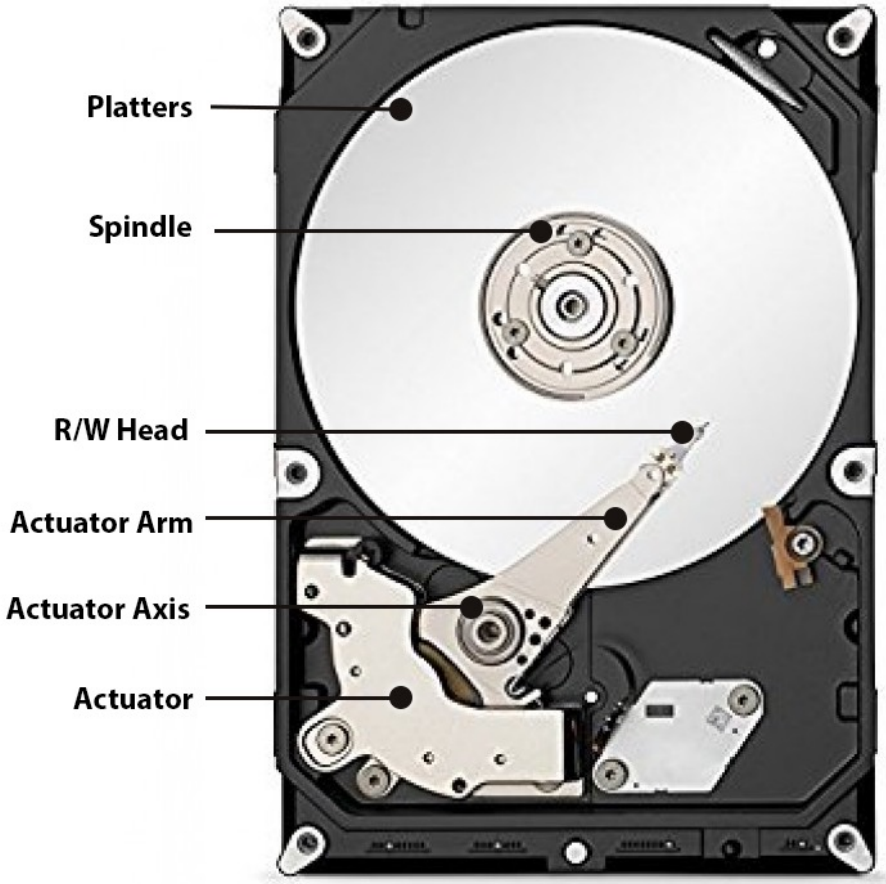
Bài 9
Mass-Storage Structure (Cấu trúc lưu trữ)

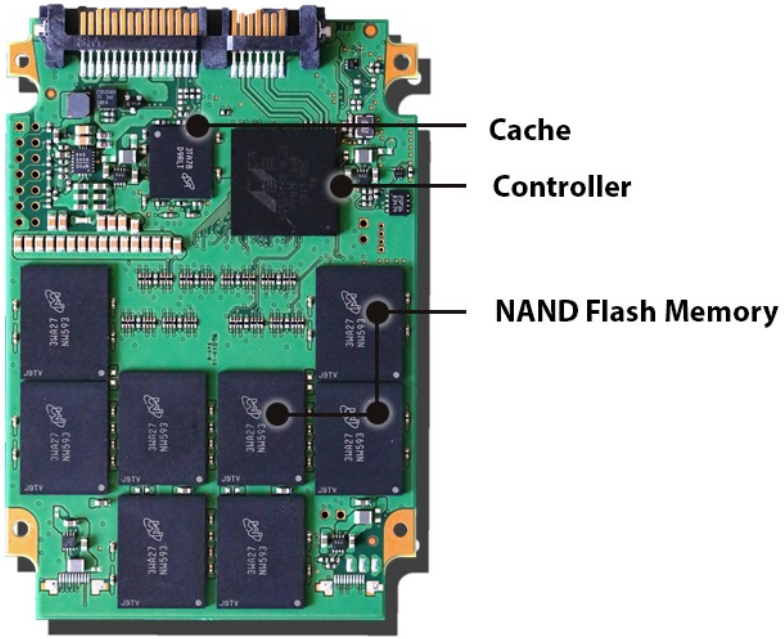
Tổng quan về cấu trúc lưu trữ

- Bộ nhớ thứ cấp gồm 2 loại phổ biến: Ổ đĩa cứng (HDD – Hard disk drive) và bộ nhớ bất biến (NVM – Nonvolatile).

HDD
3.5″



SSD
2.5″

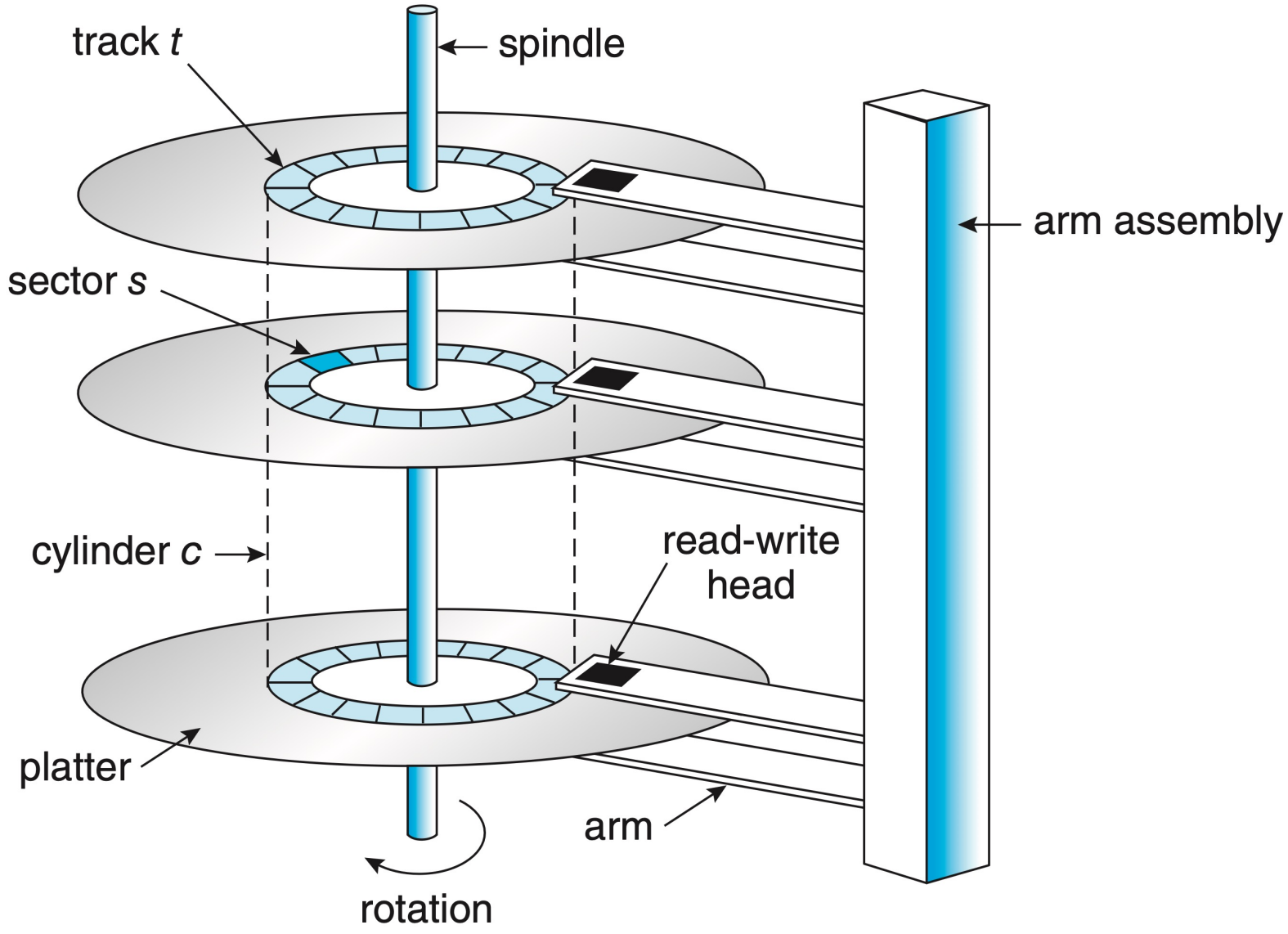


Hard disk drive (HDD)

