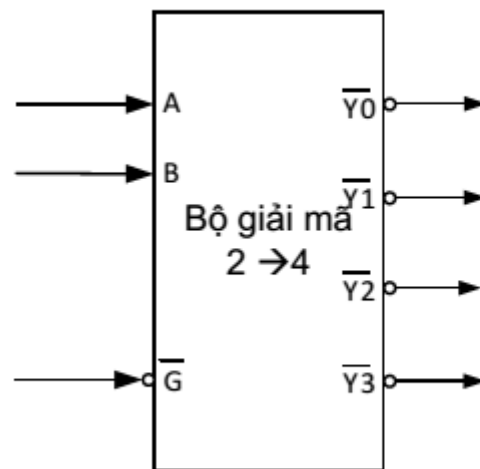
- Cho chip nhớ $2^n \times m$ bit
- Thiết kế modul nhớ $2^{n+k} \times m$ bit

*\Rightarrow Dùng 2^k chip nhớ
Bộ giải mã $k \rightarrow 2^k$*

VD:

Cho chip nhớ $2^n \times m$ bit

Thiết kế modul nhớ $2^{n+2} \times m$ bit



\bar{G}	B	A	\bar{Y}_0	\bar{Y}_1	\bar{Y}_2	\bar{Y}_3
0	0	0	0	1	1	1
0	0	1	1	0	1	1
0	1	0	1	1	0	1
0	1	1	1	1	1	0
1	x	x	1	1	1	1

Thiết kế kết hợp

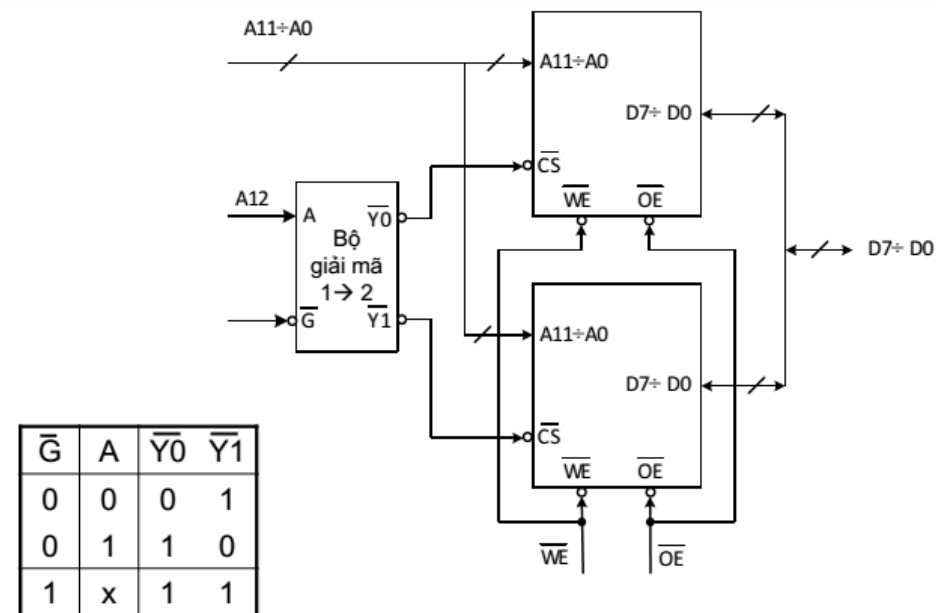
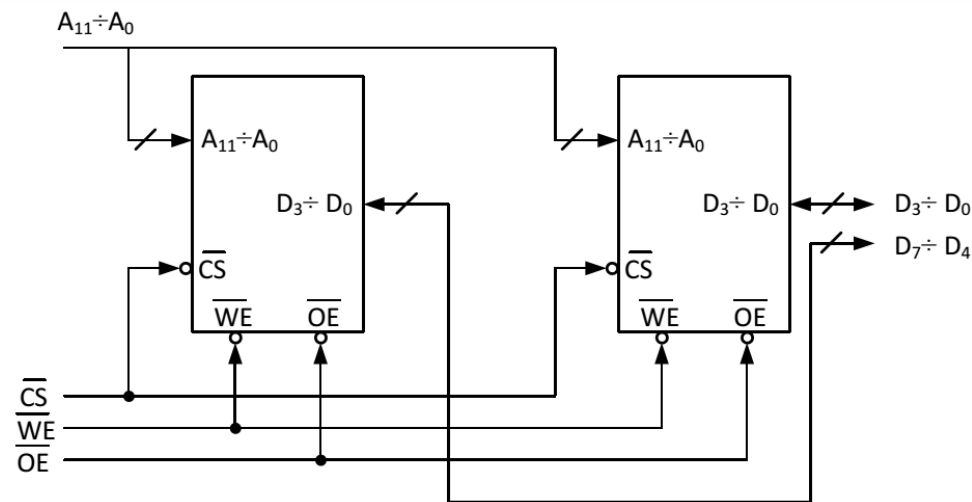
■ VD:

- Cho chip nhớ SRAM 4K x 4 bit
- Thiết kế modul nhớ 8K x 8 bit

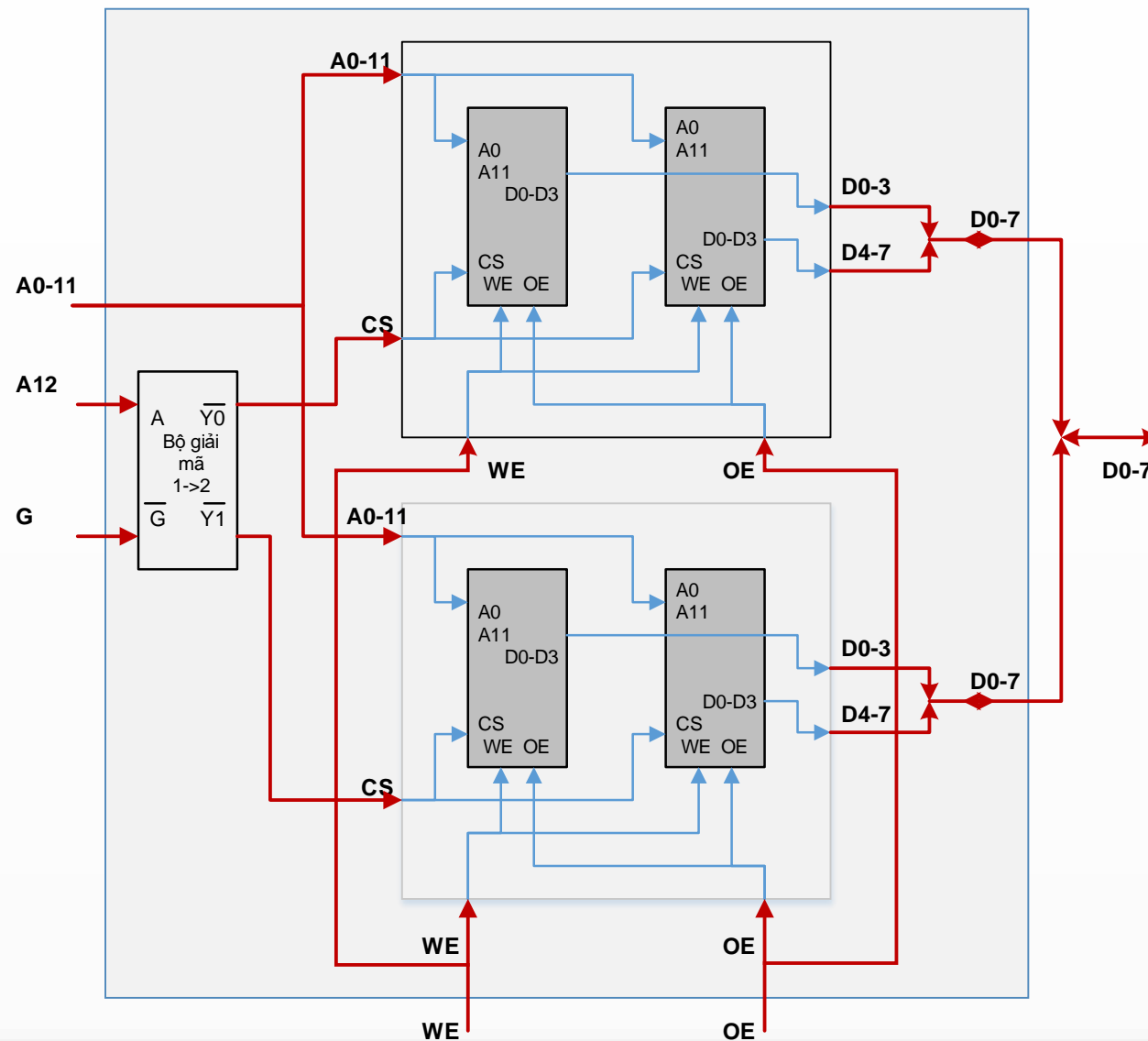
■ Giải:

