Non-Volatile Storage

- To keep data even if the system is shutdown
  - C.f. RAM (SRAM / DRAM) is volatile
    - Data stored in RAM is disappeared after the shutdown
- Historically, there are many non-volatile devices
  - Magnetic tapes
  - Magnetic drams
  - Magnetic hard disk drives (HDD)
  - Magnetic floppy disk
  - Optic disks (LD, CD, DVD, Blue-ray)
  - Magnetic optic disks (MD)
  - Flash Memory / Solid state drives (SSD)

Trends

- Currently, HDD and SSD are most popular
  - HDD: Large scale systems (Cloud centers)
  - SSD: Small size systems (even in large scale systems)
- Capacity and Price
  - HDD has larger capacity and cheap
  - SSD has smaller capacity and still expensive
    - Has limitation of write counts to flash memory
- New trends
  - New non-volatile memory devices
    - MRAM, PRAM, ReRAM, FeRAM, …

In this course

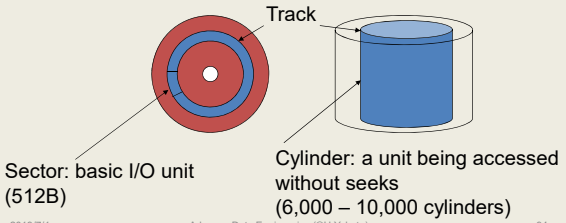
- We still focus on HDD
  - Considering large scale applications such as DWH
  - Current DBMS assume access model of HDD
- We also consider SSD and new devices

The First Hard Disk Drive

- IBM 305 RAMAC (1957)
  - Random Access Method of Accounting and Control
  - Size
    - Diameter :60 cm
    - 50 aluminum platter
  - 1200 rpm
  - Capacity
    - 5 MB (!)
  - Density
    - 0.2 Kb/inch^2

Access for a Hard Disk

- Rotation Speed: 3,000 – 15,000 rpm
  - Average rotational latency: 10ms – 2ms
- Seek Time: 1ms (next track) – 20ms (most Inside – outside)



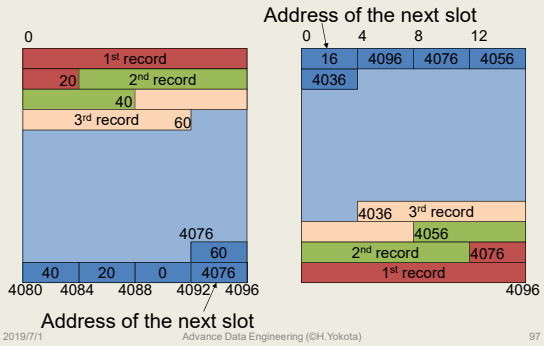
Disk Pages

- A Disk Page: A unit to assess disks
  - A sector is too small to databases
    - It requires frequent disk access commands
  - The size of a page is usually 4KB [8 sectors]
    - Occasionally 512B - 1MB
- Calculate disk access time
  - Assuming
    - Average seek time=3ms
    - Rotation Speed: 15,000 rpm
      - (i.e. average rotational latency = (60/15,000)/2 = 2ms)
    - Data transfer bandwidth 10MB/s

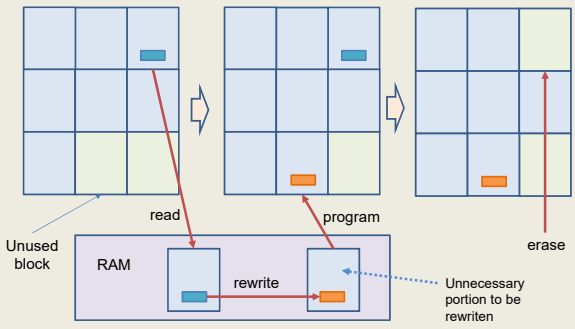
Disk Access

- Disk access time
  = seek time + rotational latency + data transfer
  - 512B page:
  - 4KB page:
  - 1MB page:
- Access to a record (corresponding to a tuple)
  - Relative record number
    - Tuple ID (TID)= Page# + Slot#
  - Effective use of disk space
    - Stuffing records and their header information from the opposite side

