Sistemas Operativos: Input/Output Disks

Pedro F. Souto (pfs@fe.up.pt)

April 28, 2012

Topics

Magnetic Disks

RAID

Solid State Disks

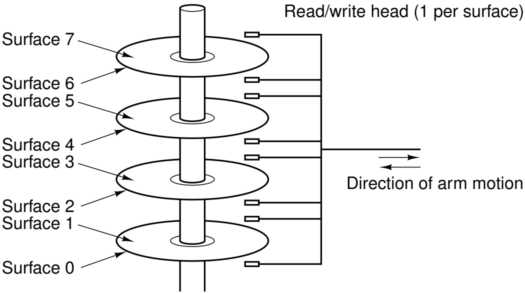
Topics

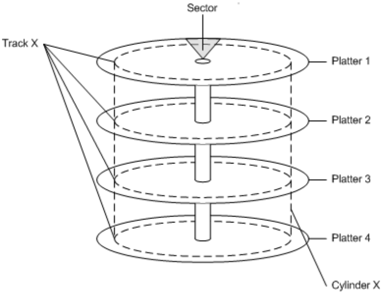
Magnetic Disks

RAID

Solid State Disks

Magnetic Disk Construction



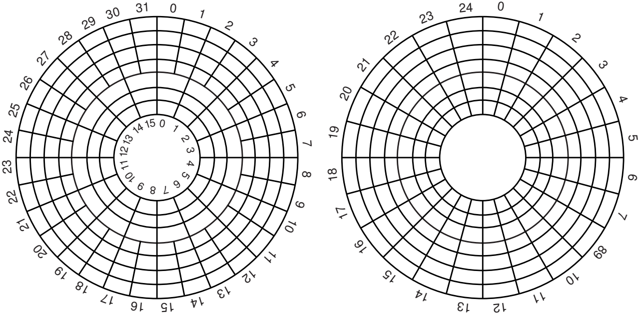


Track A concentric ring on a platter surface

Sector An arc of a track with a fixed number of bytes that is individually addressable

Cylinder A set of tracks of all platters under the heads at a given position

Disks: Physical vs. Logical Geometry



- Modern disks have several zones (2 in the left figure), each with a fixed number of sectors per track
- Nowadays, disks use **Logical Block Addressing** a scheme where sectors are numbered starting at 0
  - It is up to the disk controller to map the LBA sector number to the physical sector on the disk
  - Earlier, some disks would advertise a (logical) geometry (CHS) that might be different from the physical geometry

Modern Disk Specs: Seagate

	Cheetah 15K.7	Barracuda
Class	Enterprise	Business
Capacity		
Formatted capacity (GB)	600	3000
Discs	4	3
Heads	8	6
Sector size (B)	512	4096
Performance		
External interface	6 Gbit/s Ser. Att. SCSI	6 Gbit/s SATA
Rotational speed (rpm)	15,000	7,200
Average latency (ms)	2.0	4.17
Seek time, rd/wr (ms)	3.4/3.9	8.5/9.5
Sust. Transfer rate (MB/s)	122 to 204	<210
Cache Size (MB)	16	64
Reliability		
Non-recoverable read errors	1 sector per 1E16	1 sector per 1E14
MTBF	1,600,000	
Annual. Failure Rate (AFR)	0.55%	1%

Disk Performance Times

Seek time Time required to position the head over the track with the sector to access

- Read seeks are shorter than write seeks (see above). Why?
- Typically between 3 and 10 ms

Rotational latency Time required for the desired sector to rotate undern the head

- On average, half of the rotation time, which depends on the rotational speed (2 ms/ 4.17 ms / 5.56 ms)

Transfer time Time required to transfer the data, always a multiple of a sector

- Sustained transfer bandwidth ranges from 40 to 200 MB/s. For 40 MB/s:

Block Size (B)	Transfer Time (ms)
512	0.013
4096	0.103
1.? M	25.013

Disk Scheduling Algorithms and FCFS

Observation Seek time is one of the main factors in disk performance

Idea Order the service disk access requests so as to minimize seek time.