- Cấu tạo gồm các phiến đĩa được phủ lớp vật liệu từ, xoay quanh một trục.
- Ổ đĩa quay với tốc độ 60 → 250 lần/giây, tương ứng 5400 → 15000 RPM.
- Transfer rate (tốc độ truyền): tốc độ dữ liệu chuyển giữa ổ đĩa và máy tính.
- Positioning time (thời gian định vị): là thời gian di chuyển đầu đọc đến đúng cylinder (seek time) + thời gian tìm đến đúng sector mong muốn (rotational latency)
- Head crash: sự cố xảy ra khi đầu đọc đĩa tiếp xúc với bề mặt đĩa.
- Đĩa cứng có thể tháo rời (đặc biệt đối với máy chủ).

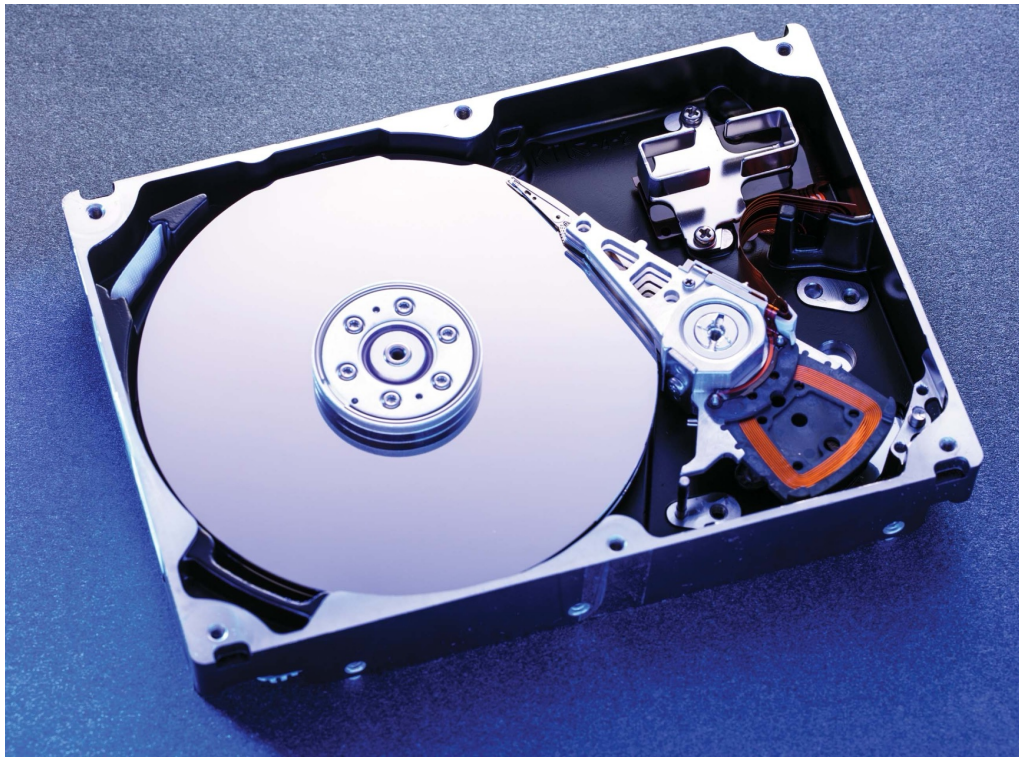
Hard disk drive (HDD)

- Các khái niệm:
  - Platter
  - Track
  - Cylinder
  - Sector (thường có kích thước 4KB)



Hard disk drive (HDD)

- Kích thước đĩa: phổ biến từ 3.5, 2.5 và 1.8 inch.
- Dung lượng từ vài chục GB → TB
- Hiệu suất:
  - Tốc độ truyền tải (lý thuyết): 6Gb/giây
  - Tốc độ truyền tải (thực tế): 1GB/giây
  - Seek time: 3ms → 12ms



Nonvolatile Memory (NVM)

- Có tên gọi khác là ổ cứng thể rắn (Solid-state-disk – SSD).
- Một số đặc điểm so với HDD:
  - Đáng tin cậy hơn.
  - Đắt hơn (tính trên MB lưu trữ)
  - Đôi khi có tuổi thọ thấp hơn.
  - Dung lượng thường thấp hơn, nhưng tốc độ cao hơn.
  - Một số loại sẽ kết nối trực tiếp vào cổng PCI để tăng tốc độ.
  - Không có bộ phận cơ học, do đó giảm thiểu seek time và rotational latency.
  - Tuổi thọ phụ thuộc số lần đọc/ghi.

Volatile Memory

- Phân biệt Nonvolatile và Volatile memory?
- Đặc điểm volatile memory: dữ liệu bị mất khi mất nguồn điện.
- Một số loại bộ nhớ volatile: RAM, cache (L1, L2, L3)
- Ưu điểm: tốc độ truy xuất dữ liệu cực nhanh.

Magnetic tape (băng từ)

- Dung lượng lưu trữ dữ liệu lớn, lên đến hàng chục TB.
- Độ tin cậy rất cao, tỉ lệ xảy ra lỗi thấp hơn ổ cứng.
- Chi phí lưu trữ thấp, không cần năng lượng.
- Tốc độ đọc/ghi thấp.
- Dữ liệu trên băng từ có thể được nén để tiết kiệm bộ nhớ.
- An toàn
- Ai đang sử dụng băng từ: Amazon, Google, Meta, Baidu, Alibaba, Tencent,….

Cấu trúc đĩa

- CHS (Cylinder – Head – Sector) là phương pháp truyền thống, dùng định vị các khối dữ liệu trên đĩa cứng, dựa vào thứ tự cylinder, head và sector.
- LBA (Logical Block Addressing) là phương pháp định vị các khối dữ liệu tốt hơn:
  - Mỗi khối dữ liệu trên đĩa cứng được gọi là một logical block và có một địa chỉ duy nhất, đây là đơn vị nhỏ nhất để truyền dữ liệu.
  - Mỗi logical block được gắn kết đến một sector vật lý trên đĩa.
  - Sector 0 là sector đầu tiên của track đầu tiên thuộc cylinder ngoài cùng.
  - Sử dụng LBA giúp đơn giản hóa quá trình định vị dữ liệu và giảm sự phức tạp khi quản lý các địa chỉ vật lý của đĩa cứng.
  - Nếu đĩa cứng có kích thước 1TB và kích thước mỗi sector là 4KB thì có tổng cộng 1TB/4KB khối dữ liệu, mỗi khối sẽ có một địa chỉ LBA riêng biệt.
  - Đối với NVM, logical block được gắn kết đến bộ chip, block và page.