- Kết hợp 2 phương pháp trên
- Chip nhớ: $2^{12} \times 4$ bit
 - 12 chân địa chỉ
 - 4 chân dữ liệu
- Modul nhớ: $2^{13} \times (2 \times 4)$ bit
 - 13 chân địa chỉ
 - 8 chân dữ liệu

=> Dùng 4 chip nhớ
1 bộ giải mã 1-2



Thiết kế kết hợp



Bài tập



- Thiết kế module nhớ 16K x 8 bit từ các chip nhớ 4K x 8 bit
- Thiết kế module nhớ 32K x 8 bit từ các chip nhớ 4K x 8 bit
- Thiết kế module nhớ 16K x 16 bit từ các chip nhớ 4K x 8 bit
- Thiết kế module nhớ 32M x 32 bit từ các chip nhớ 4M x 32 bit

Chương 7



1. Tổng quan hệ thống bộ nhớ
2. Bộ nhớ bán dẫn
3. Bộ nhớ chính
4. Bộ nhớ đệm (cache)
5. Bộ nhớ ngoài
6. Bộ nhớ ảo

Các đặc trưng cơ bản



- Tồn tại trên mọi hệ thống máy tính
- Chứa các chương trình đang thực hiện và các dữ liệu liên quan
- 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 có thể quản lý
- Việc quản lý logic bộ nhớ chính tùy thuộc vào hệ điều hành

Tổ chức bộ nhớ đan xen (interleaved memory)

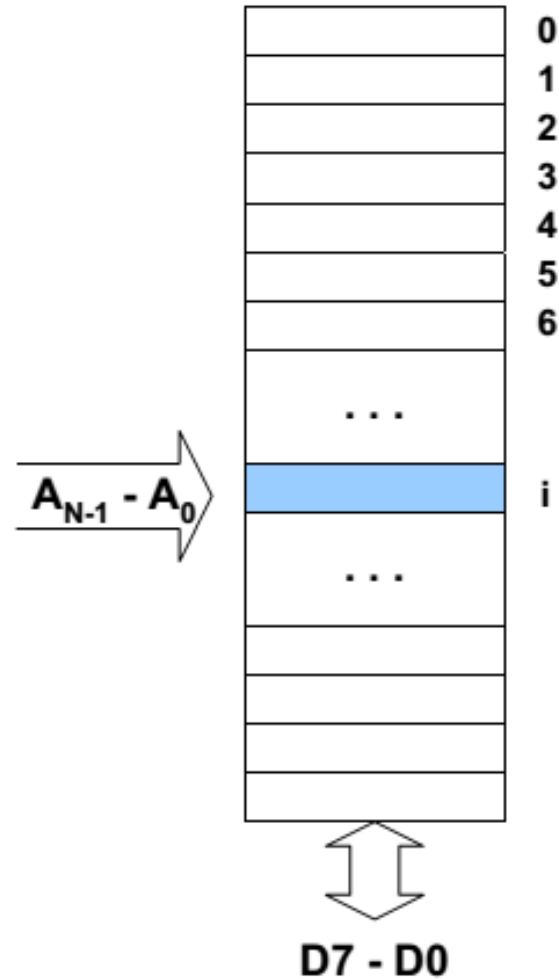


- Độ rộng của bus dữ liệu để trao đổi với bộ nhớ: $m = 8, 16, 32, 64, 128 \dots$ bit
- Các ngăn nhớ được tổ chức theo byte \Rightarrow tổ chức bộ nhớ vật lý khác nhau

m=8 bit



■ Bảng nhớ tuyến tính



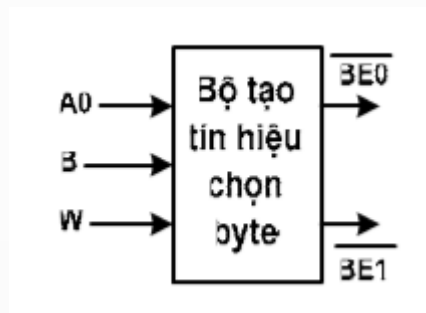
m = 16 bit



■ Hai băng nhớ đan xen

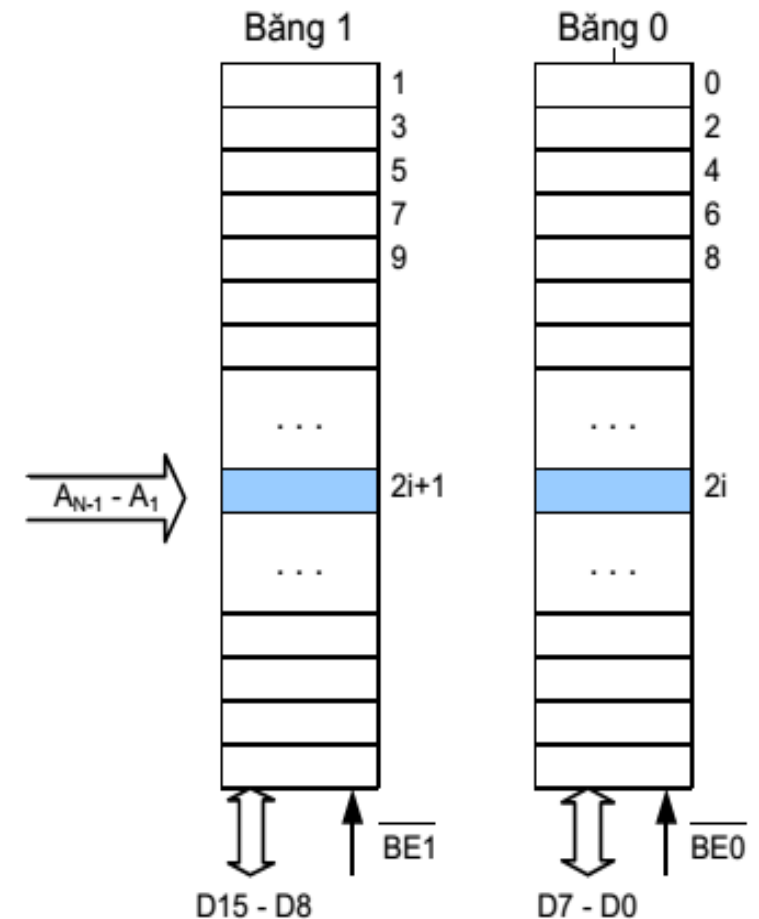
VD: Ta có 32 bit địa chỉ ta có thể quản lý bộ nhớ địa chỉ: $00000000H$ đến $FFFFFFFFH$, tức là 2^{32} byte = 4GB

Giả sử ta có 2 bank nhớ, mỗi bank 2GB. Đường tín hiệu A0 dùng để chọn Bank, A1-A31 được đưa vào bank nhớ để chọn ô nhớ



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



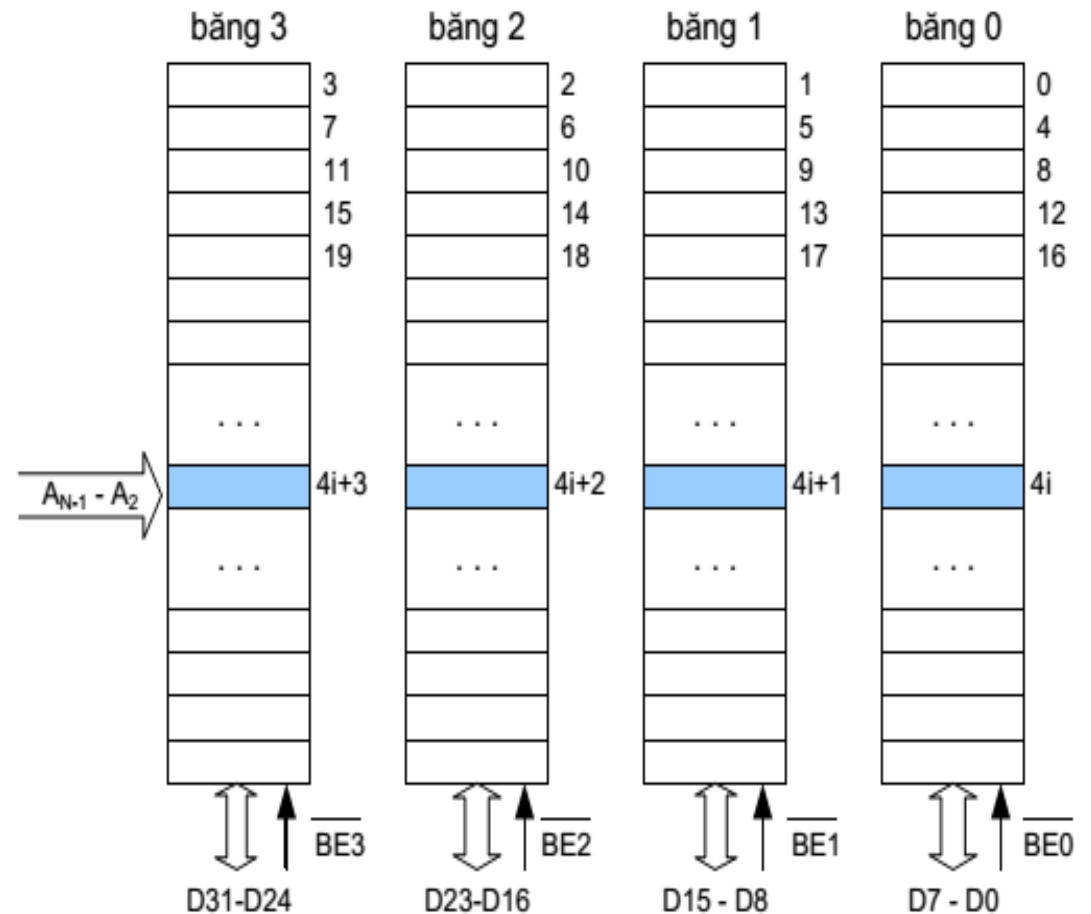
m = 32 bit



■ Bốn bảng nhớ đơn xen

VD: Ta có 32 bit địa chỉ ta có thể quản lý bộ nhớ địa chỉ: $00000000H$ đến $FFFFFFFFH$, tức là 2^{32} byte = 4GB