FCFS

Idea Process requests in the order they are submitted

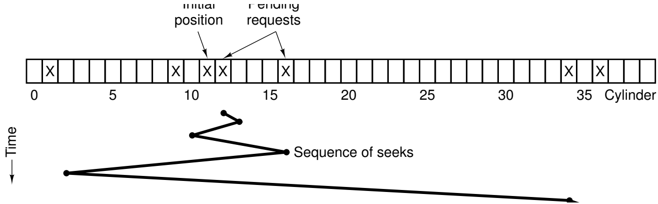
Pros

- Simple and fair

Cons

- Unnecessarily long seeks, with wild arm swings

Shortest Seek Time First: SSTF



Idea Process the request that requires the shortest seek time

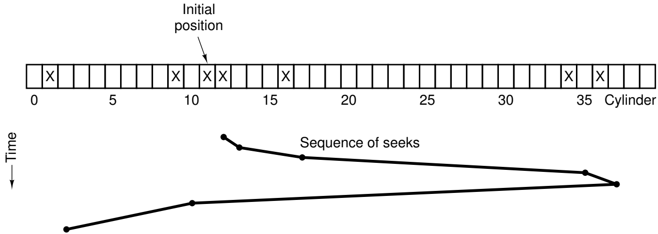
Pros

- Tries to minimize seek time ...
- ... but it is not optimal

Cons

- May lead to starvation

Elevator (SCAN)



Idea Use an algorithm similar to that used in elevators

- There is no need to go until the end (LOOK)

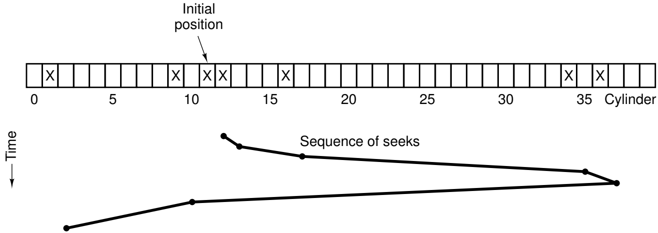
Pros

- No starvation

Cons

- Requests on the wrong end may take too much time
  - Rotational latency is of the same order as seek time

Circular SCAN (C-SCAN)



Idea Like scan, but service requests in only one direction

- There is no need to go until the end (C-LOOK)

Pros

- No starvation
- Equal treatment independent of the track

Cons

- Does nothing on the return arm movement

- Disk space management may also be important
  - But with LBA the driver does not really know much to make the best decisions

Topics

Redundant Array of Independent Disks

I was for **Inexpensive** in the original proposal, which dates back from the late 80's

Idea Store the data in a disk array so as to improve

- Performance by executing in parallel several disk operations on different disks
- Reliability by storing redundant information, so that if one disk fails, its content can be recovered

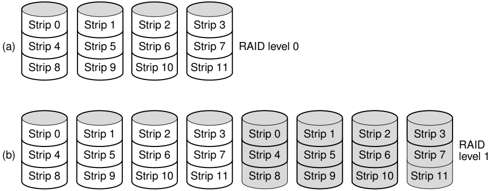
Transparency The RAID controller interfaces to the OS just as a single disk controller

- Most RAID controllers are SCSI controllers
  - Which allow the attachment of up to 7/15 devices

Cons Some:

- Controller complexity OK!
- Cost Technological breakthroughs made larger disks much more cost effective than smaller disks

RAID Levels 0 & 1



Strip Is a set of *k* consecutive sectors, for some fixed *k*

- Access to strips that are in different disks can be done in parallel

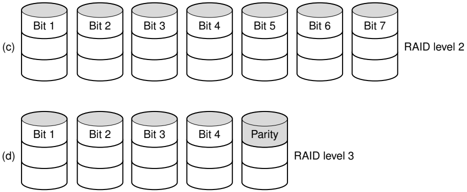
RAID 0

- Higher performance for large I/O requests, or smaller concurrent I/O requests as long as ...
- Reliability is worse than for single disk, because ...

