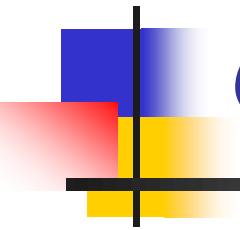
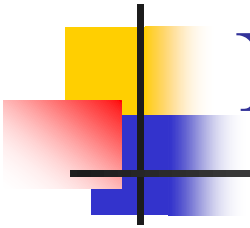


Chương 8: Hệ thống lưu trữ lớn

Tìm hiểu cơ chế quản lý thiết bị lưu trữ (thứ cấp) lớn của Hệ điều hành



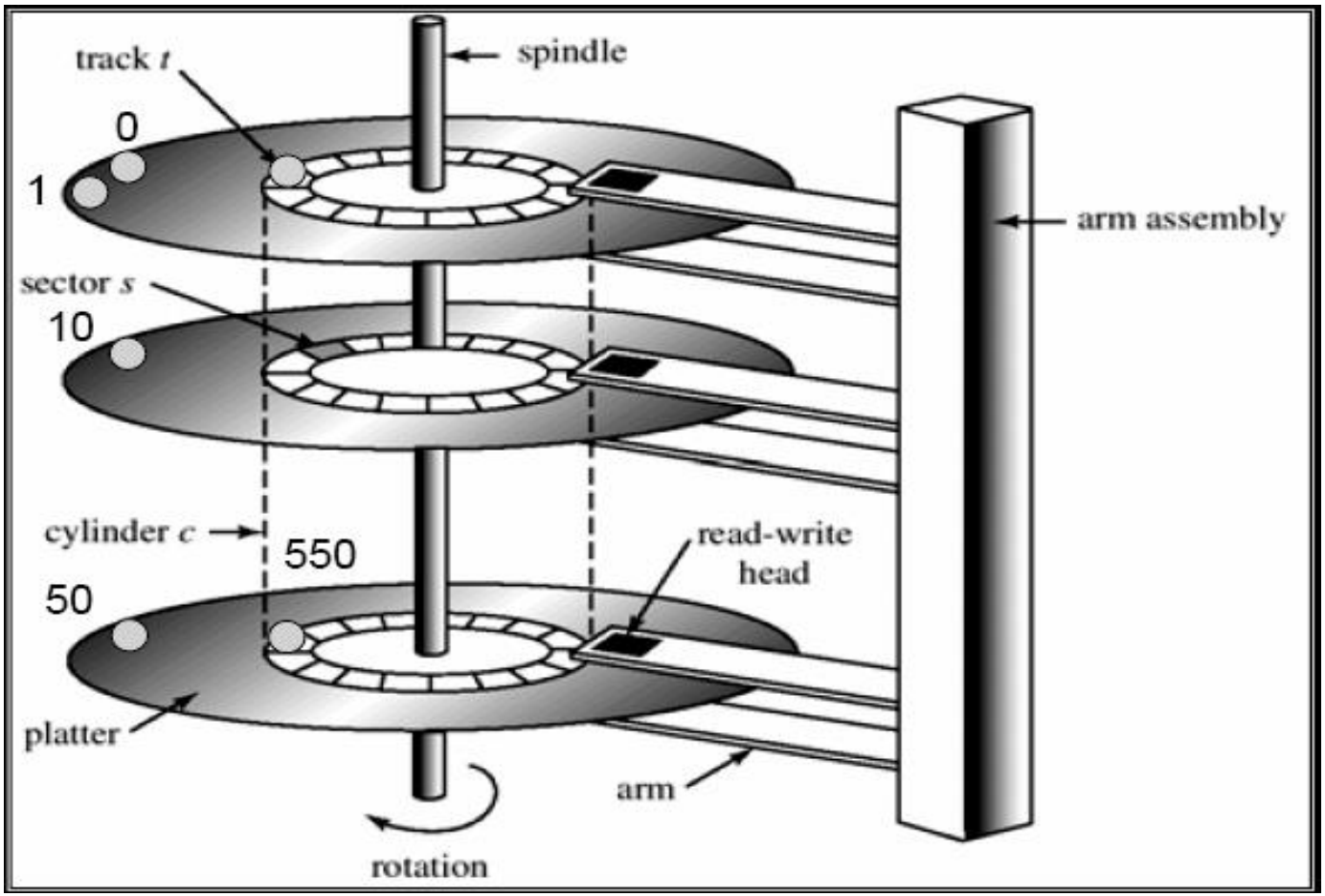
Nội dung

- Cấu trúc đĩa Disk Structure
- Lập lịch đĩa Disk Scheduling
- Quản lý đĩa Disk Management
- Quản lý không gian hoán đổi Swap-Space Management
- Cấu trúc RAID RAID Structure
- Nối kết đĩa Disk Attachment
- Các thiết bị lưu trữ cấp ba Tertiary Storage Devices

1. Cấu trúc đĩa(1)

- Đĩa được đánh địa chỉ là mảng 1 chiều của các khối logic, khối logic là đơn vị nhỏ nhất trong chuyển dữ liệu.
- Mảng 1 chiều của các khối logic được ánh xạ vào các sector của đĩa một cách tuần tự.
  - Sector 0 là sector đầu tiên của track đầu tiên trên cylinder ngoài cùng.
  - Việc ánh xạ tiếp tục theo thứ tự qua track đó, rồi đến các track còn lại trong cylinder đó, rồi đến các cylinder còn lại từ ngoài vào trong.

1. Cấu trúc đĩa(2)



2. Lập lịch đĩa - Disk scheduling (1)

- HĐH chịu trách nhiệm sử dụng các ổ đĩa một cách hiệu quả, có nghĩa đĩa phải có thời gian truy nhập nhanh và dải thông rộng.
- Thời gian truy nhập có 2 thành phần chính
  - *Thời gian định vị (Seek time):* là thời gian chuyển đầu từ tới cylinder chứa sector được yêu cầu.
  - *Trễ quay (Rotational latency):* là thời gian cộng thêm chờ đĩa quay sector được yêu cầu tới đầu từ.

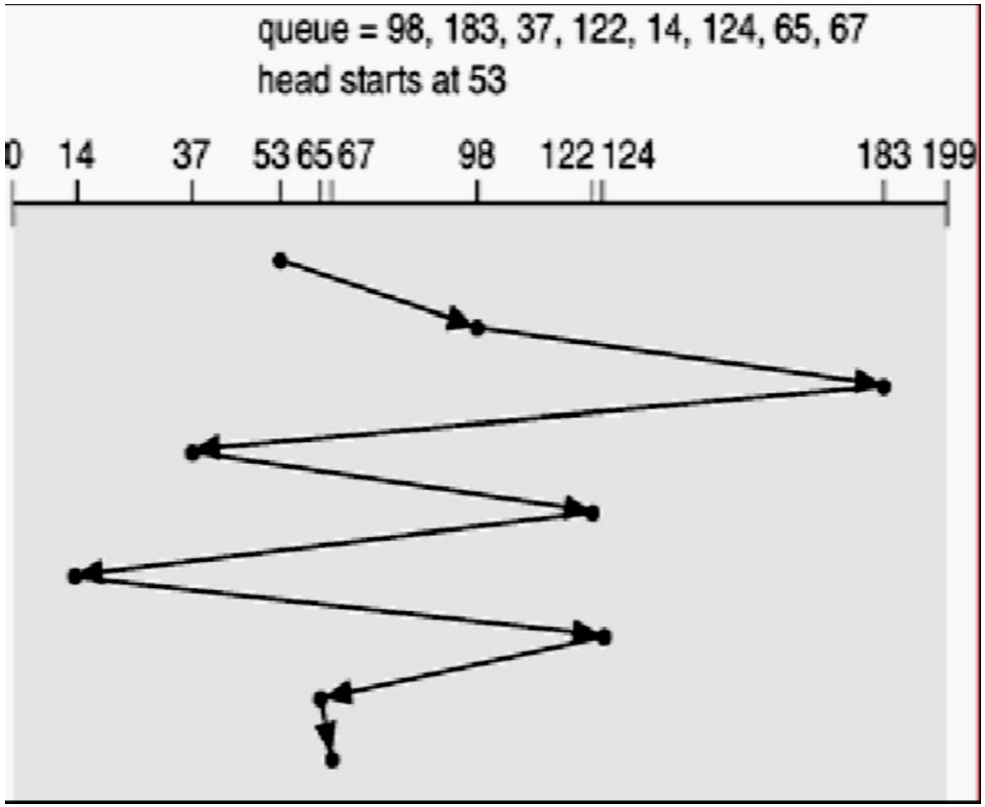
  -> Tối thiểu hóa seek time bằng cách lập lịch đĩa
- Seek time ≈ seek distance
- Dải thông đĩa (Disk bandwidth) tính bằng *tổng số byte được chuyển* chia cho *tổng thời gian giữa lần chuyển đầu tiên và lần chuyển cuối cùng*.
  - Seek time tốt hơn với mỗi yêu cầu sẽ cải thiện bandwidth.