Lập lịch HDD

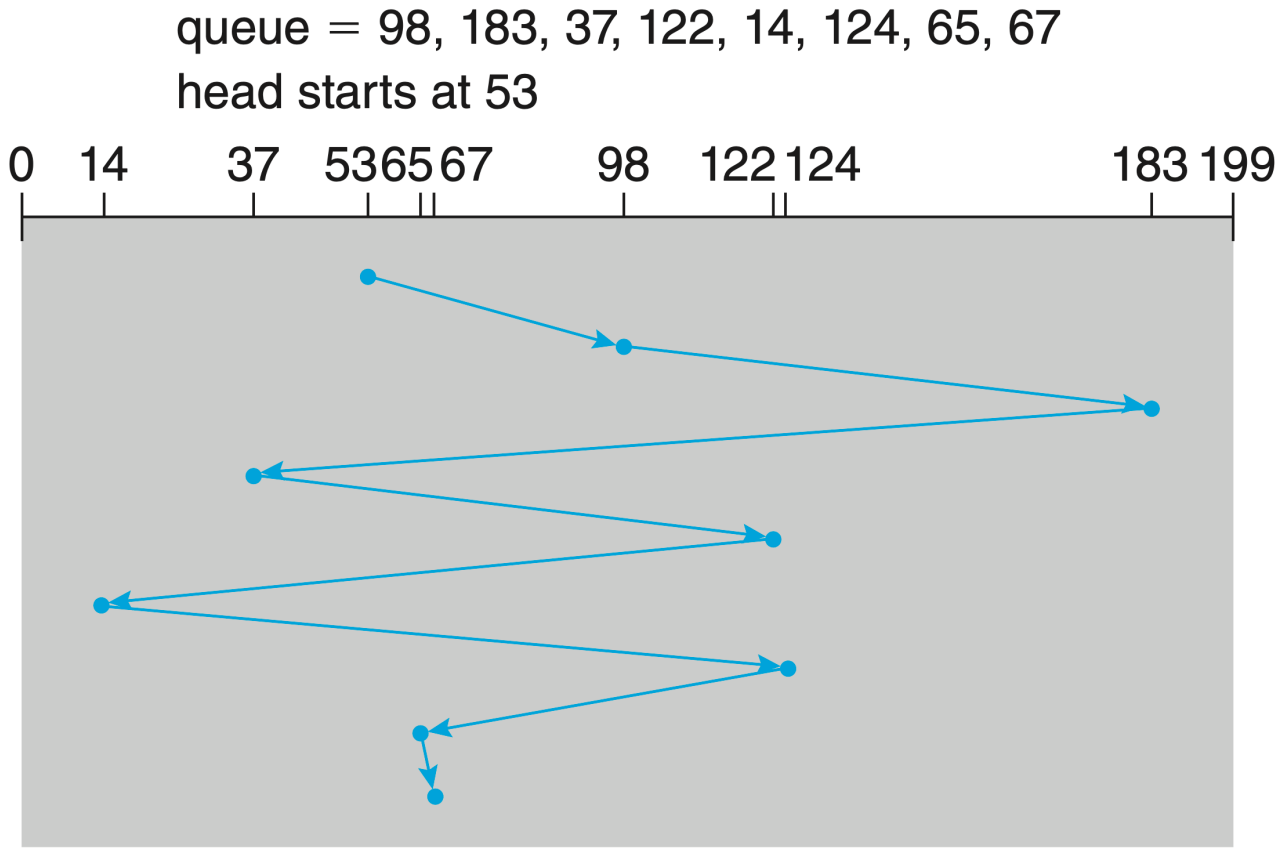
- Trách nhiệm của hệ điều hành là sử dụng phần cứng hiệu quả. Đối với HDD, OS phải giảm thiểu thời gian truy cập và tối đa hóa băng thông truyền dữ liệu.
- Đối với HDD, thời gian truy cập (access time) = seek time + rotational latency.
- Băng thông là tổng số byte được truyền / tổng thời gian từ khi yêu cầu truyền dữ liệu đầu tiên đến khi hoàn thành yêu cầu truyền cuối cùng.
- Các thuật toán lập lịch đĩa cứng được sử dụng để quyết định thứ tự truy cập các block dữ liệu trên đĩa cứng khi có nhiều yêu cầu truy xuất đến đĩa . Mục tiêu của các thuật toán này là tối ưu hóa hiệu suất của hệ thống và giảm thời gian truy cập đĩa cứng.

Lập lịch HDD

- Một số thuật toán lập lịch đĩa cứng: di chuyển đầu đọc theo
  - FCFS: thứ tự xuất hiện, first come first serverd
  - Shortest Seek Time First (SSTF): yêu cầu có thời gian di chuyển đầu đọc ngắn nhất so với vị trí hiện tại.
  - SCAN(Elevator): di chuyển từ đầu → cuối hoặc ngược lại. Khi đến đầu/cuối, đầu đọc đảo chiều và tiếp tục.
  - C-SCAN (Circular SCAN): giống SCAN nhưng chỉ di chuyển 1 hướng.
  - LOOK: tương tự SCAN, chỉ dừng lại khi không còn yêu cầu trên hướng di chuyển của đầu đọc.
  - C-LOOK: kết hợp của C-SCAN và LOOK, chỉ quét theo một hướng và không quay đầu lại khi kết thúc.
- Sử dụng website https://www.seektime.app để minh họa và tính toán kết quả.

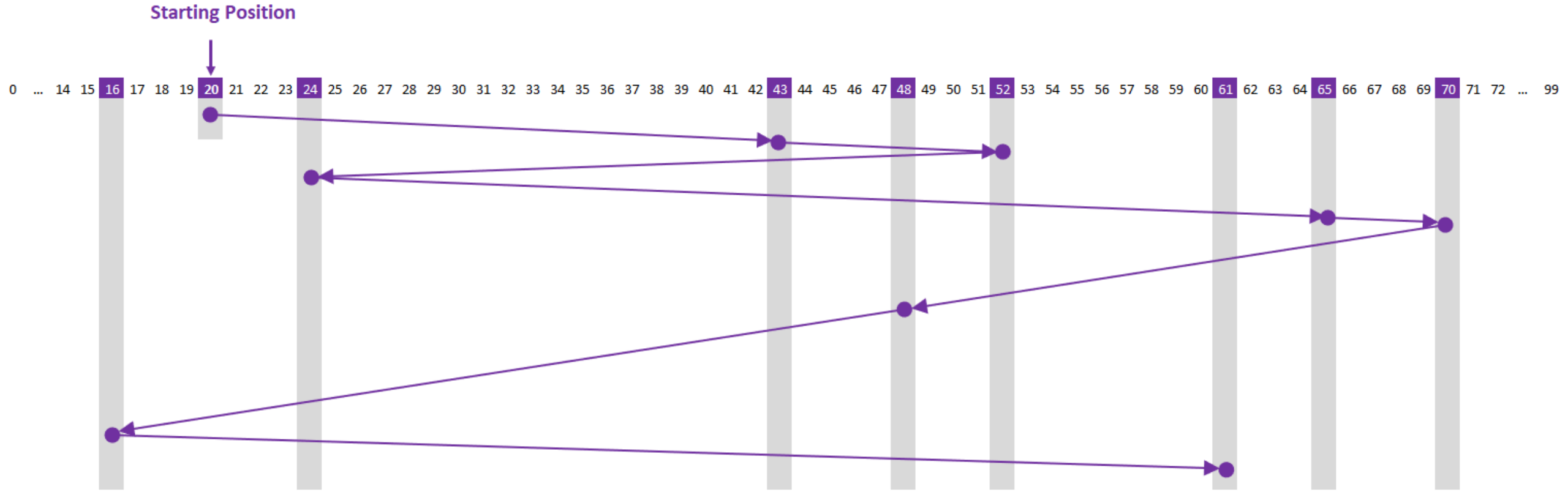
FCFS

- Phục vụ các yêu cầu theo thứ tự đến.
- Ví dụ: cho một hàng đợi với các yêu cầu đọc đĩa tại các cylinder: 98, 183, 37, 122, 14, 124, 65, 67. Nếu đầu đọc đĩa đang ở vị trí cylinder 53, nó sẽ di chuyển từ 53 → 98, sau đó đến 183, 37, 122, … cuối cùng là 67 với tổng số cylinder đã duyệt qua là 640.