Giả sử ta có 4 bank nhớ, mỗi bank 1GB. Đường tín hiệu A0, A1 dùng để chọn Bank, A2-A31 được đưa vào bank nhớ để chọn ô nhớ



Chương 7



1. Tổng quan hệ thống bộ nhớ
2. Bộ nhớ bán dẫn
3. Bộ nhớ chính
4. Bộ nhớ đệm (cache)
5. Bộ nhớ ngoài
6. Bộ nhớ ảo

Bộ nhớ đệm (Cache)



Nguyên tắc chung

Các phương pháp ánh xạ

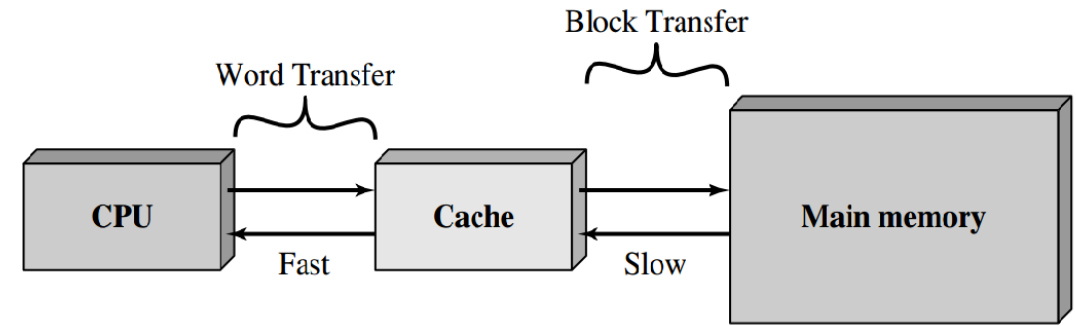
Thay thế block trong cache

Phương pháp ghi DL khi cache hit

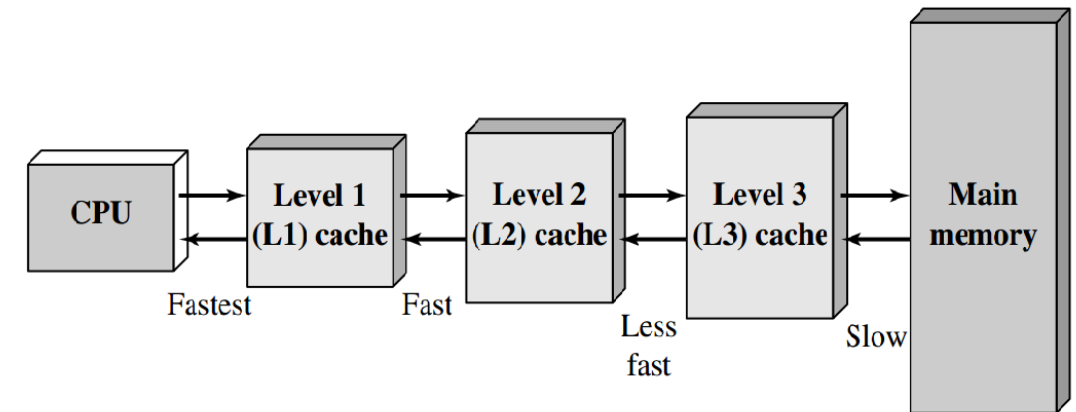
Bộ nhớ đệm (cache memory)



- Cache là bộ nhớ trung gian có tốc độ nhanh hơn bộ nhớ chính
- Cache nằm giữa CPU và bộ nhớ chính nhằm tăng tốc độ CPU truy cập bộ nhớ
- Cache thường được đặt trong CPU
- Cache có nhiều level, các level về sau dung lượng càng lớn hơn và tốc độ chậm hơn.



(a) Single cache



(b) Three-level cache organization

Cache



- Nguyên tắc nạp DL:

CPU yêu cầu DL, nó kiểm tra trong cache

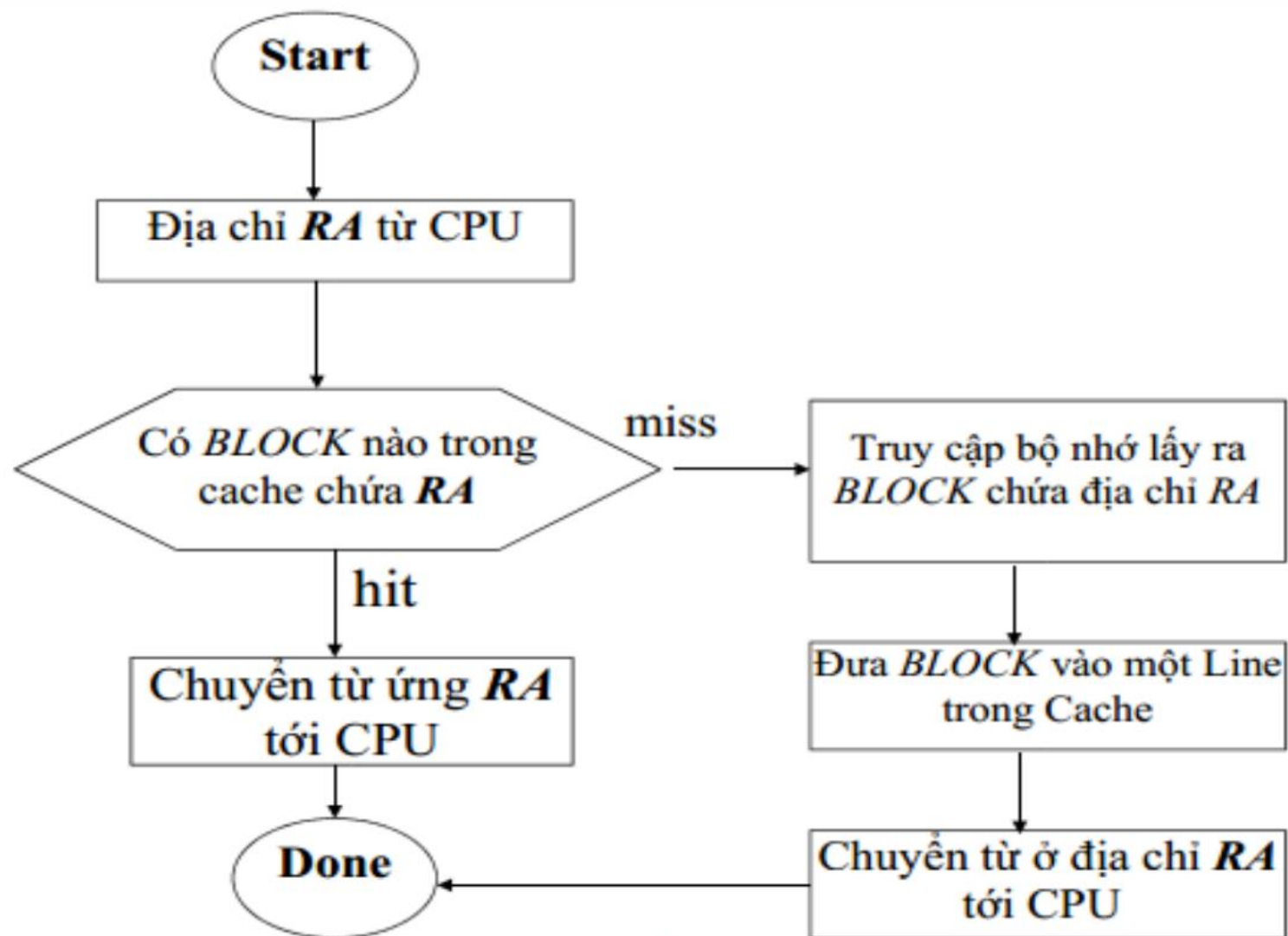
- Nếu có: CPU lấy DL từ cache (Cache hit)
- Không có: Nạp block tương ứng trong main memory vào cache, sau đó DL được chuyển vào CPU