2. Lập lịch đĩa - Disk scheduling (2)

- Khi tiến trình cần thực hiện vào-ra với đĩa, nó phát 1 system call tới HĐH, HĐH cần xác định:
  - thao tác là input hay output
  - địa chỉ đĩa và địa chỉ bộ nhớ (nguồn và đích)
  - số byte cần chuyển
- Nếu ổ đĩa và mạch điều khiển sẵn sàng, yêu cầu có thể được thực hiện ngay. Trái lại, nó được đưa vào queue của đĩa để chờ được phục vụ.
- Có một số giải thuật lập lịch sự phục vụ các yêu cầu vào-ra đĩa cho một thứ tự tốt.
- Chúng ta minh họa chúng với một request queue (0-199). 98, 183, 37, 122, 14, 124, 65, 67
  - Đầu từ đĩa đang ở cylinder 53

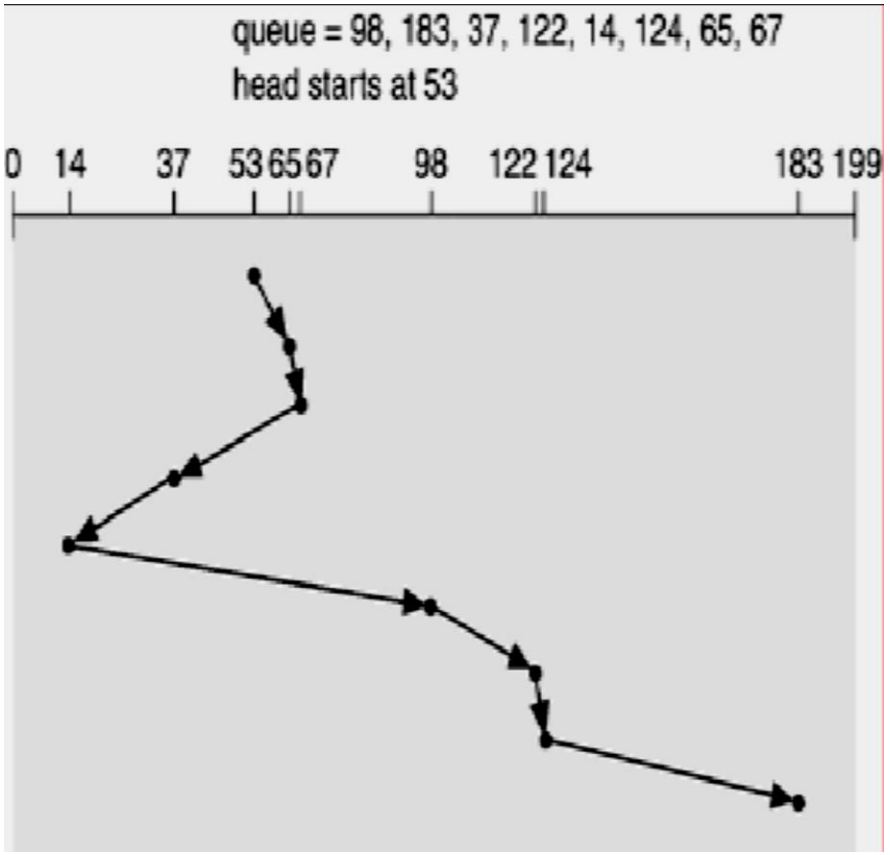
2.1. FCFS – Fist Come, First Served

- Tổng quãng đường di chuyển của đầu từ là 640 cylinder.



2.2. SSTF – Shortest Seek Time First

- Chọn yêu cầu với seek(định vị, tìm) time nhỏ nhất từ vị trí đầu từ hiện thời.
- Tổng quãng đường di chuyển của đầu từ là 236 cylinder.

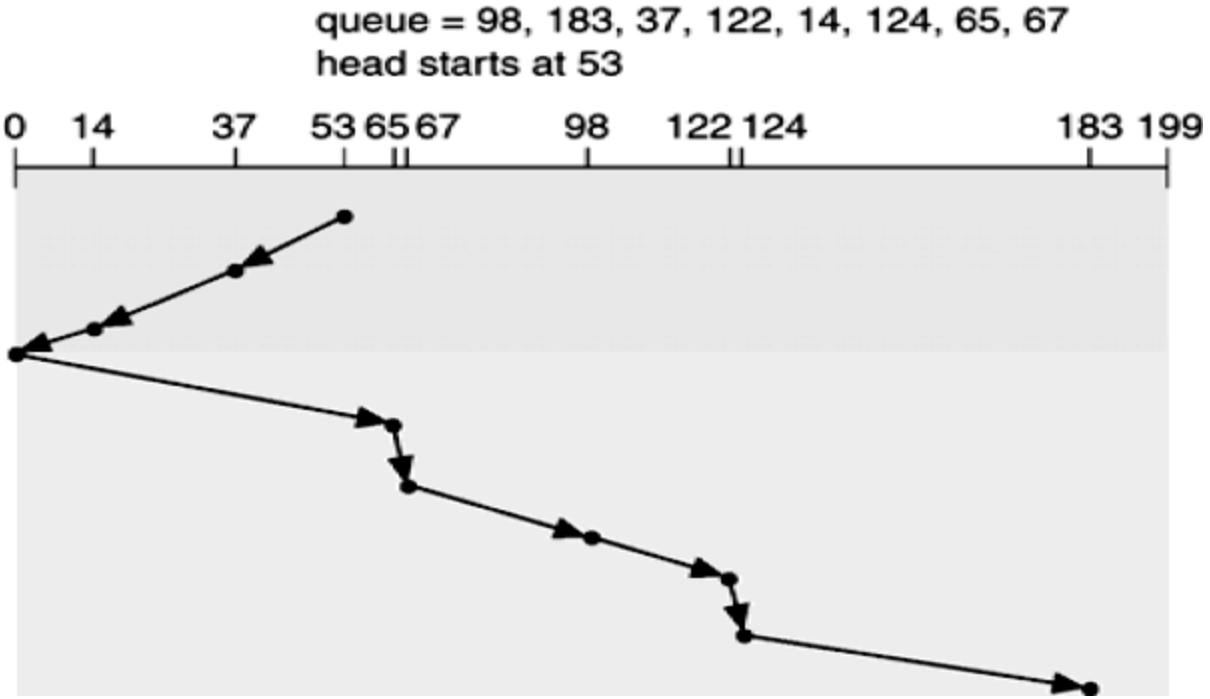


2.3. SCAN(1)

- Disk arm bắt đầu tại một đầu của đĩa, tiến dần tới đầu còn lại, phục vụ yêu cầu khi nó đến mỗi cylinder, tại đầu còn lại hướng di chuyển của đầu từ sẽ đảo ngược và việc phục vụ tiếp tục.
- Cần biết thêm hướng di chuyển của đầu từ
- Còn gọi là giải thuật thang máy - *elevator algorithm*.