FCFS

- Ví dụ: cho một hàng đợi với các yêu cầu đọc đĩa tại các cylinder: 43,52,24,65,70,48,16,61. Biết đầu đọc đang ở cylider 20, tính tổng số cylinder đã duyệt qua.

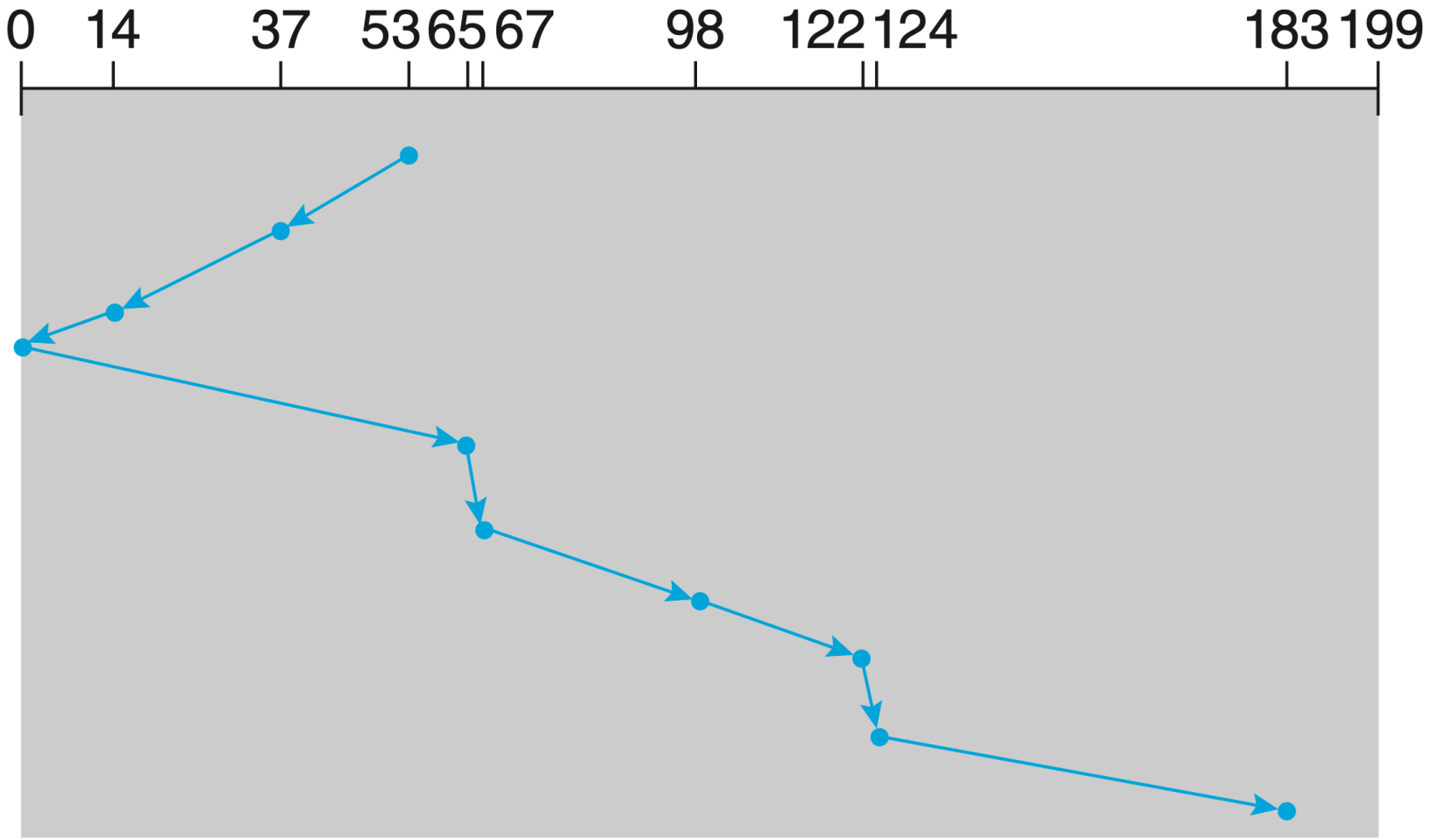


- KQ = |20-43| + |43-52| + |52-24| + |24-65| + |65-70| + |70-48| + |48-16| + |16-61| = 205

SCAN

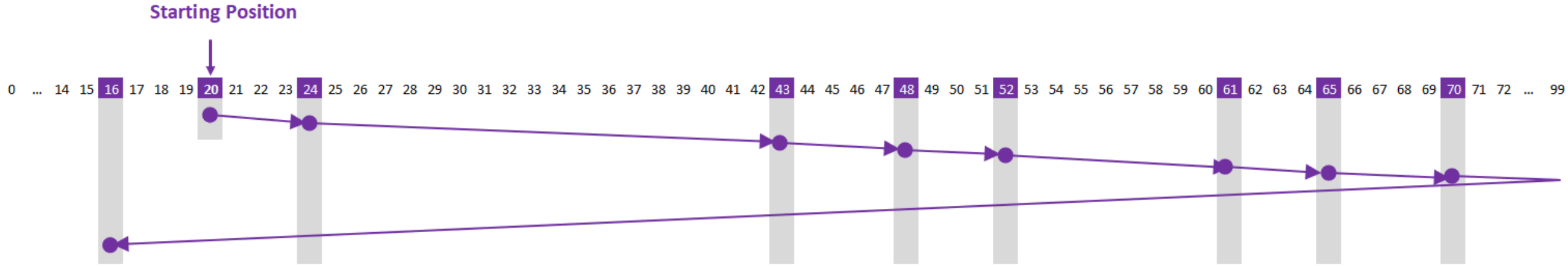
queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53



SCAN

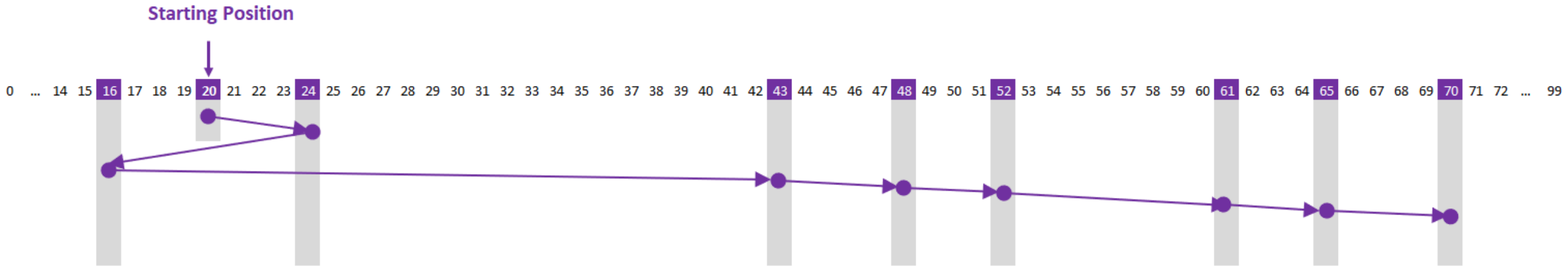
- Queue = 43,52,24,65,70,48,16,61. Đầu đọc ở cylinder 20, di chuyển hướng right.