- Tính địa phương: Trong một khoảng thời gian đủ nhỏ, CPU thường chỉ tham chiếu các thông tin trong một khối nhớ cục bộ

- Ví dụ:

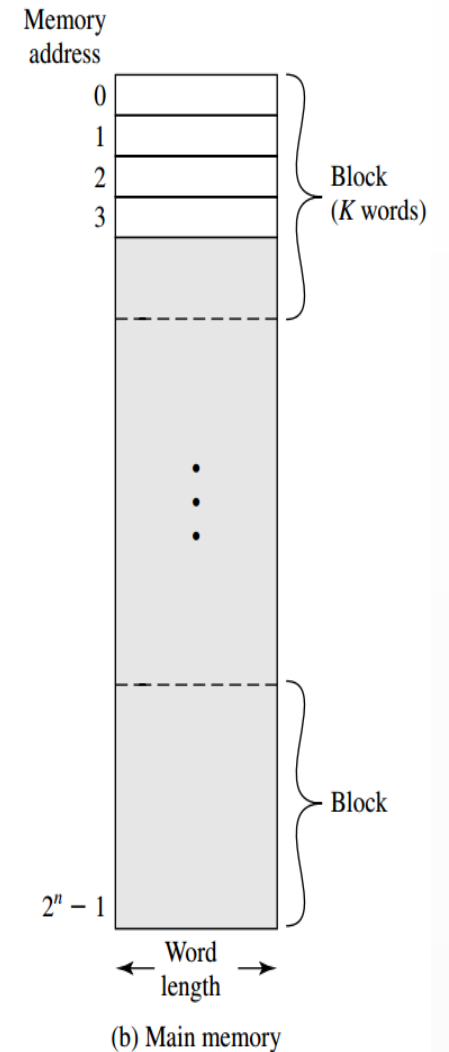
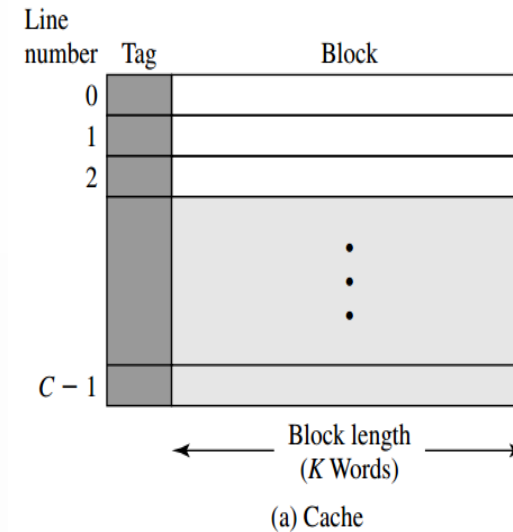
- Cấu trúc chương trình tuần tự
- Vòng lặp có thân nhỏ
- Cấu trúc dữ liệu mảng

Sơ đồ đọc DL



Cấu trúc của cache/memory

- Bộ nhớ chính và cache được chia thành các khối có kích thước bằng nhau
 - Bộ nhớ chính: $B_0, B_1, B_2, \dots, B_{p-1}$ (p Blocks)
 - Bộ nhớ cache: $L_0, L_1, L_2, \dots, L_{c-1}$ (c Lines)
 - Kích thước Block /Line: 8, 16, 32, 64, 128 byte
- Mỗi Line trong cache được gắn thẻ nhớ (Tag), Tag cho biết **Line** tương ứng **Block** nào trong bộ nhớ.



Cache/Main Memory Structure

Ví dụ



- Cho bộ nhớ 4GB, cache 256KB, kích thước Block/line 4byte

Tính C, M.

- Ta có:
 - $4\text{GB} = 2^{32}\text{byte}$
 - $M = 2^{32}/4 = 2^{30}$ (Blocks)
 - $256\text{KB} = 2^{18}\text{ byte}$
 - $C = 2^{18}/4 = 2^{16}$ (Lines)

Bộ nhớ Cache

Tag	Line 1
	Line 2
	Line 3
	...
	Line C

Bộ nhớ chính

Block 1
Block 2
Block 3
Block 4
...
Block M-2
Block M-1
Block M

Bộ nhớ đệm (Cache)



Nguyên tắc chung

Các phương pháp ánh xạ

Thay thế block trong cache

Phương pháp ghi DL khi cache hit

Các phương pháp ánh xạ



- 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 được nạp duy nhất vào một Line của cache**
- **Địa chỉ CPU được chia làm 3 trường**
 - **Trường Word:** Gồm W bit xác định một từ nhớ trong Block hay Line
 $2^W = \text{kích thước của Block hay Line}$
 - **Trường Line:** Gồm L bit xác định một trong số các Line trong cache
 $2^L = \text{số Line trong cache} = c$
 - **Trường Tag:** Gồm T bit
 $T = N - (W+L)$

$B_0 \Rightarrow L_0$	Lặp
$B_1 \Rightarrow L_1$	$B_c \Rightarrow L_0$
....	$B_{c+1} \Rightarrow L_1$
$B_{c-1} \Rightarrow L_{c-1}$
<u>Line</u>	<u>Block</u>
0	0,c,2c,3c,...
1	1, c+1, 2c+1,...
...	
c-1	c-1,2c-1,3c-
1,...	

Ánh xạ trực tiếp



Tag của cache:

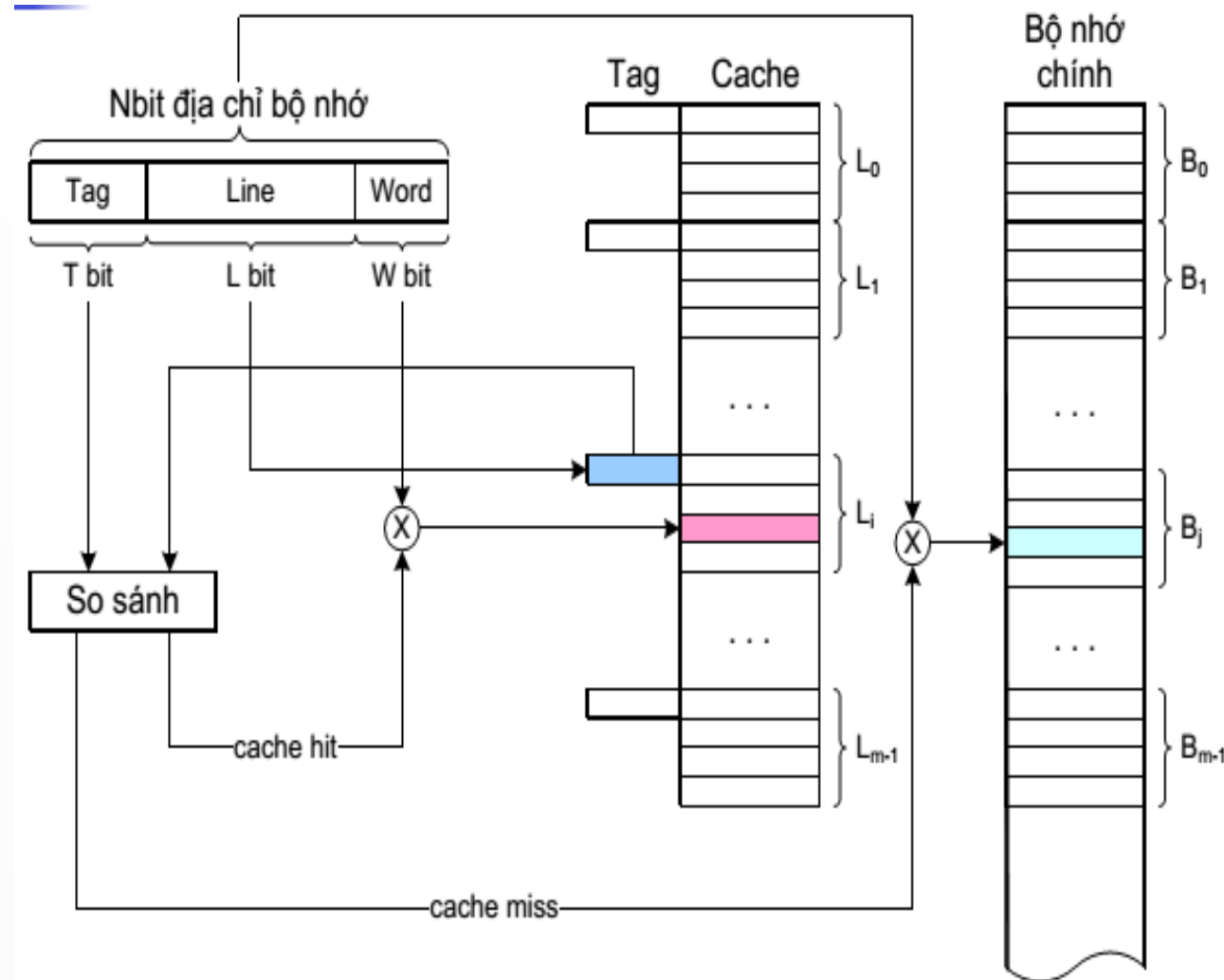
Chứa T bit cao của đ/c của Block nhớ tương ứng