2.3. SCAN(2)

- Tổng quãng đường di chuyển của đầu từ là 236 cylinder.
- Tổng quãng đường di chuyển của đầu từ là bao nhiêu nếu nó di chuyển theo hướng ngược lại?

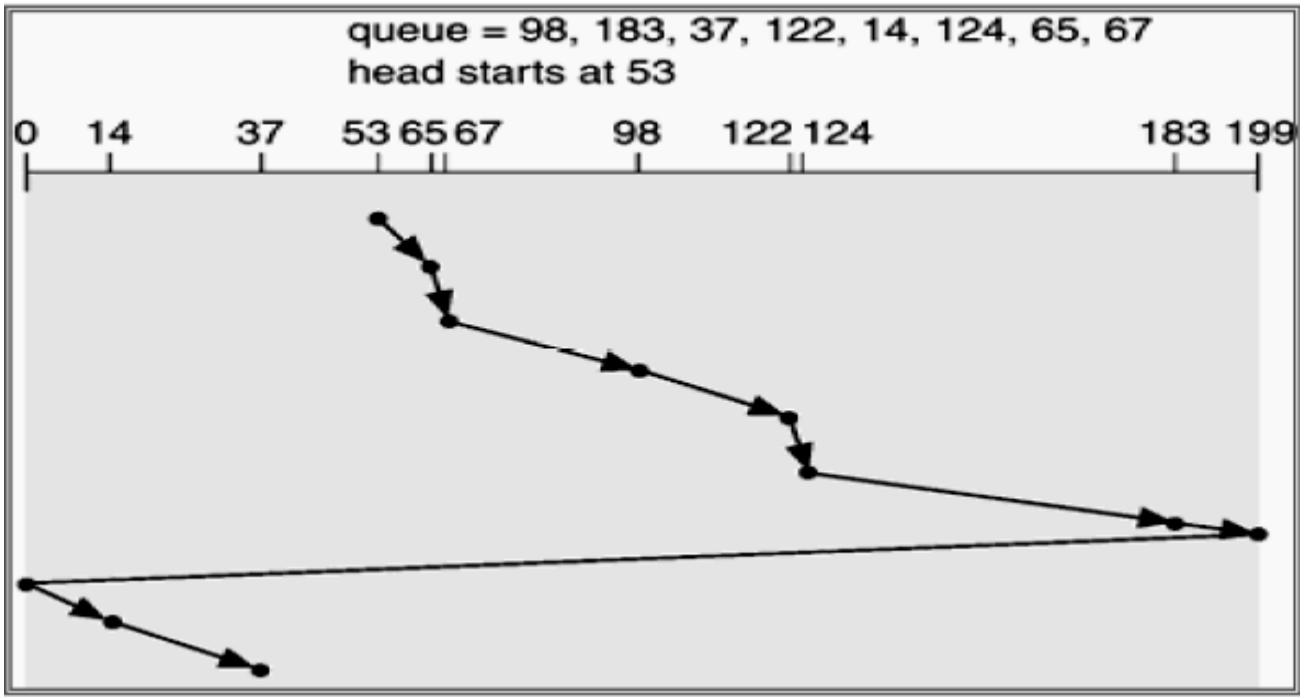


2.4. C-SCAN - Circular SCAN(1)

- Tương tự như SCAN, nhưng có thời gian chờ đồng đều hơn so với SCAN.
- Đầu từ di chuyển từ một đầu đĩa tới đầu còn lại, phục vụ yêu cầu khi nó đến. Tuy nhiên, khi nó đến đầu kia thì lập tức quay về điểm đầu đĩa mà không phục vụ yêu cầu nào trên hành trình quay về đó.

2.4. C-SCAN - Circular SCAN(2)

- Tổng quãng đường di chuyển của đầu từ là 382 cylinder.
- Tổng quãng đường di chuyển của đầu từ là bao nhiêu nếu nó di chuyển theo hướng ngược lại?

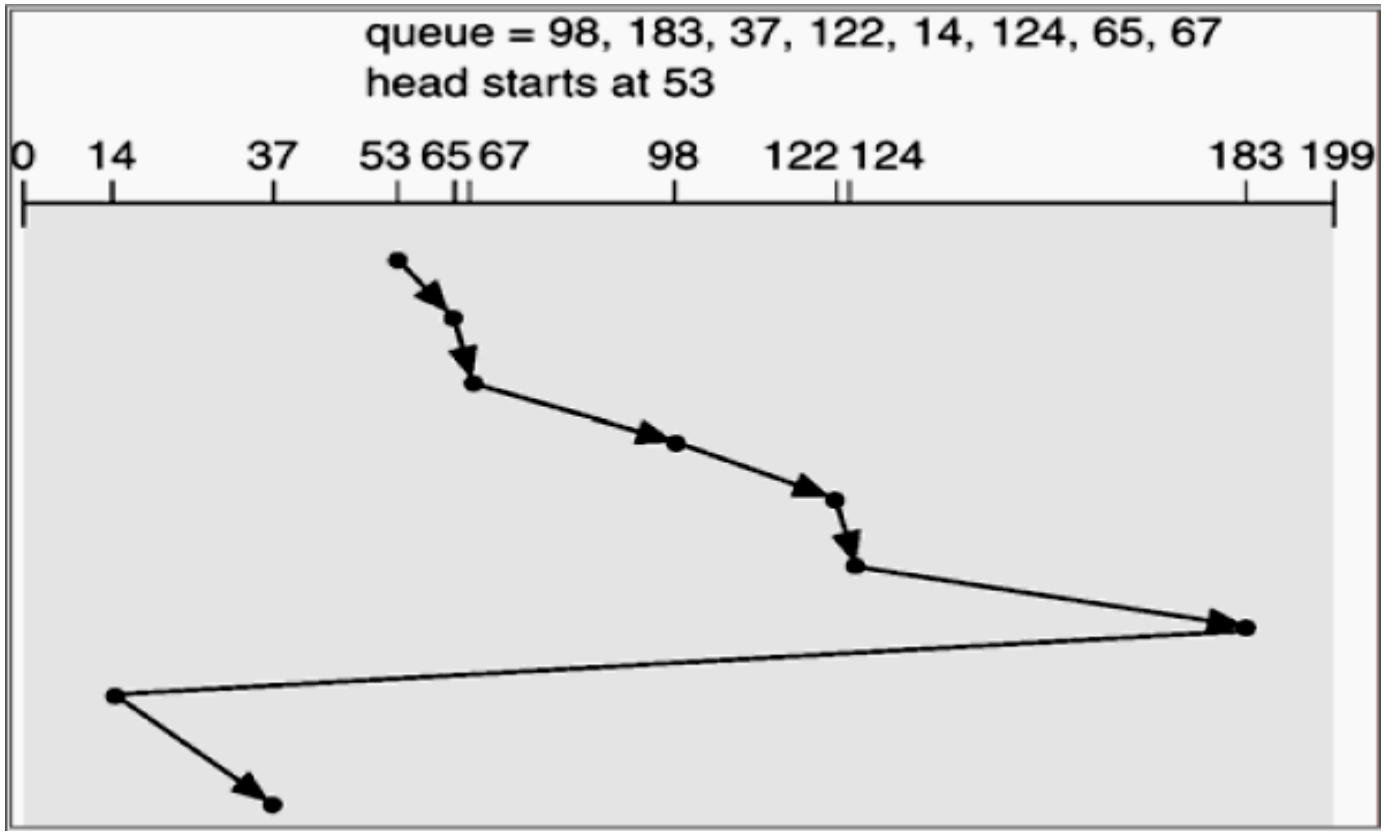


2.5. LOOK và C-LOOK(1)

- Là phiên bản tương ứng của SCAN và C-SCAN
- Arm chỉ đi đến yêu cầu cuối cùng trên mỗi hướng rồi lập tức đảo hướng mà không đi tất cả quãng đường lãng phí đến tận cùng đĩa.
- Gọi là LOOK vì nó tìm kiếm một yêu cầu trước khi tiếp tục di chuyển trên hướng đi.