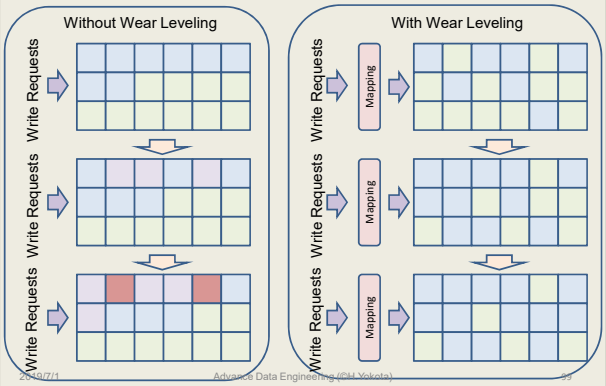
Configuration of a Page



Rewriting (NAND) Flash Memory



Wear Leveling

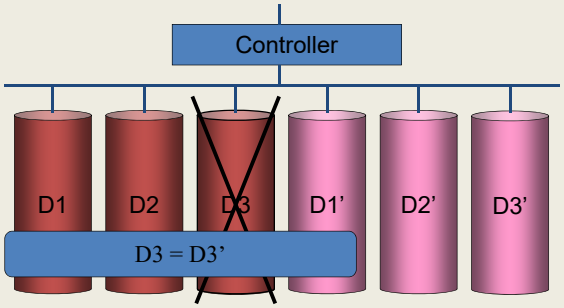


New Non-Volatile Memories

- Remove or reduce the limitation of write count
  - MRAM: Magneto-Resistive RAM
  - PRAM/PCM: Phase Change RAM
  - ReRAM: Resistance RAM
  - FeRAM: Ferroelectric RAM
- Still expensive or research level

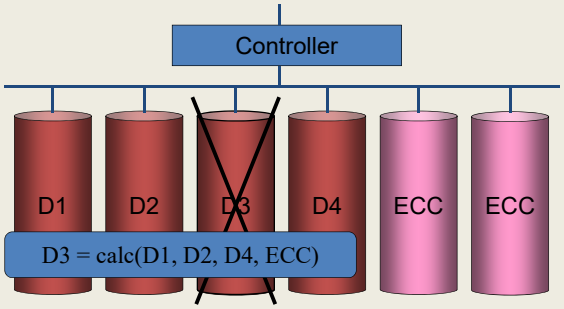
RAID 1

- Mirroring (Write into both D1 and D1')



RAID 2

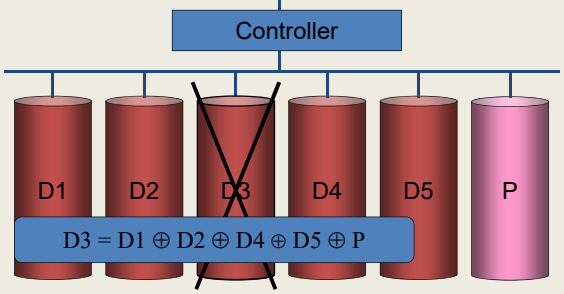
- ECC: Error Correction Code



RAID 3

- Parity Code

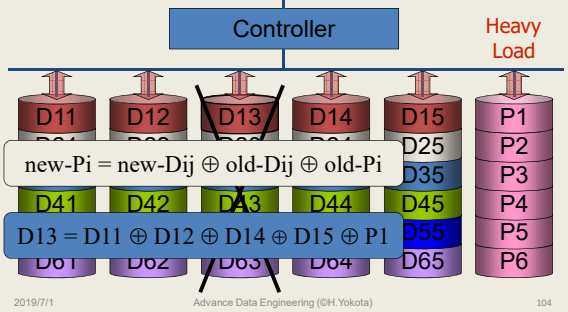
$$P = D1 \oplus \ldots \oplus Dj \oplus \ldots \oplus Dn$$



RAID 4

- Parity Code

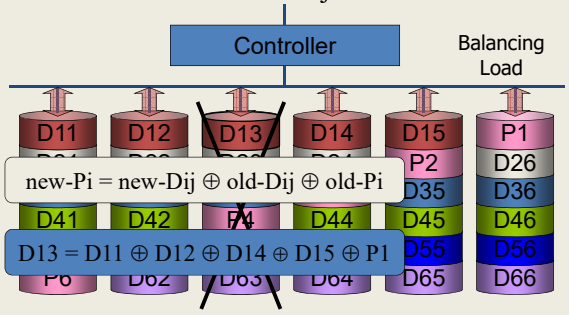
$$Pi = Di1 \oplus \ldots \oplus Dij \oplus \ldots \oplus Din$$