CPU phát ra một đ/c

- L bit của trường Line x/đ Line tương ứng
- Nội dung Tag ở Line được ss với T bit cao của đ/chỉ
 - Giống: cache hit
 - Khác: cache miss
- W bit xác định word trong line

Ưu: Bộ so sánh đơn giản

Nhược: Xác suất cache hit thấp



Ánh xạ trực tiếp



- **VD:** Giả sử bộ nhớ được đánh địa chỉ theo byte
 - Không gian main memory: 4GB
 - Cache: 128KB
 - Block: 8 byte
 - Xác định kích thước các trường trong địa chỉ?
- **Ta có:**
 - Cache: 128 KB = 2^{17} byte
 - Block: 8 byte = 2^3 byte $\Rightarrow W=3$
 - Số Line trong Cache: $2^{17}/2^3 = 2^{14}$
 $\Rightarrow L=14$
 - Main memory: 4GB = 2^{32} byte $\Rightarrow N=32$
 $\Rightarrow T = N - (W + L) = 32 - 17 = 15$

Tag	Line	Word
15	14	3

Các phương pháp ánh xạ



- 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 liên kết toàn phần



- **Mỗi Block có thể nạp vào bất kỳ Line nào**
- Địa chỉ của CPU chia thành 2 trường
 - **Trường Word:** Gồm W bit, xác định word trong Line
 - **Trường Tag:** T bit, dùng để xác định Block của bộ nhớ chính
 $T = N - W$
- **Ưu:** Xác suất cache hit cao
- **Nhược:**
 - So sánh đồng thời với tất cả các Tag \Rightarrow mất nhiều thời gian
 - Bộ so sánh phức tạp \Rightarrow Ít sử dụng

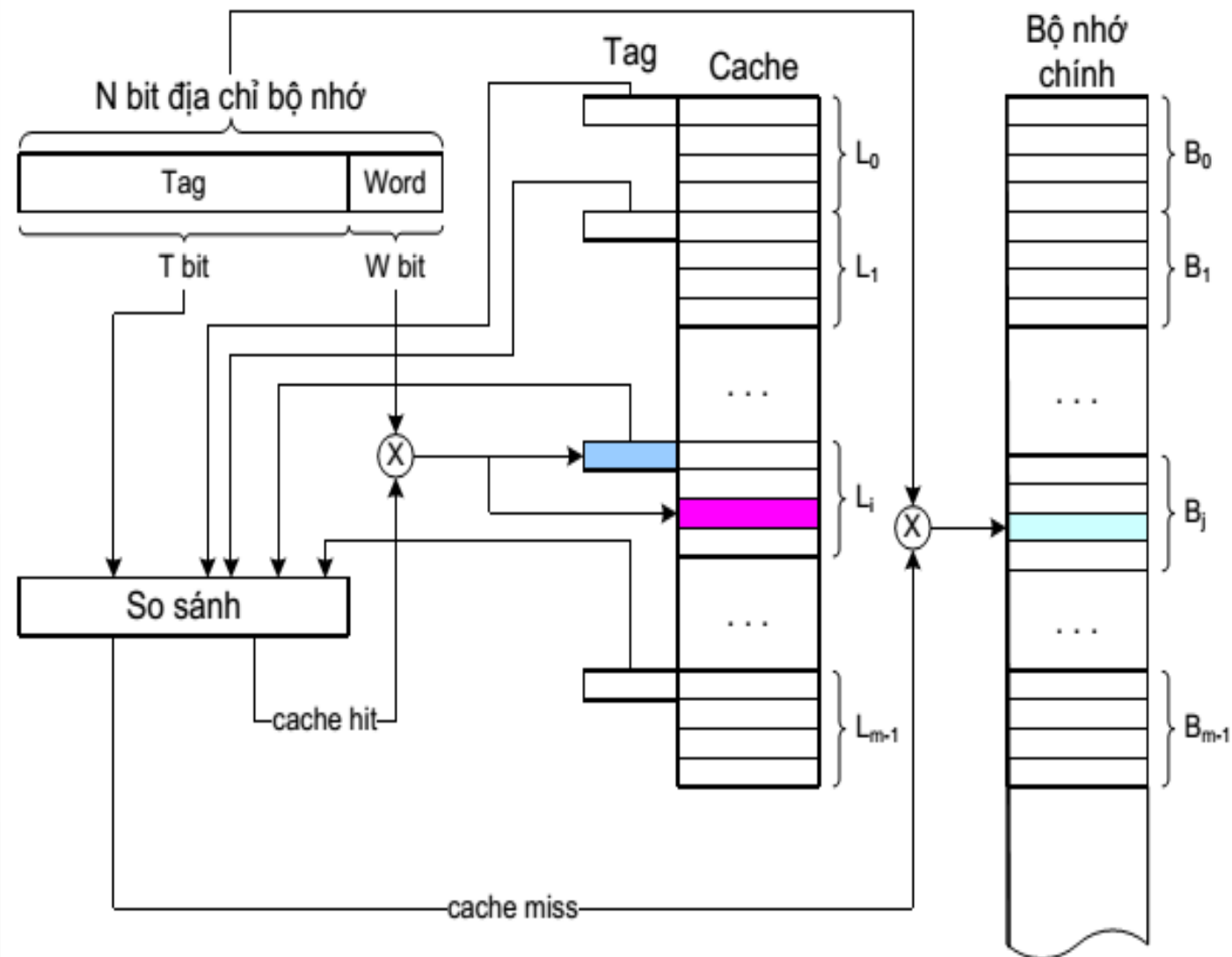
Ánh xạ liên kết toàn phần

■ Tag của cache:

Là T bit cao của đ/c của Block nhớ

■ CPU phát ra một đ/c

- Nội dung Tag tất cả các Line được ss với T bit cao của đ/chỉ
 - Giống: cache hit
 - Khác: cache miss
- W xác định word trong line



Ánh xạ liên kết toàn phần



- VD: Giả sử bộ nhớ được đánh địa chỉ theo byte
 - Không gian main memory: 4GB
 - Cache: 128KB
 - Block: 8 byte
 - Xác định kích thước các trường trong địa chỉ?
- Ta có:
 - Cache: 128 KB = 2^{17} byte
 - Block: 8 byte = 2^3 byte $\Rightarrow W=3$
 - Main memory: 4GB = 2^{32} byte $\Rightarrow N=32$
$$\Rightarrow T = N - W = 32 - 3 = 29$$