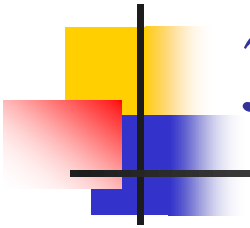
2.5. LOOK và C-LOOK(2): C-LOOK

- Tổng quãng đường di chuyển của đầu từ là 322 cylinder.



2.6. Lựa chọn một giải thuật lập lịch đĩa

- SSTF phổ biến và có hiệu quả tốt.
- SCAN và C-SCAN thực hiện tốt hơn đối với các hệ thống đặt tải lớn lên đĩa.
- Hiệu năng phụ thuộc vào số lượng và loại yêu cầu.
- Các yêu cầu phục vụ đĩa có thể bị ảnh hưởng bởi phương thức phân phối file.
- Giải thuật lập lịch đĩa có thể được viết như một module riêng của HĐH, cho phép thay thế khi cần thiết.
- Giải thuật SSTF hoặc LOOK là một lựa chọn hợp lý làm mặc định.



3. Quản lý đĩa

- Disk Formatting
- Boot Block
- Bố trí đĩa của MS-DOS
- Bad Blocks

3.1. Disk Formatting(1)

- *Low-level formatting*, hay *physical formatting* – chia một đĩa thành các sector (cung) để mạch điều khiển đĩa có thể đọc và ghi.
- Cấu trúc dữ liệu của mỗi sector:
  - 1 header và 1 trailer: chứa thông tin được sử dụng bởi disk controller, ví dụ số hiệu sector, mã sửa lỗi (ECC: errorcorrecting code)
  - Vùng chứa dữ liệu: thường 512 byte; hoặc 256 hay 1024 byte.

3.1. Disk Formatting(2)

- Để sử dụng đĩa lưu giữ file, HĐH vẫn cần phải ghi cấu trúc dữ liệu của chính nó trên đĩa.
  - 1. Phân chia đĩa thành một hay nhiều nhóm các cylinder, được gọi là Partition. HĐH có thể xử lý mỗi Partition như một đĩa độc lập.
  - *2. Logical formatting* hay "making a file system": HĐH ghi lên đĩa cấu trúc dữ liệu hệ thống file ban đầu, có thể gồm các thư mục rỗng ban đầu, bản đồ không gian bộ nhớ tự do và đã sử dụng (bảng FAT – File Allocation Table).

3.2. Boot Block(1)

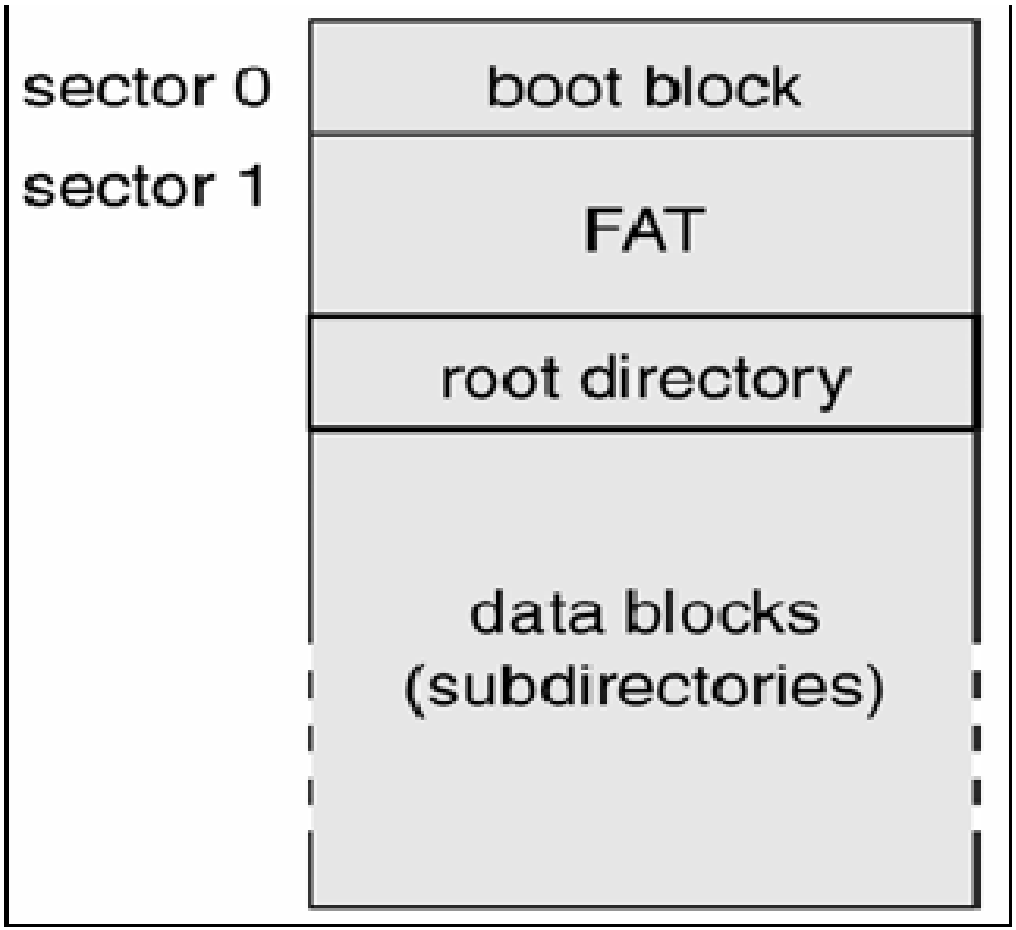
- Để MT bắt đầu chạy (khi bật máy, khi khởi động lại), cần có một chương trình khởi tạo để chạy: chương trình mồi – bootstrap.
- Bootstrap khởi động tất cả các bộ phận của máy tính, từ các thanh ghi trong CPU đến các mạch điều khiển thiết bị và nội dung của bộ nhớ chính, sau đó bắt đầu chạy HĐH.
- Để thực hiện công việc của mình, chương trình bootstrap tìm nhân (kernel) của HĐH trên đĩa để nạp vào bộ nhớ, rồi nhảy đến một địa chỉ bắt đầu sự thực hiện của HĐH

3.2. Boot Block(2)

- Bootstrap được lưu trong ROM:
  - ROM không cần khởi tạo,
  - Ở tại vị trí cố định processor có thể bắt đầu thực hiện khi được khởi động
  - ROM không bị ảnh hưởng bởi virus máy tính
- Hầu hết HĐH chỉ chứa chương trình mồi rất nhỏ trong boot ROM, giúp cho việc nạp chương trình mồi đầy đủ từ đĩa.