- Số cylinder đi qua: 162

Shortest Seek Time First (SSTF)

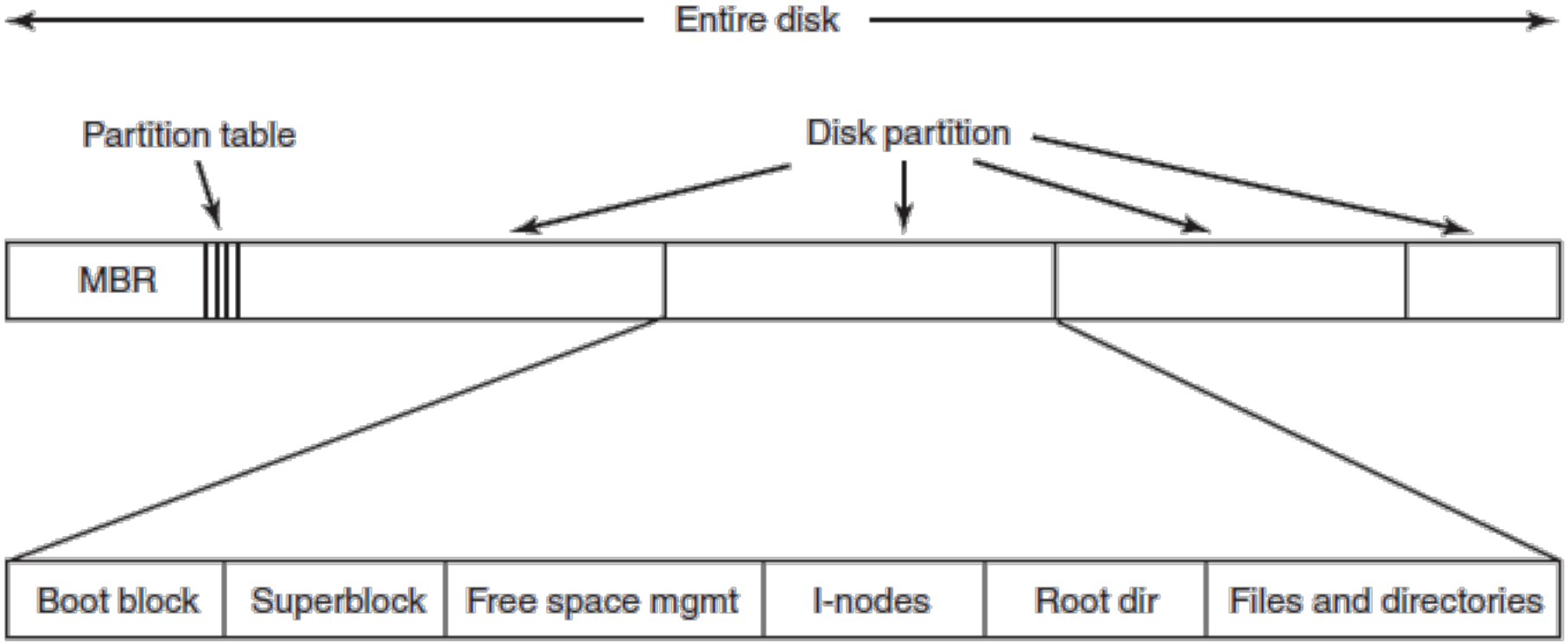
- Queue = 43,52,24,65,70,48,16,61. Đầu đọc ở cylinder 20.



- Số cylinder đi qua: |24-20| + |24-16| + |70-16| = 66
- Hoặc có thể bắt đầu từ cylinder 20 → 16 → 24

- Định dạng đĩa cấp thấp: tạo ra các track và sector cơ bản.
  - Mỗi sector có cấu trúc dữ liệu gồm: header – data – trailer
  - Header và trailer chứa các thông tin dành riêng cho disk controller như chỉ số sector và error-correcting code (ECC)
  - Khi disk controller ghi dữ liệu lên một sector, ECC được cập nhật với giá trị được tính dựa trên dữ liệu ghi.
  - Khi disk controller đọc dữ liệu từ một sector, giá trị ECC được tính lại và so sánh với giá trị đã lưu trước đó nhằm phát hiện lỗi.

- Phân vùng đĩa (partitioning)
  - Chia đĩa thành nhiều phân vùng (partition), mỗi phân vùng gồm nhiều block liên tục.
  - Có bảng phân vùng để quản lý các phân vùng, mỗi phân vùng xem như một ổ đĩa logic.
  - Phân vùng được định dạng với một hệ thống file (file system): FAT32, NTFS, ext4…