Tag	Word
29	3

Các phương pháp ánh xạ



- 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 liên kết tập hợp



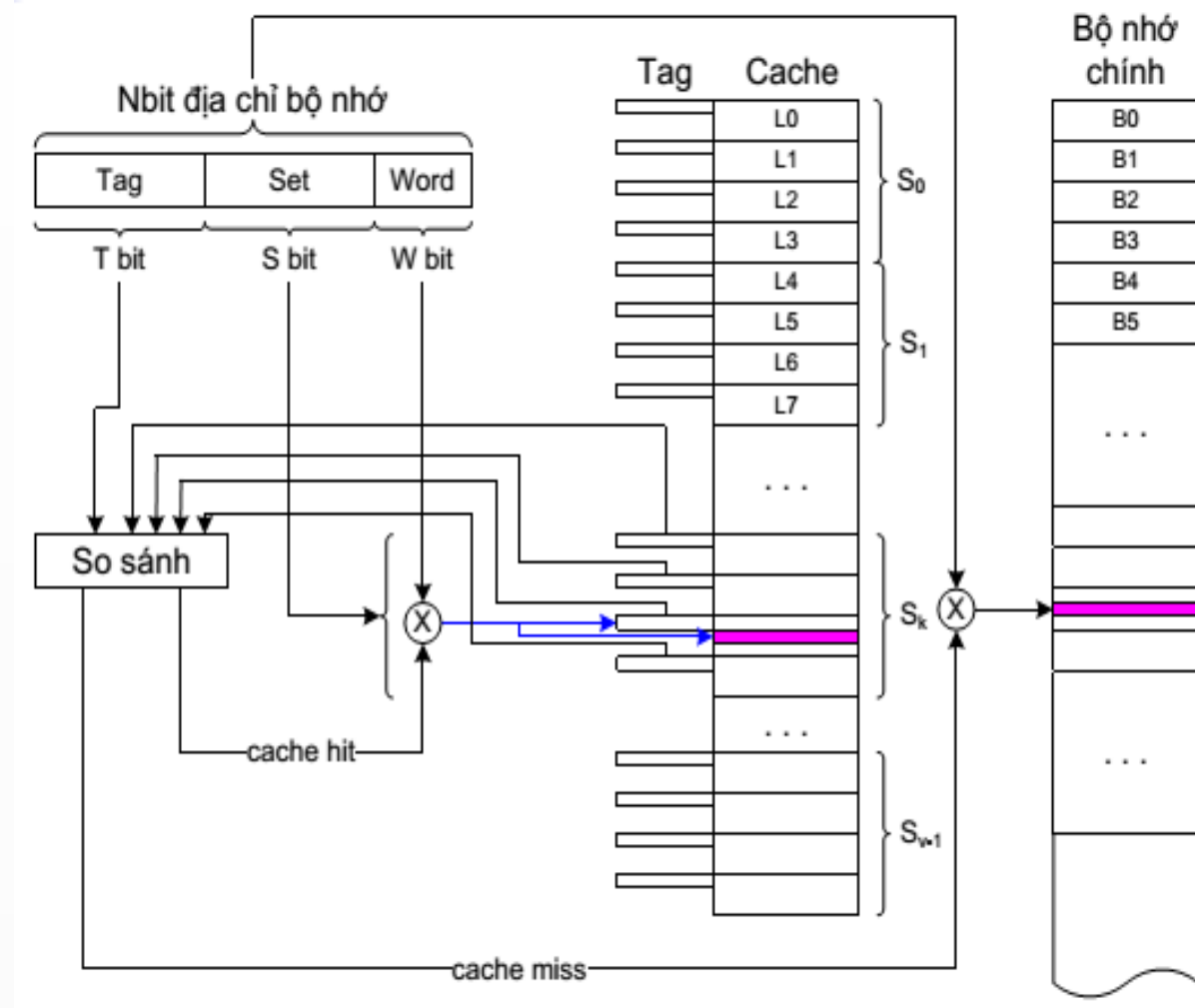
- Dung hòa hai phương pháp trên
 - Cache được chia thành các Tập (Set)
 - Mỗi một Set chứa một số Line
 - **Mỗi Block chỉ được ghi vào các Line nằm trong 1 set tương ứng**
- Địa chỉ CPU được chia làm 3 trường
 - **Trường Word:** Gồm **W** bit xác định một từ nhớ trong Block hay Line
 2^W = kích thước của Block hay Line
 - **Trường Set:** Gồm **S** bit xác định set tương ứng
 2^S = số Set trong cache
 - **Trường Tag:** Gồm **T** bit
 $T = N - (W+S)$

- Ví dụ:
 - 4 Line/Set => 4-way associative mapping
- Thông dụng: 2,4,8,16 Lines/Set
- Nguyên tắc ánh xạ:
 $B_0 \Rightarrow S_0$
 $B_1 \Rightarrow S_1$
 $B_2 \Rightarrow S_2$
....

Ánh xạ liên kết tập hợp



- **Tag của mỗi Line trong cache:**
Chứa T bit cao của đ/c của Block nhớ tương ứng
- **CPU phát ra một đ/c**
 - S bit của trường Set x/đ Set tương ứng
 - Nội dung Tag của các Line trong Set được ss với T bit cao của đ/chỉ
 - Giống: cache hit
 - Khác: cache miss
 - W bit xác định word trong line
- **Ưu/ Nhược:** Dung hòa 2 pp trên



Ánh xạ liên kết tập hợp



- VD: Giả sử bộ nhớ được đánh địa chỉ theo byte
 - Không gian main memory: 4GB
 - Cache: 128KB
 - Block: 8 byte
 - Xác định kích thước các trường trong địa chỉ của tổ chức ánh xạ liên kết tập hợp 4-đường?

- Ta có:
 - 4 Line/Set => 4-way associative mapping
 - Cache: 128 KB = 2^{17} byte
 - Block: 8 byte = 2^3 byte => $W=3$
 - Số Set trong Cache: $2^{17}/2^3/4 = 2^{12}$
=> $S=12$
 - Main memory: 4GB = 2^{32} byte => $N=32$
=> $T = N-(W+S) = 32-15=17$

Tag	Set	Word
17	12	3

Bài tập



- **VD:** Giả sử bộ nhớ được đánh địa chỉ theo byte
 - Không gian main memory: 8GB
 - Cache: 256KB
 - Block: 16 byte
 - Xác định kích thước các trường trong địa chỉ trong các trường hợp:
 1. Ánh xạ trực tiếp
 2. Ánh xạ liên kết toàn phần
 3. Ánh xạ liên kết tập hợp 8- đường

Bộ nhớ đệm (Cache)



Nguyên tắc chung

Các phương pháp ánh xạ

Thay thế block trong cache

Phương pháp ghi DL khi cache hit

Thay thế Block trong cache



Với ánh xạ trực tiếp:

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

Thay thế Block trong cache



Với ánh xạ liên kết: Cần có thuật toán

- **Random**: Thay thế ngẫu nhiên
- **FIFO** (First In First Out): Thay thế Block nào nằm lâu nhất
- **LFU** (Least Frequently Used): Thay thế Block nào có số lần truy nhập ít nhất trong cùng một khoảng thời gian
- **LRU** (Least Recently Used): Thay thế Block có thời gian lâu nhất không được tham chiếu tới

Kết quả tốt nhất: **LRU**

Bộ nhớ đệm (Cache)



Nguyên tắc chung

Các phương pháp ánh xạ

Thay thế block trong cache

Phương pháp ghi DL khi cache hit

Phương pháp ghi dữ liệu khi cache hit



- Ghi xuyên qua (Write-through):
 - Ghi đồng thời cả cache và bộ nhớ chính
 - Tốc độ chậm
- Ghi sau (Write-back):
 - Chỉ ghi ra cache
 - Ghi trả về bộ nhớ chính khi Line trong cache bị thay thế
 - Tốc độ nhanh

Chương 7



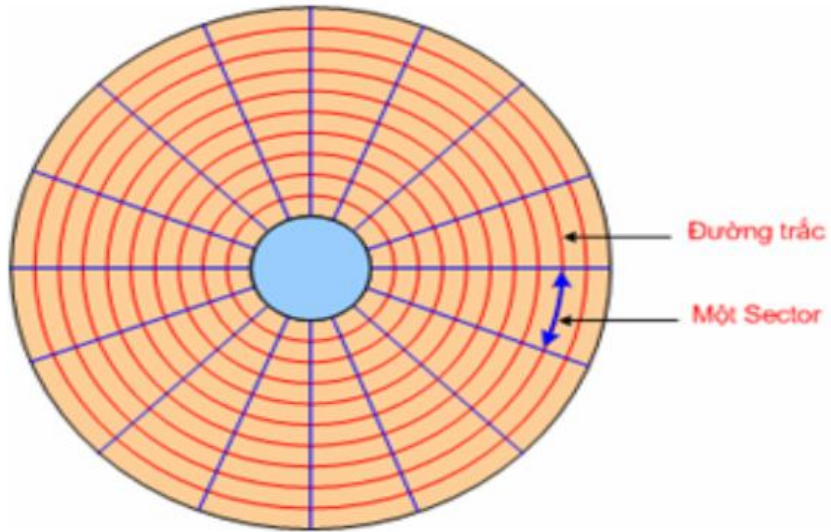
1. Tổng quan hệ thống bộ nhớ
2. Bộ nhớ bán dẫn
3. Bộ nhớ chính
4. Bộ nhớ đệm (cache)
5. Bộ nhớ ngoài
6. Bộ nhớ ảo

Bộ nhớ ngoài

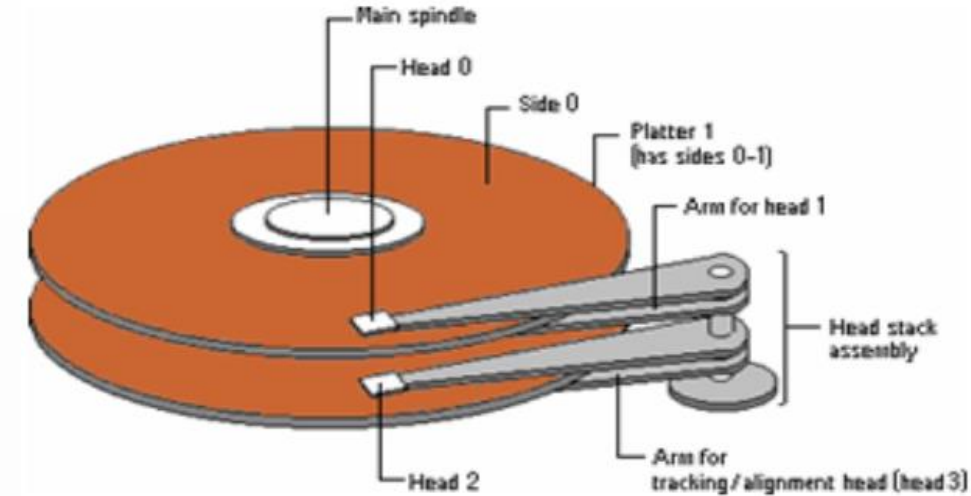


- Tồn tại dưới dạng các thiết bị lưu trữ
- Các kiểu bộ nhớ ngoài
 - Băng từ: Ít sử dụng
 - Đĩa từ: Ổ đĩa cứng HDD (Hard Disk Drive)
 - Đĩa quang: CD, DVD
 - Bộ nhớ Flash:
 - Ổ nhớ thể rắn SSD (Solid State Drive)
 - USB flash
 - Thẻ nhớ