  -> Chương trình mồi đầy đủ có thể được thay đổi dễ dàng
- Chương trình mồi đầy đủ được chứa trong một partition gọi là boot blocks, tại một vị trí cố định trên đĩa.
- Đĩa có một boot partition được gọi là boot disk hay system disk

3.3. Bố trí đĩa của MS-DOS



3.4. Bad Blocks(1)

- Trên các đĩa đơn giản, vd đĩa IDE, các bad block được xử lý thủ công bằng lệnh **format** của MS-DOS:
  - Thực hiện format logic, quét đĩa để tìm các bad block.
  - Nếu tìm thấy bad block, một giá trị đặc biệt được ghi vào phần tử tương ứng trong bảng FAT để báo cho các thường trình phân phối (allocation routine) không sử dụng block đó nữa.
  - Nếu các block trở thành bad trong khi hoạt động bình thường, có thể chạy một chương trình đặc biệt như **chkdsk** để tìm các bad block và xử lý chúng như ở trên.
  - Dữ liệu trên các bad block thường bị mất

3.4. Bad Blocks(2)

- Trên các đĩa phức tạp, vd đĩa SCSI, việc phục hồi bad block thông minh hơn:
  - Mạch điều khiển duy trì một danh sách các bad block trên đĩa. DS này được khởi tạo khi format cấp thấp tại nhà máy và được cập nhật trong suốt sự tồn tại của đĩa.
  - Format cấp thấp cũng thiết lập các sector dự phòng (spare sector) vô hình đối với HĐH. Mạch điều khiển có thể thay thế mỗi bad sector một cách logic bởi một trong số các sector dự phòng.

    -> **Sector sparing** (kỹ thuật dự phòng sector) hay **forwarding**
    - HĐH cố gắng đọc block 87. Mạch điều khiển (controller) tính toán ECC và thấy sector đó là bad. Nó thông báo cho HĐH biết.
    - Ở lần khởi động tiếp theo, một lệnh đặc biệt được chạy để ra lệnh cho Mạch điều khiển SCSI thay thế bad sector, ví dụ bởi sector 202.
    - Sau đó, mỗi khi hệ thống yêu cầu block 87, mạch điều khiển sẽ thông dịch yêu cầu sang địa chỉ của sector 202.

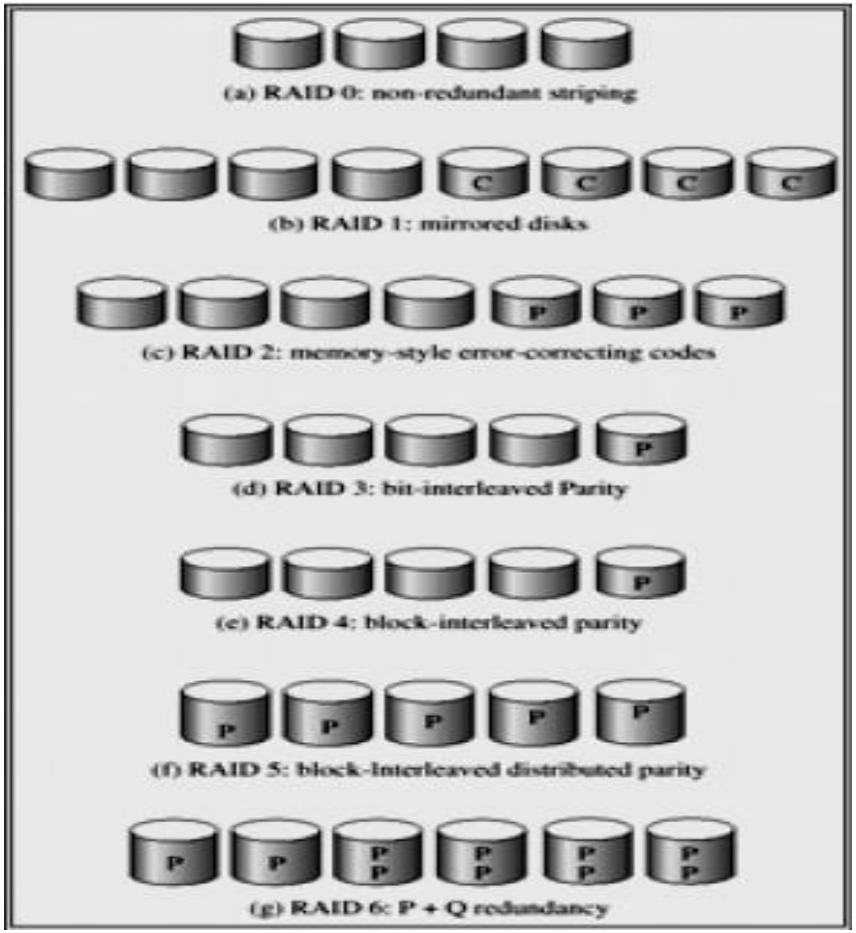
4. Quản lý không gian hoán đổi (Swap-Space)

- Swap-space: Bộ nhớ ảo sử dụng không gian đĩa như là sự mở rộng của bộ nhớ chính -> tăng dung lượng, tăng tốc độ.
- Swap-space có thể được tạo ra:
  - Từ hệ thống file bình thường: swap-space là một file lớn do thường trình hệ thống file tạo, đặt tên và phân phối bộ nhớ.
    - Phương pháp này dễ thực hiện nhưng không hiệu quả: định vị cấu trúc thư mục và cấu trúc dữ liệu trên đĩa tốn nhiều thời gian và truy nhập đĩa nhiều lần hơn.
  - Trong 1 partition riêng (phổ biến hơn): không có cấu trúc thư mục và file trên đó,
    - 1 trình quản lý bộ nhớ hoán đổi riêng điều khiển việc phân phối và thu hồi các block. Nó sử dụng các giải thuật để tối ưu tốc độ hơn là để lưu trữ hiệu quả.

5. Cấu trúc RAID(1)

- **RAID** – Redundant Arrays of Inexpensive/Independent Disk
- Sử dụng nhiều đĩa như một đơn vị lưu trữ
- Cải thiện hiệu năng và độ tin cậy của hệ thống lưu trữ bằng cách lưu trữ các dữ liệu dư thừa.
- RAID được phân cấp thành 7 mức:
  - RAID mức 0: dữ liệu được phân ra nhiều ổ đĩa nhưng không có ổ dự phòng.
  - RAID mức 1: dữ liệu được phân vào 1 dãy những ổ đĩa và dữ liệu trong mỗi ổ lại được chuyển vào một ổ đĩa lưu trữ để dự phòng.
  - RAID mức 2, 3, 4: dữ liệu được chia cho nhiều đĩa và thông tin parity được phát sinh và ghi vào 1 đĩa riêng biệt. Mỗi cấp độ có những phương pháp khác nhau khi ghi dữ liệu lên đĩa.
  - RAID mức 5: dữ liệu và các đoạn mã parity được ghi lên tất cả các ổ đĩa trong dãy các ổ đĩa. Mức này cho phép ghi nhanh vì thông tin parity được lan ra tất cả các ổ đĩa mà không phải là ghi lên từng ổ riêng biệt.