RAID 5

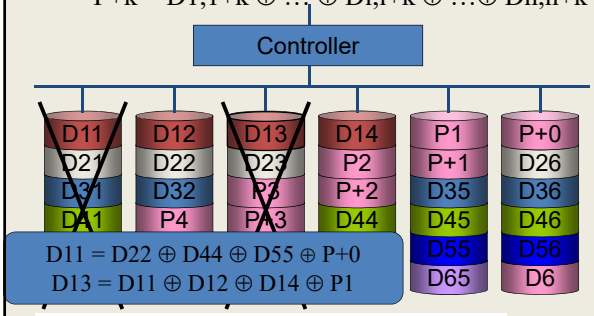
- Parity Code

$$Pi = Di1 \oplus \ldots \oplus Dij \oplus \ldots \oplus Din$$



RAID 6

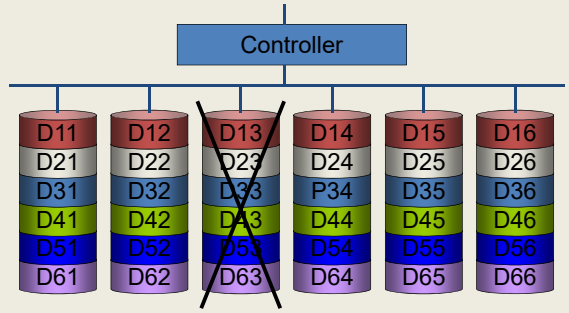
$$Pi = Di,1 \oplus \ldots \oplus Di,j \oplus \ldots \oplus Di,n$$
$$P+k = D1,1+k \oplus \ldots \oplus Di,i+k \oplus \ldots \oplus Dn,n+k$$



Other RAID6s uses Reed-Solomon Coding, EVENODD Coding

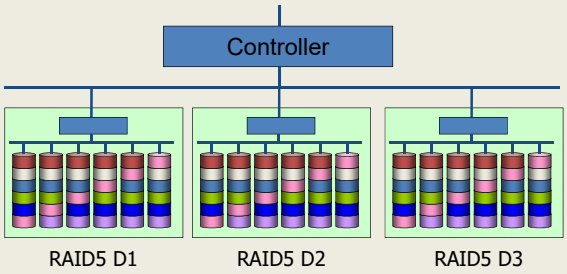
RAID 0

- No Fault-Tolerancy



RAID 50

- Combination of RAID 5 and RAID 0

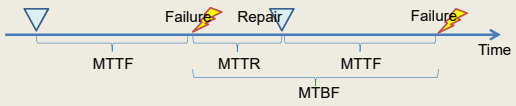


Comparison of RAID

	Small Read	Small Write	Large Read	Large Write	Storage Efficiency
RAID0	1	1	1	1	1
RAID1	1	1/2	1	1/2	1/2
RAID3	1/M	1/M	(M-1)/M	(M-1)/M	(M-1)/M
RAID5	1	max(1/M, 1/4)	1	(M-1)/M	(M-1)/M
RAID6	1	max(1/M, 1/6)	1	(M-2)/M	(M-2)/M

Reliability

- MTTF: Mean Time To Failure
- MTTR: Mean Time To Repair
- MTBF: Mean Time Between Failure



- Catalog Spec. of Recent HDD
  - More than 1,000,000 hour (more than 100 (114) years)
- In the case of RAID
  - MTTDL: Mean Time To Data Loss

Reliability of RAID 0, 1

- RAID 0

$$MTTDL_{RAID0} = \frac{MTTF_{disk}}{N}$$

- RAID 1 (10)

$$MTTDL_{RAID1} = \frac{MTTF_{disk}}{N} \times \frac{MTTF_{disk}}{MTTR_{disk}} = \frac{MTTF_{disk}^2}{N \times MTTR_{disk}}$$

  – If $MTTR_{disk}$ = 20 hours, 5×10^4 yeas for 100 disks

Reliability of RAID 3-5

- RAID 3-5

$$MTTDL_{RAID3\text{-}5} = \frac{MTTF_{disk}^2}{M \times (M\text{-}1) \times MTTR_{disk}}$$

  – If MTTR = 20 hours, more than 5×10^4 yeas for RAID 3-5 (M =10)

- RAID 30-50

$$MTTDL_{RAID30\text{-}50} = \frac{MTTF_{disk}^2}{G \times M \times (M\text{-}1) \times MTTR_{disk}}$$

  – If G = 10, M = 10 (N =100) , more than 5×10^3 years
    – worse than RAID10

Reliability of RAID 6

- RAID 6

$$MTTDL_{RAID6} = \frac{MTTF_{disk}^3}{M \times (M\text{-}1) \times (M\text{-}2) \times MTTR_{disk}^2}$$

  – Enough long MTTDL

- RAID 60

$$MTTDL_{RAID60} = \frac{MTTF_{disk}^3}{G \times M \times (M\text{-}1) \times (M\text{-}2) \times MTTR_{disk}^2}$$

  – Enough long MTTDL