Ổ đĩa cứng (HDD – Hard Disk Drive)



Bề mặt của đĩa cứng, tín hiệu ghi trên các đường tròn đồng tâm gọi là Track, mỗi Track được chia làm nhiều Sector



HDD



■ Đĩa từ:

- HDD gồm nhiều đĩa từ làm bằng nhôm, gốm, thủy tinh, xếp chồng lên nhau và gắn vào cùng một trục, khi mô tơ quay tất cả các đĩa đều quay, vận tốc 5400-7200 vòng/phút
- Đĩa từ gồm 2 mặt
- Đĩa được chia thành nhiều track, mỗi track chia thành nhiều sector
- VD: Đĩa 10GB: 7000 track/ mặt, 200 sector/ 1 track
- Tăng dung lượng: Chia thành nhiều Track, vòng càng lớn càng nhiều track hơn, nhiều sector hơn, đòi hỏi TB có độ chính xác cao hơn
- Dung lượng:

Số Sides x số tracks x số sector trung bình trên track x 512 byte

HDD



- Đầu từ đọc/ ghi: Mỗi mặt đĩa có 2 đầu từ đọc/ghi
- Mô tơ điều khiển đầu từ: Giúp đầu từ dịch chuyển trên mặt đĩa
- Mạch điều khiển: Giúp điều khiển
 - Tốc độ quay
 - Dịch chuyển đầu từ
 - Mã hóa và giải mã các tín hiệu đọc ghi
- Ưu/ nhược:
 - Dung lượng lớn, giá thành rẻ
 - Tốc độ truy nhập chậm

HDD



- **Hiệu năng:**
 - Thời gian truy nhập
 - Dung lượng lưu trữ
- **Thời gian truy nhập:**
 - Thời gian định vị đầu từ
 - Thời gian tìm kiếm DL
 - Thời gian truyền DL

SSD (Solid State Drive)



- Được sinh ra để cạnh tranh với HDD
- Bộ nhớ bán dẫn được cấu thành từ nhiều chip nhớ flash
- Tốc độ truy nhập rất nhanh
- Tiêu thụ năng lượng ít
- An toàn, độ bền cao
- Tuy nhiên dung lượng thấp: 128, 256GB, 512GB, 1TB
- Chi phí rất cao



Đĩa quang



- Dùng công nghệ lazer để khắc vào mặt đĩa
 - CD (Compact Disc)
Dung lượng thông dụng 650MB
 - DVD (Digital Versatile Disk)
Dung lượng có thể lên 17GB nếu là đĩa 2 mặt
- Một số qui tắc tên
 - CD/DVD-ROM: Là loại đĩa chỉ để đọc
 - CD/DVD-W: Là loại đĩa chỉ ghi được 1 lần
 - CD/DVD-RW: Là loại đĩa chỉ ghi được nhiều lần

Hệ thống lưu trữ dung lượng lớn: RAID



- Redundant Array of Inexpensive Disks
- (Redundant Array of Independent Disks)
- Tập các ổ đĩa cứng vật lý được OS coi như một ổ logic duy nhất => dung lượng lớn
- Dữ liệu được lưu trữ phân tán trên các ổ đĩa vật lý => truy cập song song (nhanh)
- Lưu trữ thêm thông tin dư thừa, cho phép khôi phục lại thông tin trong trường hợp đĩa bị hỏng => an toàn thông tin
- Phổ biến: RAID 0 – 6
- Lý do:
 - Dự phòng
 - Hiệu quả cao
 - Giá thành hạ

RAID 0, 1, 2



RAID 0

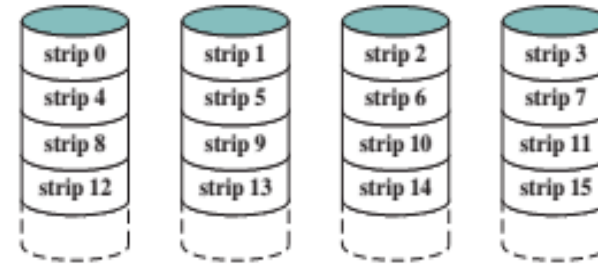
- Dùng kỹ thuật *stripping*
- Thực chất ko phải là 1 cấp độ của RAID
- Ko có dự phòng DL
- Tăng thời gian truy cập bộ nhớ
- Ko làm mất dung lượng nhớ

RAID 1

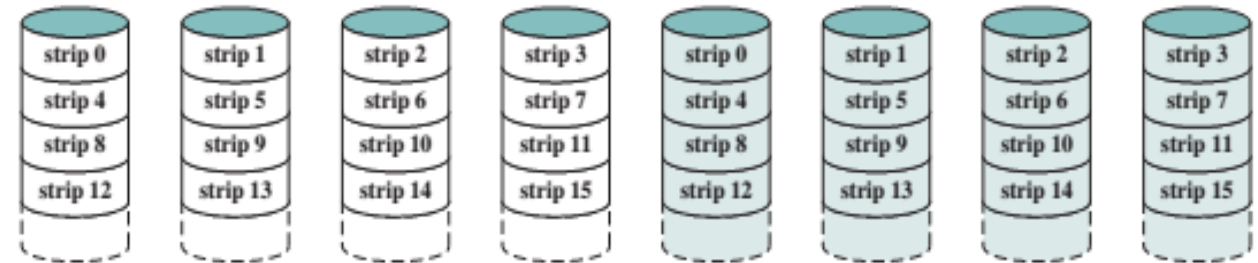
- Kỹ thuật *Mirroring*
- Dung lượng dự phòng lớn
- Nếu chỉ 2 ổ cứng ko tăng hquả thực thi

RAID 2:

- Dùng kỹ thuật sửa lỗi Hamming
- Dùng 1 bit phát hiện lỗi, 2 bit để sửa lỗi
- Mất nhiều dung lượng cho dự phòng



(a) RAID 0 (Nonredundant)



(b) RAID 1 (Mirrored)



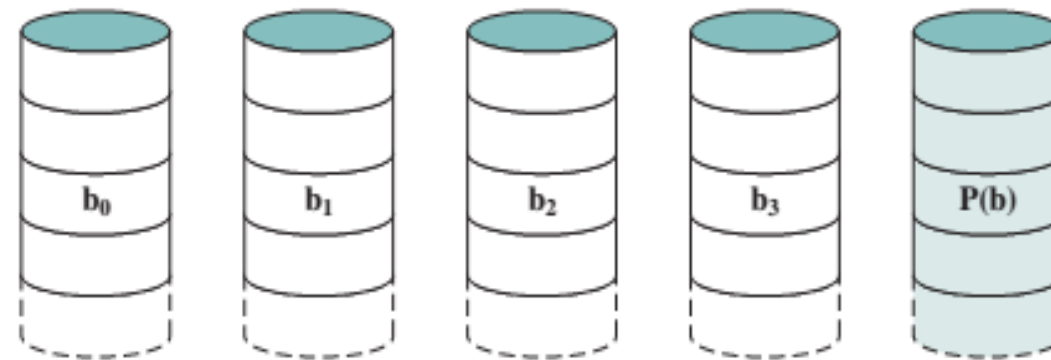
(c) RAID 2 (Redundancy through Hamming code)

RAID 3 & 4

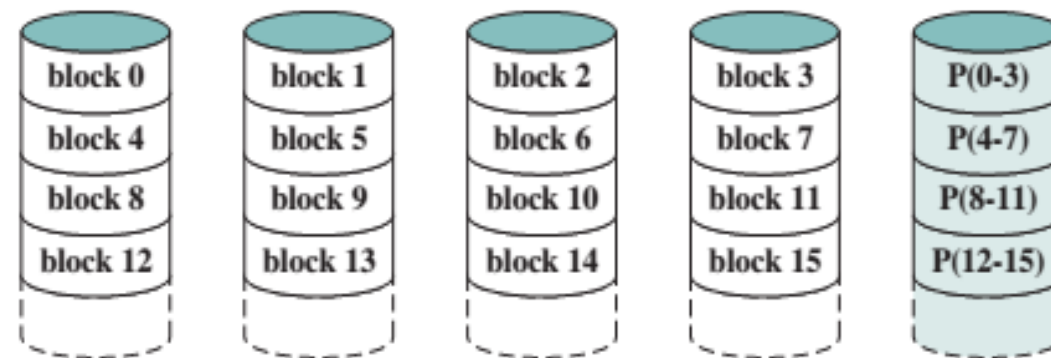


RAID 3

- Dùng kỹ thuật “*bit parity*”
- Mỗi khi đọc: DL được đọc đồng thời các đĩa, nếu phát hiện lỗi nó tự sửa lỗi