5. Cấu trúc RAID(2): Các mức

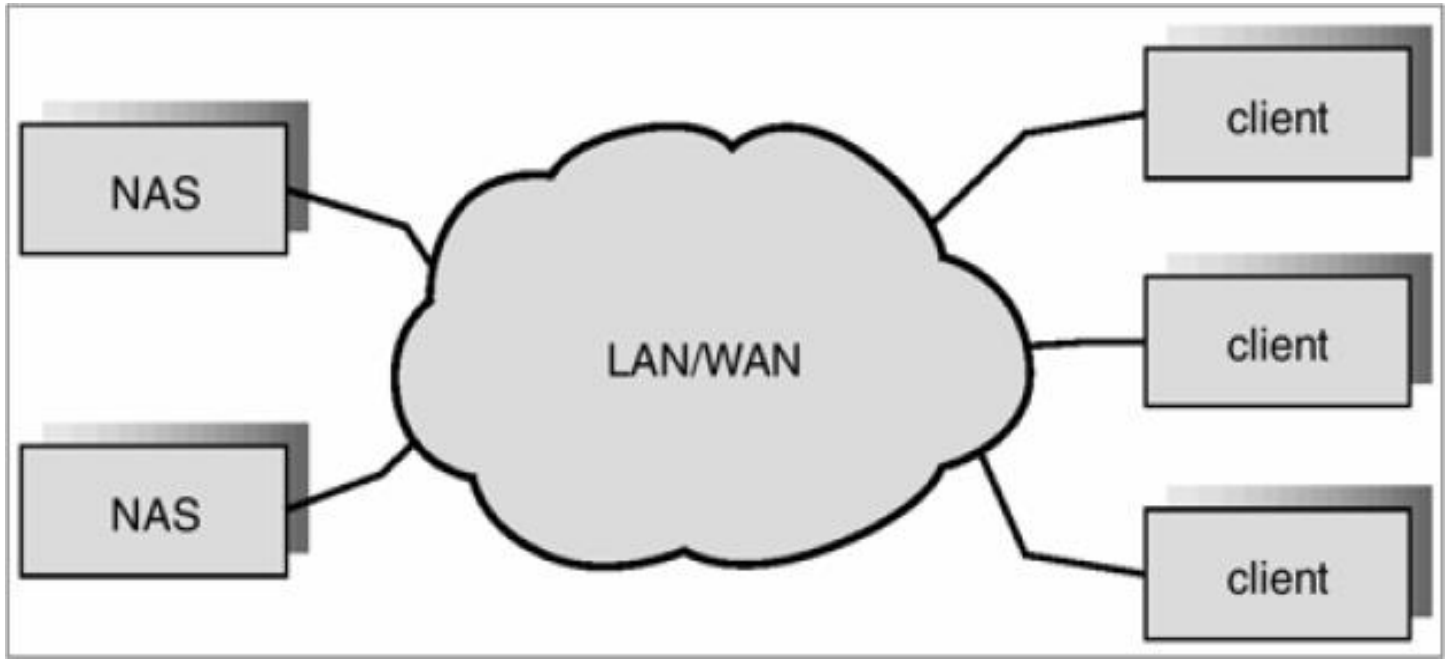


6. Nối kết đĩa(1)

- Các đĩa có thể được nối kết theo 1 trong 2 cách:
- **1. Host-attached storage:** nối kết thông qua một cổng vàora, sử dụng một số kỹ thuật:
  - Kiến trúc I/O bus (IDE-Integrated Drive Electronics, ATAAdvanced Technology Attachment): hỗ trợ tối đa 2 ổ đĩa trên mỗi I/O bus, sử dụng trong các máy PC.
  - Kiến trúc SCSI (Small Computer System Interface): hỗ trợ tối đa 16 thiết bị/1 bus (1 card điều khiển, 15 thiết bị lưu trữ)
  - Kiến trúc Fibre Channel (FC): sử dụng cáp quang hoặc cáp đồng, có thể kết nối hàng triệu thiết bị (2^{24}) trên mạng.

6. Nối kết đĩa(2)

- **2. Network-attached storage (NAS):** nối kết các thiết bị nhớ thông qua một kết nối mạng



7. Cấu trúc lưu trữ cấp ba
7.1. Các thiết bị nhớ cấp ba

- Giá thành rẻ là đặc điểm nổi bật của bộ nhớ cấp ba.
- Nói chung, bộ nhớ cấp ba gồm các thiết bị khả chuyển (*removable media*)
- Ví dụ phổ biến: đĩa mềm, đĩa CD, flash disk.

7.1.1. Removable Disks(1)

- Floppy disk – đĩa mềm dẻo mỏng được phủ lớp từ và được bảo vệ bên ngoài bởi một vỏ bằng chất dẻo.
  - Có dung lượng khoảng 1.4 MB;
- Removable magnetic disk được tạo bằng kỹ thuật tương tự
  - Có dung lượng hơn 1 GB.
  - Có tốc độ nhanh gần như hard disks, nhưng rất dễ bị hỏng.

7.1.1. Removable Disks(2)

- Đĩa từ-quang (magneto-optic disk) ghi dữ liệu trên một mặt đĩa cứng được phủ lớp từ.
  - Ổ đĩa có 1 cuộn dây sinh từ trường. Tại nhiệt độ thường, từ trường quá rộng và yếu nên không thể từ hóa các bit trên đĩa → Sử dụng phương pháp đốt Laser: đầu đĩa chiếu 1 tia laser lên mặt đĩa để ghi một bit.
  - Ánh sáng Laser cũng được sử dụng để đọc dữ liệu.
  - Đầu đọc đĩa từ-quang bay xa mặt đĩa hơn nhiều so với đầu đọc đĩa từ, và lớp từ được phủ một lớp bảo vệ dầy bằng kính hoặc chất dẻo để chống sự phá hủy của đầu đọc.

7.1.1. Removable Disks(3)

- Các đĩa quang (optical disk) không sử dụng hiện tượng từ tính; chúng sử dụng các chất liệu đặc biệt có thể bị biến đổi bằng ánh sáng laser thành các điểm sáng và tối.
- Vd: đĩa đổi pha:
  - Đĩa được phủ chất liệu có thể đông cứng thành trạng thái kết tinh hoặc vô định hình.
  - Trạng thái kết tinh trong suốt hơn nên tia laser sáng hơn khi đi qua chất liệu đổi pha và bật ra khỏi lớp phản chiếu.
  - Ổ đĩa đổi pha sử dụng ánh sáng laser ở 3 cường độ: cường độ yếu để đọc dữ liệu, cường độ trung bình để xóa đĩa (làm tan rồi đông cứng lại), cường độ mạnh làm tan chất liệu thành tr.thái vô định hình để ghi dữ liệu.
  - VD: re-recordable CD-RW và DVD-RW

7.1.1. Removable Disks(4)

- Các loại trên là read-write disk: dữ liệu có thể sửa đổi nhiều lần.
- WORM Disks
  - Các WORM disk ("Write Once, Read Many Times") được ghi chỉ 1 lần.
  - Màng nhôm mỏng được xen giữa 2 mặt đĩa bằng kính hay chất dẻo.
  - Để ghi 1 bit, ổ đĩa sử dụng ánh sáng laser để đốt một hố nhỏ qua lớp nhôm; thông tin có thể bị phá hủy vì không thể thay đổi.
  - Rất bền và đáng tin cậy.
  - Vd: CD-R, DVD-R (Digital Video/Versatile Disk – Recordable)
- *Read Only* disks, ví dụ CD-ROM và DVD, sử dụng kỹ thuật tương tự (hố được ép chặt, không phải bị đốt).

7.1.2. Tapes

- So với đĩa, băng rẻ hơn và lưu chứa nhiều dữ liệu hơn, nhưng sự truy nhập ngẫu nhiên chậm hơn nhiều.
- Băng là một giải pháp tiết kiệm cho những mục đích không yêu cầu truy nhập nhanh, vd:
  - lưu trữ bản sao của dữ liệu trên đĩa.
  - sử dụng trong những trung tâm siêu máy tính lớn để lưu trữ lượng thông tin khổng lồ phục vụ nghiên cứu khoa học.