RAID 1 RAID 0 with mirroring

- Read load can be distributed over all disks with the desired data
- Highly reliable and recovery from a disk failure is straightforward

RAID Levels 2 & 3



RAID 2 More space efficient than RAID 1

- Splits stream in chunks of fixed size (nibbles in the fig.)
- Each chunk is stored using Hamming code ((7,4) in the fig.), 1 bit per drive
- Can recover from the failure of one disk

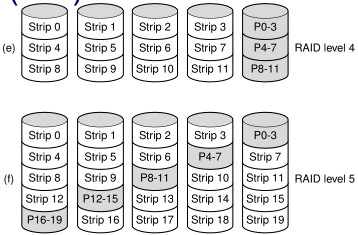
RAID 3 Use just a parity bit rather than a 7-bit Hamming code

- Lower cost at the expense of lower reliability
- Still able to recover damaged disk content, as long as one can identify it

Both 2 and 3

- Higher throughput than RAID 0 or 1
- But does not support concurrent I/O operations
- Require synchronized disks (hard)

RAID Levels (3/3)



RAID 4 RAID 0 with one additional drive to store a "parity strips"

- Additional drive allows to rebuild one crashed drive
- Parity drive may be a bottleneck
  - Write to a strip requires reading and writing from at least two drives

RAID 5 RAID 4 with the parity strip distributed over all the drives

RAID 6 General term used to refer to a RAID scheme that is able to tolerate two simultaneous disk failures

- The method used to achieve that is not prescribed

Topics

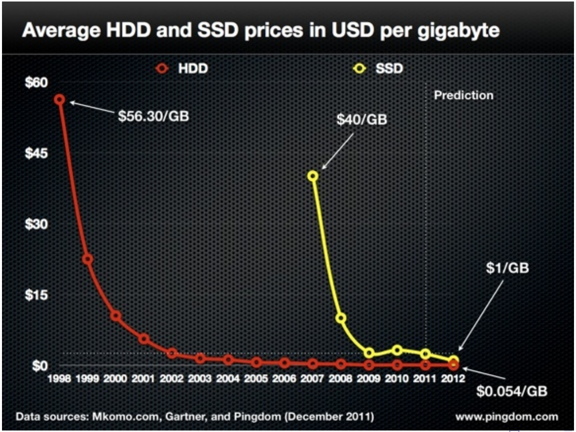
Magnetic Disks

RAID

Solid State Disks

Solid State Disks: FLASH memory

- Non-volatile RAM
- Relies on Moore's law for increasing chip density
- In 2012, SSD have reached the magic cost of 1 USD/GB
  - This is the cost of HDD about 10 years ago
  - Will the cost of SSD replicate the same evolution as that of HDD cost?



SSD vs. Magnetic Disks

+ No moving parts
  - More reliable mechanically
  - More shock-resistant
+ Faster access than disk
- 20 times more expensive than disk (see chart)

SSD Organization

Pages Access (read/write) unit

- Typical size: 512-4096 bytes

Blocks Set of pages

- Erasing unit
  - Rewriting a page requires erasing its block
    - Can write only 0's
    - Require an erase (all 1's) before writing
- Typical size: 16-256 KB

Modern SSD Specs: Intel

	710 Series	520 Series	320 Series
Capacity			
Launch Quarter	Q3'11	Q1'12	Q1'11
Max Formatted capacity (GB)	300	480	300
Performance			
External interface (SATA)	3 Gbit/s	6 Gbit/s	3 Gbit/s
Latency time, rd/wr (us)	75/85	80/85	75/90
Sequential rd/wr (MB/s)	270/210	550/520	270/205
Random Access (IOPS)			
8GB span		50,000/42,000	39,500/39,500
100% span	38,500/2,000		23,000/400
Reliability			
Non-recoverable read errors	1 sect./1E16	1 sect./1E16	1 sect./1E16
MTBF	2,000,000	1,200,000	1,200,000

SSD Technical Challenges and Solutions

Limited lifetime

- Number of writes is limited to a few tens of thousands
- By spreading the writes evenly, these problems can be minorated, but number of blocks is much smaller than number of pages

Rewriting performace limitations must erase block

Solution The SSD controller can minorate some of these problems

- Thus, this is mostly transparent to the OS