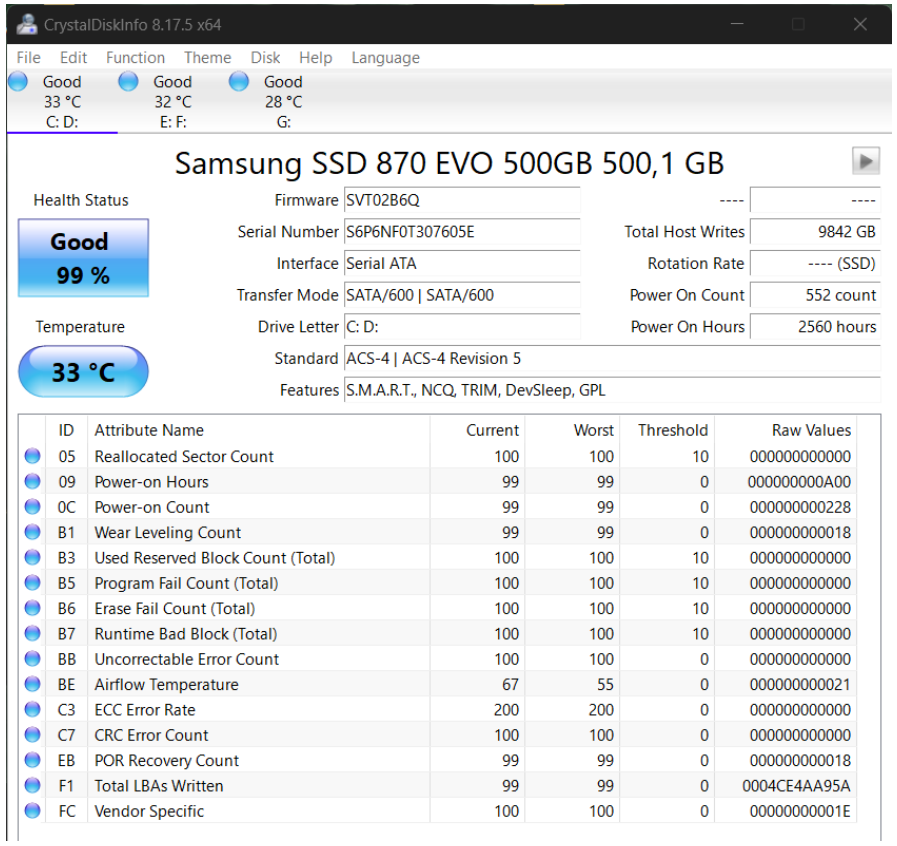


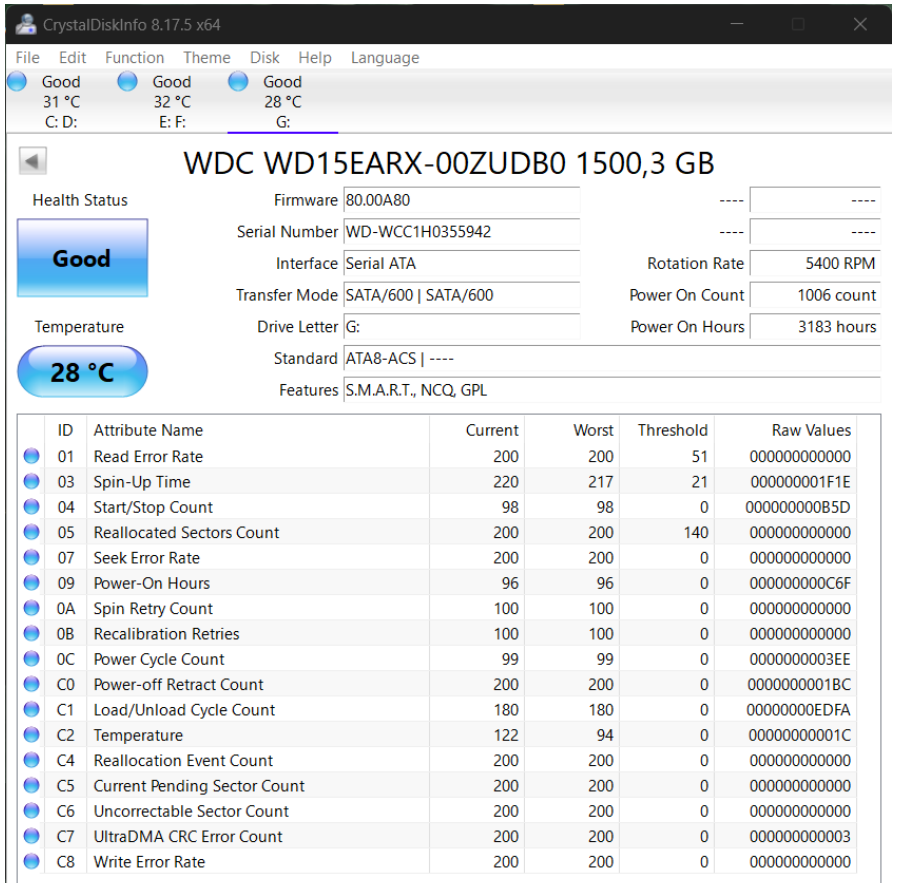
- Raw disk: partition không có hệ thống file.
  - I/O lên raw disk được gọi là raw I/O: đọc/ghi trực tiếp các block, không dùng các dịch vụ của File System như buffer cache, file locking…
  - Một số hệ thống CSDL (như Oracle) sử dụng raw disk

- Quản lý không gian tráo đổi (swap space):
  - Swap space là không gian đĩa được sử dụng để mở rộng không gian nhớ khi sử dụng bộ nhớ ảo.
  - Mục tiêu: cung cấp hiệu suất cao nhất cho hệ thống quản lý bộ nhớ ảo.
  - Giải pháp của các hệ điều hành:
    - ✓ Linux: cần 1 partition riêng (hoặc swap file) dành cho swap
    - ✓ Windows: sử dụng swap file pagefile.sys
    - ✓ macOS: hỗ trợ swap bằng cả file và partition.

Định dạng, phân vùng, raw disk

- Quản lý các khối bị lỗi: trường hợp tồn tại một số sector bị lỗi:
  - Ngay sau khi xuất xưởng: tự sửa bằng cách thay thế với các sector, track dự trữ.
  - Phát hiện sau một thời gian sử dụng: hệ điều hành thông báo để disk controller sửa (sử dụng sector, track dự trữ) hoặc đánh dấu lỗi.



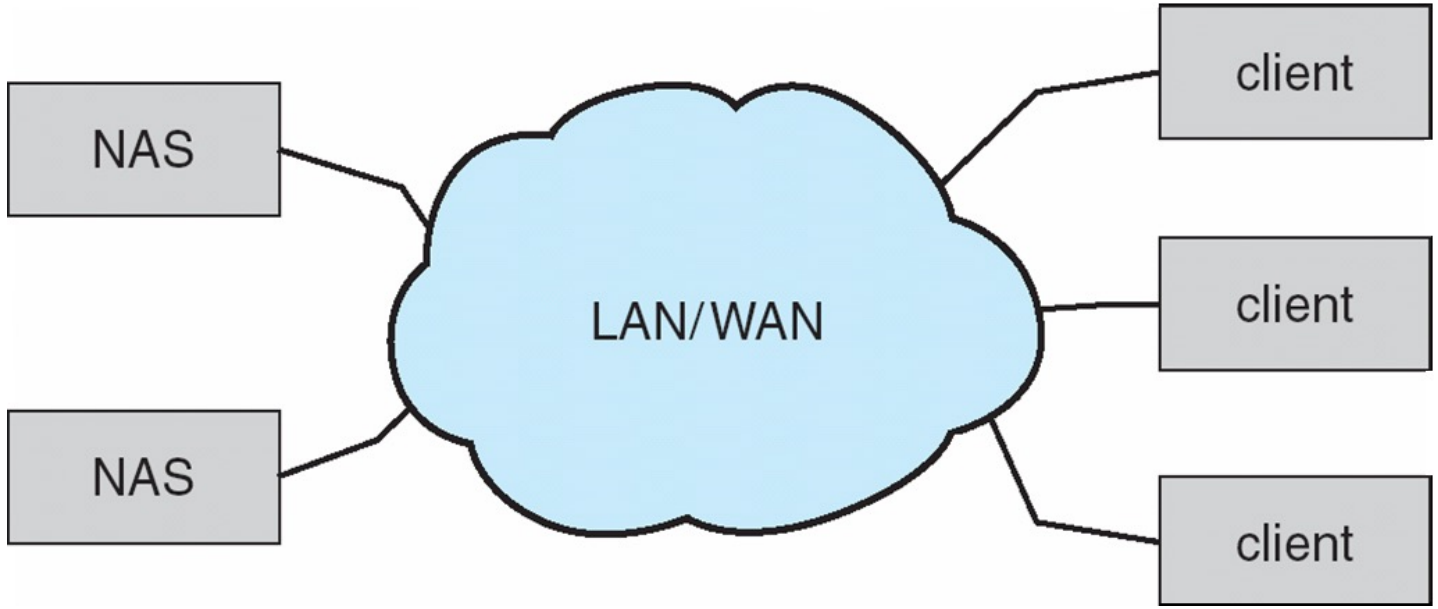


Kết nối thiết bị lưu trữ

- Máy tính kết nối với thiết bị lưu trữ theo 3 cách:
  - host-attached
  - network-attached
  - cloud
- Host-attached thực hiện thông qua các cổng I/O với nhiều công nghệ khác nhau như USB, firewire, thunderbolt, fibre

Kết nối thiết bị lưu trữ

- Network-attached storage (NAS):
  - Là một loại thiết bị lưu trữ dữ liệu được kết nối vào mạng và cung cấp khả năng chia sẻ tập tin và dữ liệu với các thiết bị khác trong cùng một mạng.



Kết nối thiết bị lưu trữ

- Cloud storage:
  - Tương tự như NAS nhưng có thể truy cập qua Internet. VD: Google Drive, Dropbox, Microsoft OneDrive, iCloud, Amazon S3.
  - Ưu điểm của cloud storage:
    - ✓ Truy cập mọi nơi
    - ✓ Đồng bộ dữ liệu
    - ✓ Tăng dung lượng lưu trữ dễ dàng
    - ✓ Bảo mật dữ liệu
    - ✓ Dễ dàng chia sẻ dữ liệu
    - ✓ Dễ dàng phục hồi dữ liệu