7.1.3. Công nghệ tương lai?

- Công nghệ lưu trữ ảnh giao thoa laser (Holographic storage):
  - Sử dụng ánh sáng laser để ghi các bức ảnh giao thoa laser (holographic photograph) trên các thiết bị đặc biệt.
  - Các bức ảnh đen trắng là một mảng 2 chiều các pixel, mỗi pixel biểu diễn 1 bit: 0-đen; 1-trắng
  - Một bức ảnh có thể lưu hàng triệu bit dữ liệu
  - Tất cả các pixel được truyền rất nhanh với tốc độ ánh sáng laser → tốc độ truyền dữ liệu rất cao
  - Là công nghệ lưu trữ đầy hứa hẹn của tương lai.
- Hệ cơ khí vi điện tử (Micro electronic mechanical system - MEMS): chế tạo các chip điện tử để sản xuất các thiết bị lưu trữ nhỏ, nếu thành công sẽ cung cấp công nghệ lưu trữ dữ liệu không khả biến, nhanh hơn đĩa từ và rẻ hơn DRAM bán dẫn.

7.2. Các công việc của HĐH
7.2.1. Giao diện ứng dụng

- Hầu hết các HĐH quản lý các đĩa khả chuyển giống như các đĩa cố định – một đĩa mới phải được format và tạo một hệ thống file rỗng trên đó.
- Các thao tác cơ bản với ổ đĩa: read, write, seek.
- Các băng được coi là phương tiện lưu trữ thô, vd: ứng dụng không mở 1 file trên băng mà mở toàn bộ ổ băng.
- Các thao tác cơ bản với tape:
  - **locate**: định vị băng vào 1 khối (block) xác định
  - **read position**: trả về vị trí hiện thời của đầu băng (block number)
  - Các ổ băng là các thiết bị "chỉ ghi thêm“ (append-only).
  - Một dấu EOT (End Of Tape) được đặt sau mỗi khối vừa được ghi.

7.2.2. Đặt tên file

- Vấn đề đặt tên các file trên các thiết bị nhớ khả chuyển là đặc biệt khó khi chúng ta muốn ghi dữ liệu lên nó trên một máy tính và rồi sử dụng nó trên một máy tính khác:
  - Tìm đường dẫn đến file, kiểu ổ đĩa có tương thích?
  - Thứ tự lưu trữ các byte dữ liệu khác nhau, vd:
    - các bộ VXL Intel 80x86 và Pentium: lưu trữ kiểu little-endian
    - các bộ VXL Motorola 680x0 và RISC: lưu trữ kiểu big-endian
- Các HĐH hiện nay nói chung để mặc vấn đề trên, phụ thuộc vào các ứng dụng và người sử dụng để tìm ra cách truy nhập, hiển thị dữ liệu.
- Một số thiết bị nhớ khả chuyển (CD, DVD) được tiêu chuẩn hóa tốt để tất cả các máy tính sử dụng chúng theo một cách chung.

7.3. Hiệu năng
7.3.1. Tốc độ(1)

- Hai mặt của tốc độ trong bộ nhớ cấp ba là dải thông (bandwidth) và trễ truy nhập (latency).
- Dải thông (số byte/giây).
  - Dải thông liên tục (Sustained bandwidth) – tốc độ dữ liệu trung bình trong suốt quá trình truyền lớn; được tính bằng số byte/thời gian truyền.

    Là tốc độ dữ liệu khi dòng dữ liệu truyền thực sự
  - Effective bandwidth (Dải thông có hiệu lực) – tính trung bình trên toàn bộ thời gian vào-ra.

    Là tốc độ dữ liệu tổng thể của ổ đĩa.

  Dải thông của ổ đĩa thường được hiểu là sustained bandwith

  Các removable disk: 0.25 - 5 MB/s; Các tape: 0.25 - 30 MB/s

7.3.1. Tốc độ(2)

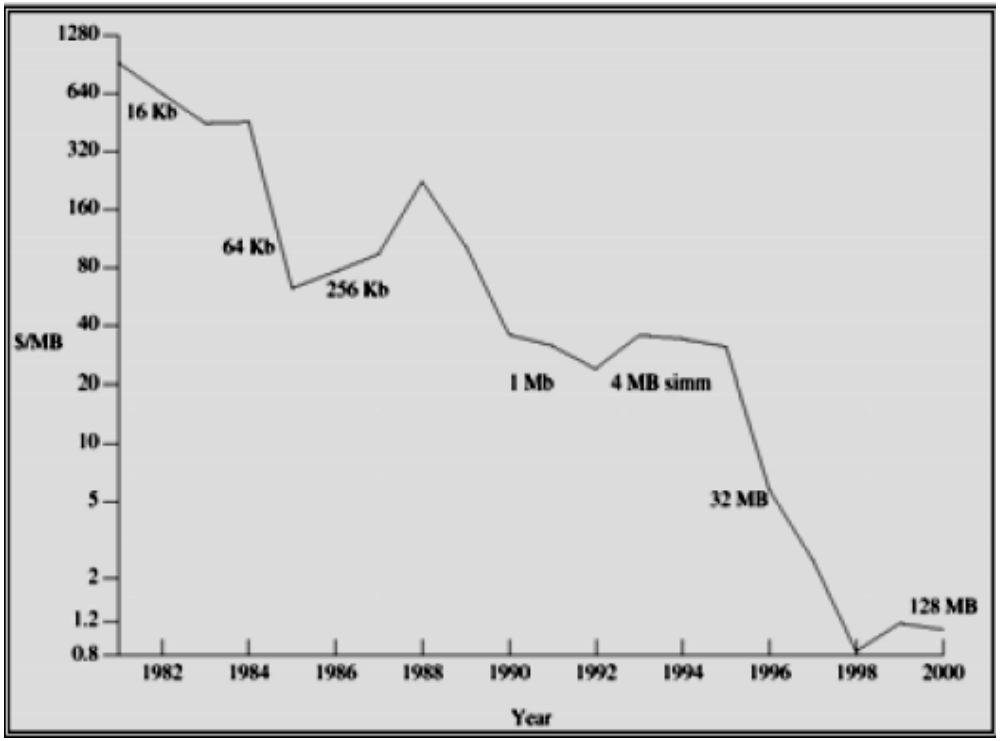
- Trễ truy nhập – khoảng thời gian cần thiết để định vị dữ liệu.
  - Thời gian truy nhập của ổ đĩa – dịch arm tới cylinder được chọn và đợi trễ quay (rotational latency); < 35 ms.
  - Truy nhập trên băng đòi hỏi phải cuộn ống băng cho đến khi khối được chọn chạm đến đầu băng; tốn hàng chục hoặc hàng trăm giây.
  - Nói chung truy nhập ngẫu nhiên trên băng chậm hơn trên đĩa hàng nghìn lần.

7.3.2. Độ tin cậy

- Một ổ đĩa cố định có vẻ đáng tin cậy hơn một ổ đĩa hay ổ băng khả chuyển.
- Một đĩa quang có vẻ đáng tin cậy hơn một đĩa từ hay băng từ.
- Sự rơi đầu từ trong một đĩa cứng cố định thường phá hủy dữ liệu, trong khi đó hỏng ổ băng hay ổ đĩa quang thường không làm hỏng đến dữ liệu.