(d) RAID 3 (Bit-interleaved parity)



(e) RAID 4 (Block-level parity)

RAID 5 & 6

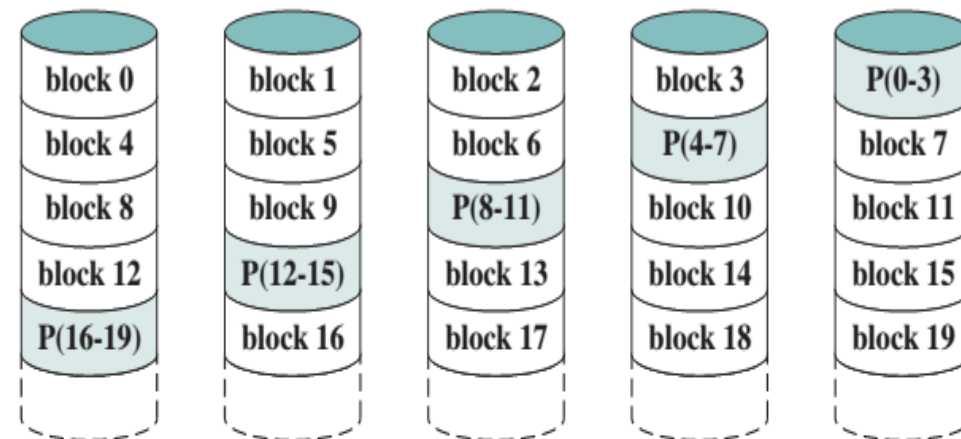


■ RAID 5

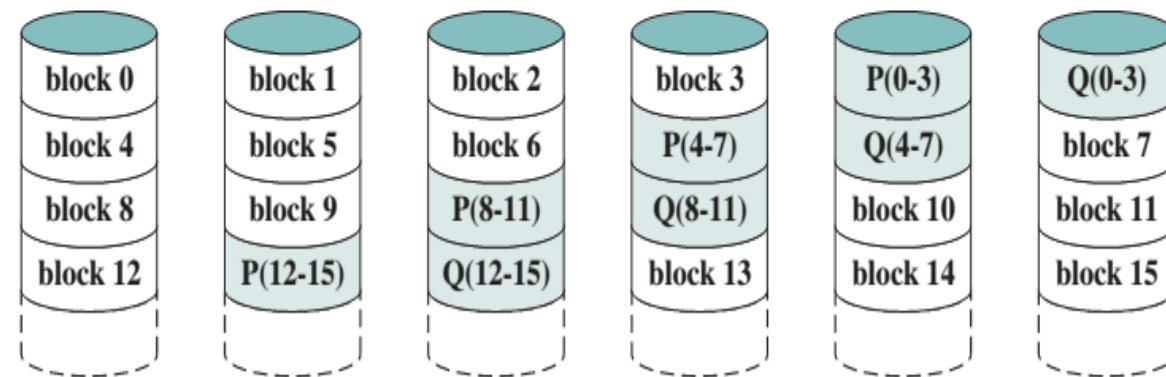
- Dùng cho phiên bản máy để bàn
- Dùng kỹ thuật “Parity”, khắc phục cổ chai RAID 4
- Tăng dung lượng lưu trữ so với RAID 1

■ RAID 6

- Nâng cấp RAID 5
- Có 2 nơi để lưu trữ dự phòng
- Tăng khả năng sửa lỗi, cho phép hỏng 2 ổ 1 lúc
- Giảm dung lượng lưu trữ, chỉ sd khi DL là rất quan trọng

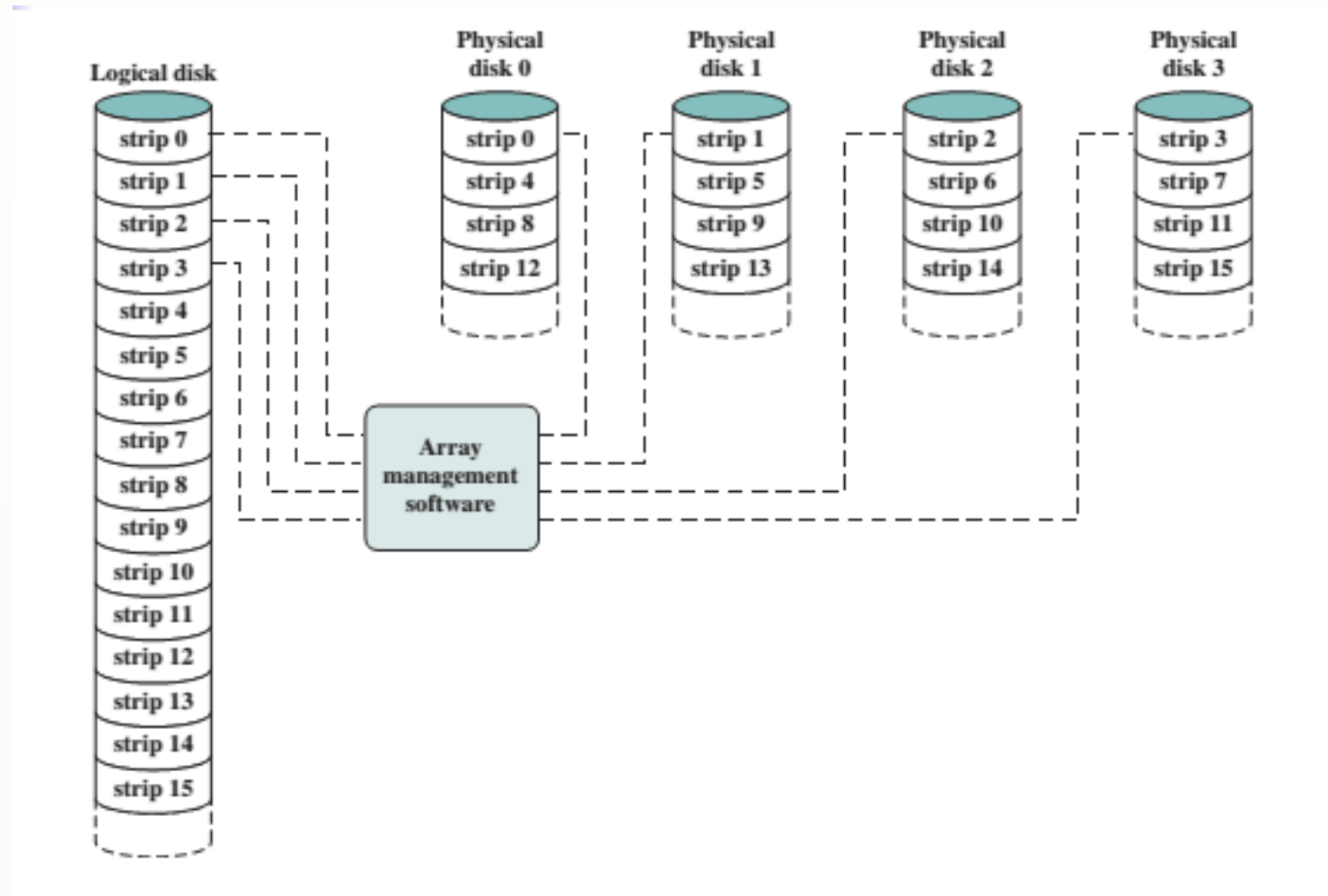


(f) RAID 5 (Block-level distributed parity)



(g) RAID 6 (Dual redundancy)

Ảnh xạ dữ liệu của RAID 0



RAID Level



Category	Level	Description	Disks Required	Data Availability	Large I/O Data Transfer Capacity	Small I/O Request Rate
Striping	0	Nonredundant	N	Lower than single disk	Very high	Very high for both read and write
Mirroring	1	Mirrored	$2N$	Higher than RAID 2, 3, 4, or 5; lower than RAID 6	Higher than single disk for read; similar to single disk for write	Up to twice that of a single disk for read; similar to single disk for write
Parallel access	2	Redundant via Hamming code	$N + m$	Much higher than single disk; comparable to RAID 3, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
	3	Bit-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
Independent access	4	Block-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 5	Similar to RAID 0 for read; significantly lower than single disk for write	Similar to RAID 0 for read; significantly lower than single disk for write
	5	Block-interleaved distributed parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 4	Similar to RAID 0 for read; lower than single disk for write	Similar to RAID 0 for read; generally lower than single disk for write
	6	Block-interleaved dual distributed parity	$N + 2$	Highest of all listed alternatives	Similar to RAID 0 for read; lower than RAID 5 for write	Similar to RAID 0 for read; significantly lower than RAID 5 for write

Chương 7



1. Tổng quan hệ thống bộ nhớ
2. Bộ nhớ bán dẫn
3. Bộ nhớ chính
4. Bộ nhớ đệm (cache)
5. Bộ nhớ ngoài
6. Bộ nhớ ảo

HẾT CHƯƠNG 7