Storage Array

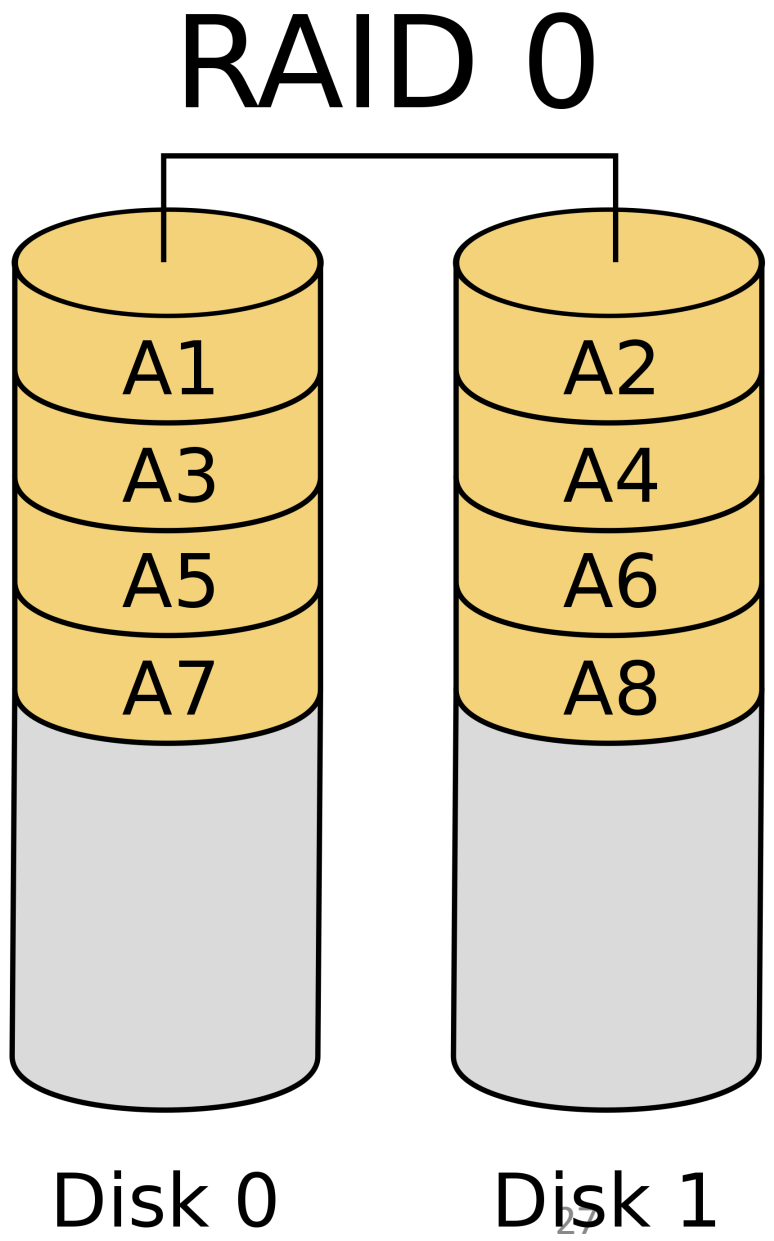
- Storage Array là một hệ thống lưu trữ dữ liệu tập trung và có thể mở rộng, được thiết kế để cung cấp dung lượng lưu trữ lớn và hiệu suất cao.
- Một Storage Array thường bao gồm nhiều ổ đĩa cứng được kết hợp với các tính năng như RAID (Redundant Array of Independent Disks), phục vụ mục đích cung cấp khả năng lưu trữ an toàn và tăng cường hiệu suất.
- Storage Array có nhiều đặc điểm: khả năng mở rộng dễ dàng, hiệu suất cao, bảo vệ dữ liệu tốt nhờ RAID, quản lý tập trung, khả năng sao lưu và khôi phục dữ liệu…
- Storage Array thường được triển khai trong các môi trường doanh nghiệp với nhu cầu lưu trữ dữ liệu lớn và hiệu suất cao.

RAID (Redudant Arrays of Independent Disks)

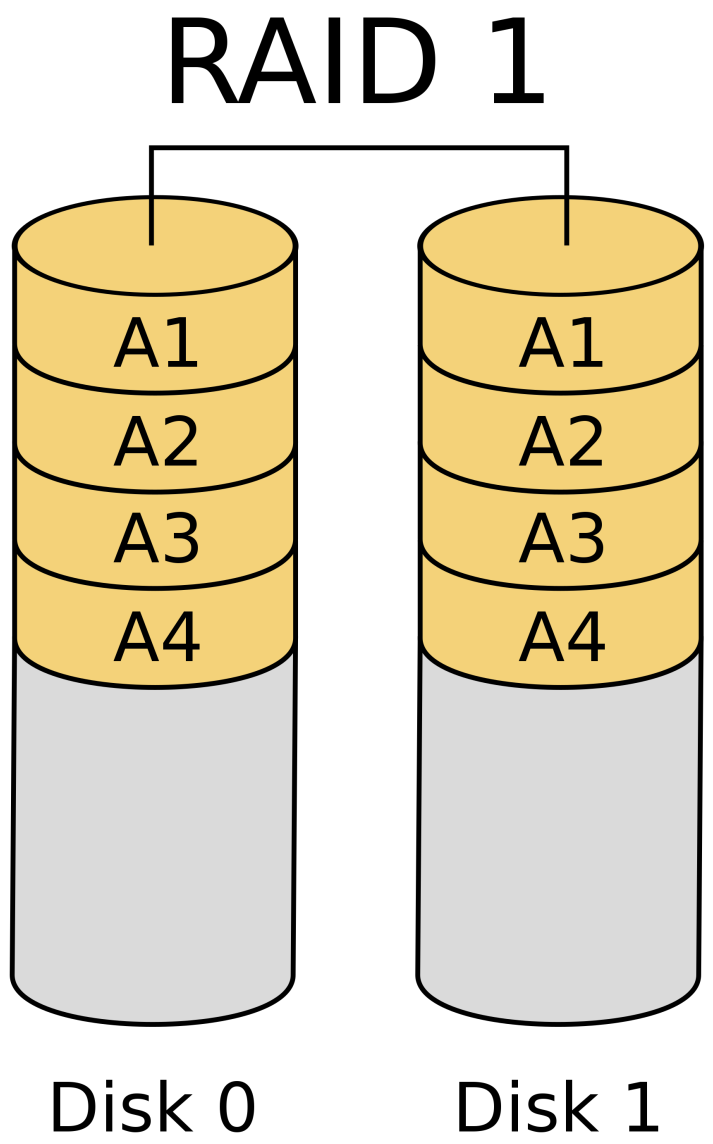
- RAID là hình thức ghép nhiều ổ đĩa cứng vật lý thành một hệ thống ổ đĩa cứng có chức năng gia tăng tốc độ đọc/ghi dữ liệu hoặc nhằm tăng thêm sự an toàn của dữ liệu chứa trên hệ thống đĩa hoặc kết hợp cả hai yếu tố trên.
- RAID được chia thành 7 cấp độ (từ 0 → 6), hầu hết đều được xây dựng từ hai cấp độ cơ bản là RAID 0 và RAID 1:
  - RAID 0 (Striping): Dữ liệu được chia thành các "stripes" (đoạn) và ghi đồng thời lên nhiều ổ đĩa.
  - RAID 1 (Mirroring): Dữ liệu được sao chép đồng thời lên ít nhất hai ổ đĩa khác nhau.

RAID (Redudant Arrays of Independent Disks)

- RAID 0 cần ít nhất 2 ổ đĩa (có thể sử dụng 1 ổ đĩa). Tổng quát ta có n đĩa (n >= 2) và các đĩa là cùng loại.
- Dữ liệu sẽ được chia ra nhiều phần bằng nhau. Ví dụ dùng 02 ổ cứng 80GB thì hệ thống đĩa có 160GB.
- Ưu điểm: Tăng tốc độ đọc / ghi đĩa: mỗi đĩa chỉ cần phải đọc/ghi 1/n lượng dữ liệu được yêu cầu. Lý thuyết thì tốc độ sẽ tăng n lần.
- Nhược điểm: Tính an toàn thấp. Nếu một đĩa bị hư thì dữ liệu trên tất cả các đĩa còn lại sẽ không còn sử dụng được. Xác suất để mất dữ liệu sẽ tăng n lần so với dùng ổ đĩa đơn.