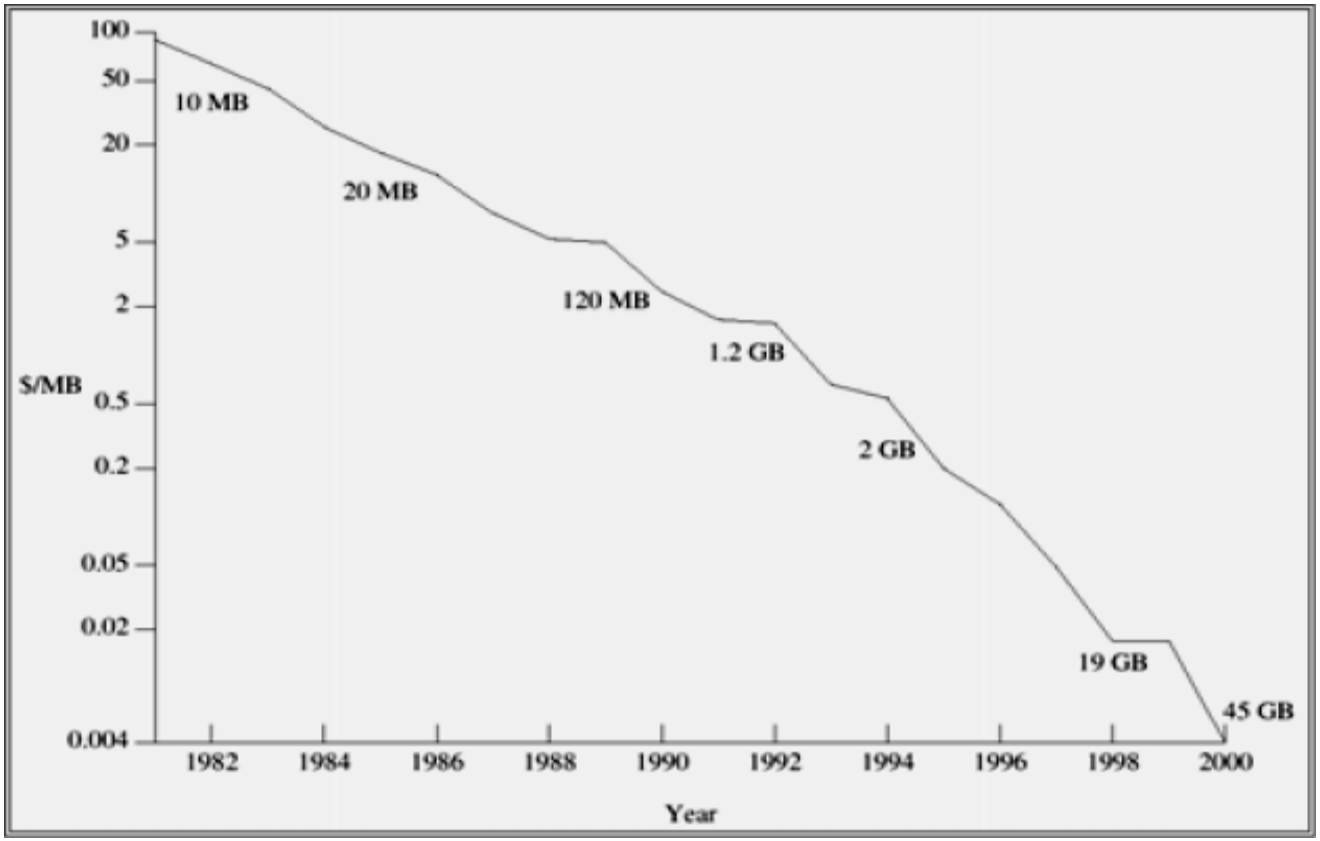
7.3.3. Giá thành (USD/1 MB)(1)

- Bộ nhớ chính (Main memory) đắt hơn nhiều so với bộ nhớ cấp hai và cấp ba.
- Giá mỗi MB DRAM, 1981-2000



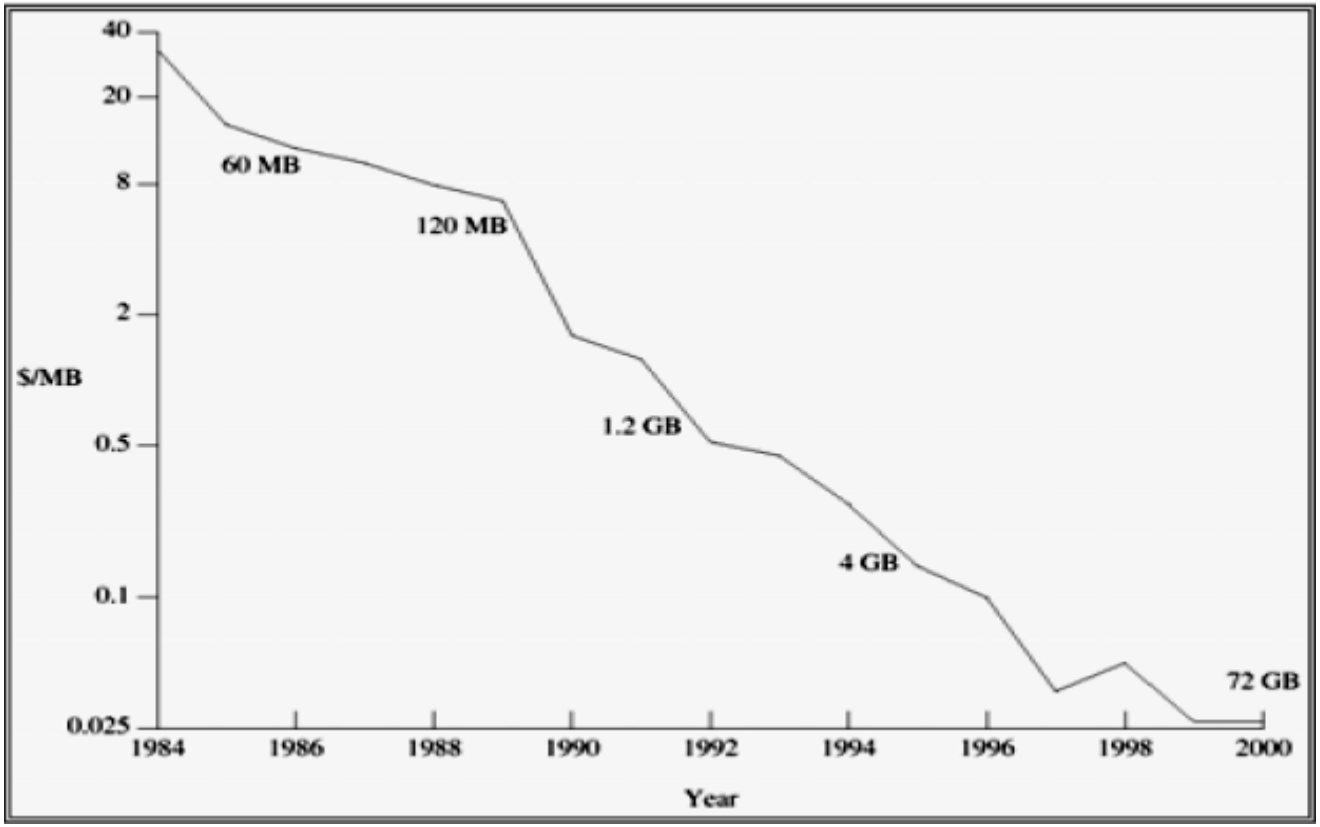
7.3.3. Giá thành (USD/1 MB)(2)

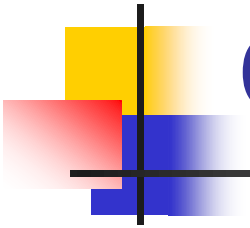
- Giá mỗi MB đĩa từ cứng, 1981-2000



7.3.3. Giá thành (USD/1 MB)(3)

- Giá mỗi MB ổ băng, 1984-2000





Q & A

- List câu hỏi