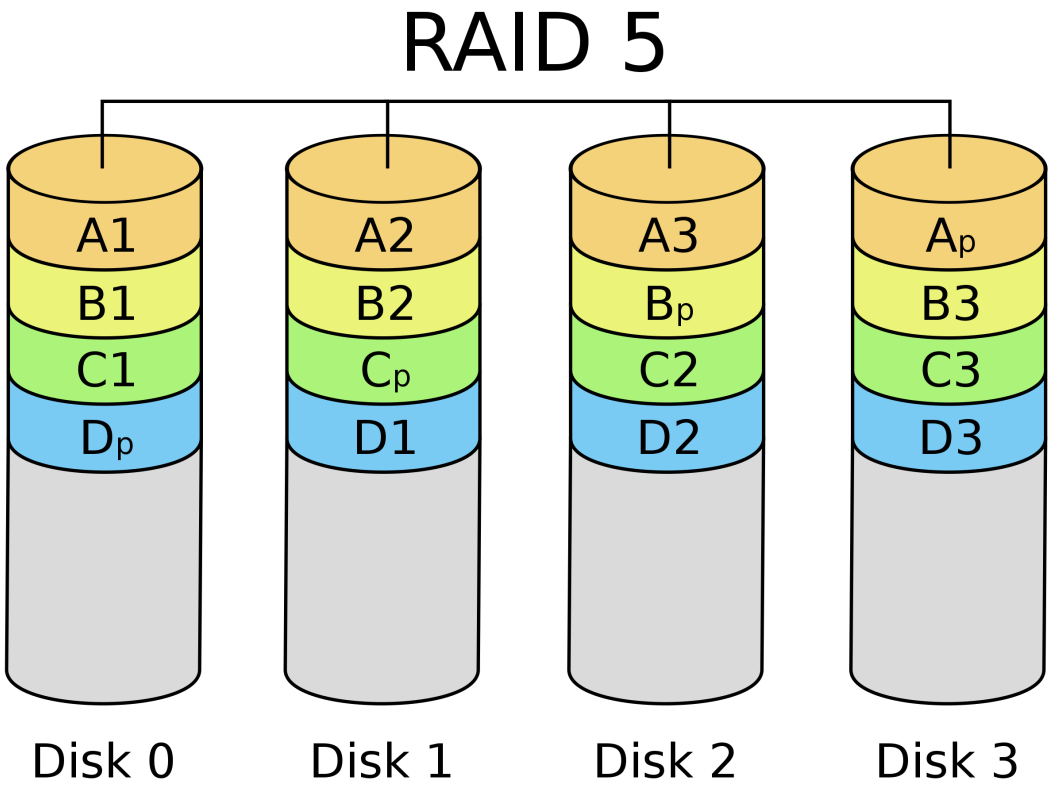


- RAID 1 đòi hỏi ít nhất hai đĩa cứng để làm việc. Dữ liệu được ghi vào 2 ổ giống hệt nhau (Mirroring). Trong trường hợp một ổ bị trục trặc, ổ còn lại sẽ tiếp tục hoạt động bình thường.
- Người dùng có thể thay thế ổ đĩa bị hỏng mà không phải lo lắng đến vấn đề thông tin thất lạc.
- Đối với RAID 1, hiệu năng không phải là yếu tố hàng đầu.
- Dung lượng cuối cùng của hệ thống RAID 1 bằng dung lượng của ổ đơn (hai ổ 80GB chạy RAID 1 sẽ cho hệ thống nhìn thấy duy nhất một ổ RAID 80GB).



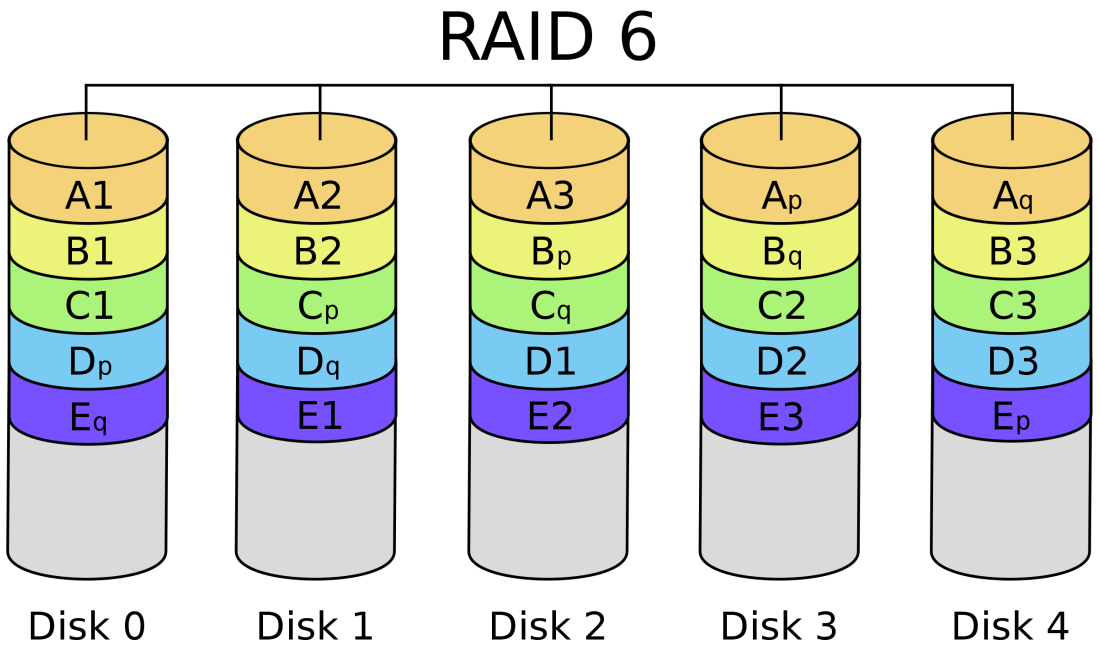
RAID (Redudant Arrays of Independent Disks)

- RAID 5 là sự cải tiến của RAID 0, có cung cấp cơ chế khôi phục dữ liệu.
- Giả sử dữ liệu A được phân tách thành 3 phần A1, A2, A3, khi đó dữ liệu được chia thành 3 phần chứa trên các ổ đĩa cứng 0, 1, 2 (giống như RAID 0). Phần ổ đĩa cứng thứ 3 chứa Parity (Ap) của A1 A2 A3 để khôi phục dữ liệu có thể sẽ mất ở ổ đĩa cứng 0, 1, 2.
- Dữ liệu B được chia thành B1 B2 B3 và Parity của nó là Bp, theo thứ tự B1 B2 B3 được lưu trữ tại ổ 0 1 3, và Bp được lưu trữ tại ổ 2. Các Parity được lưu trữ tuần tự trên các ổ đĩa cứng.
- RAID 5 cho phép tối đa có 1 ổ cứng bị chết tại một thời điểm và yêu cầu các ổ cứng tham gia RAID phải có dung lượng bằng nhau.
- Dung lượng khi dùng RAID 5 được tính bằng cách: (Dung lượng của 1 ổ cứng) x [(Số lượng các ổ cứng tham gia RAID) - 1]
- Yêu cầu tối thiểu của RAID 5 là có ít nhất 3 ổ đĩa cứng.



RAID (Redudant Arrays of Independent Disks)

- RAID 6 là dạng RAID thường được sử dụng trong các doanh nghiệp.
- Tương tự như RAID 5, nhưng RAID 6 vượt trội hơn bởi khả năng sử dụng đến hai khối parity và yêu cầu tối thiểu 4 ổ đĩa.
- Nếu có hai ổ đĩa chết cùng một lúc, hệ thống vẫn có thể tiếp tục hoạt động.
- Về mặt hiệu suất, RAID 6 có tốc độ ghi không bằng RAID 5 do phải tính toán nhiều khối parity phức tạp hơn, nhưng có tốc độ đọc ngẫu nhiên nhanh hơn do dữ liệu được stripe qua nhiều ổ đĩa hơn.
- Giống như RAID 5, hiệu suất RAID 6 có thể được điều chỉnh bằng cách thay đổi kích thước stripe.



Tham khảo

- Operating System Concepts
- Hệ điều hành – ThS. Lương Minh Huấn
- https://www.javatpoint.com/os-disk-scheduling
- https://www.101computing.net/disk-scheduling-algorithms/
- https://www.geeksforgeeks.org/disk-scheduling-algorithms/
- https://www.seektime.app/