W4118: RAID

Instructor: Junfeng Yang

References: Modern Operating Systems (3rd edition), Operating Systems Concepts (8th edition), previous W4118, and OS at MIT, Stanford, and UWisc

RAID motivation

- Performance
  - Disks are slow compared to CPU
  - Disk speed improves slowly compared to CPU
- Reliability
  - In single disk systems, one disk failure → data loss
- Cost
  - A single fast, reliable disk is expensive

RAID idea

- RAID idea: use redundancy to improve performance and reliability
  - Redundant array of cheap disks as one storage unit
  - Fast: simultaneous read and write disks in the array
  - Reliable: use parity to detect and correct errors

- RAID can have different redundancy levels, achieving different performance and reliability
  - Seven different *RAID levels* (0-6)

Evaluating RAID

- Cost: check disk capacity / total capacity
  - *Storage utilization*: data capacity / total capacity
- Reliability
  - Tolerance of disk failures
- Performance
  - (Large) sequential read, write, read-modify-write
  - (Small) random read, write, read-modify-write
  - Speedup over a single disk

Computing cost

- D = number of data disks in a RAID group
- C = number of check disks in a RAID group

- Cost = C/(D+C)

Computing reliability

- N = total number of disks
- D = number of data disks in a RAID group
- C = number of check/parity disks in a RAID group

- MTTF(disk) = mean time to failure for a disk
  - Estimated as MTTF (in years) = 1 / AFR (annual failure rate)
  - Ex) 114 years (1M hours) = 1 / 0.88%
  - Source: "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?", FAST'07
- MTTR = mean time to repair for a failed disk

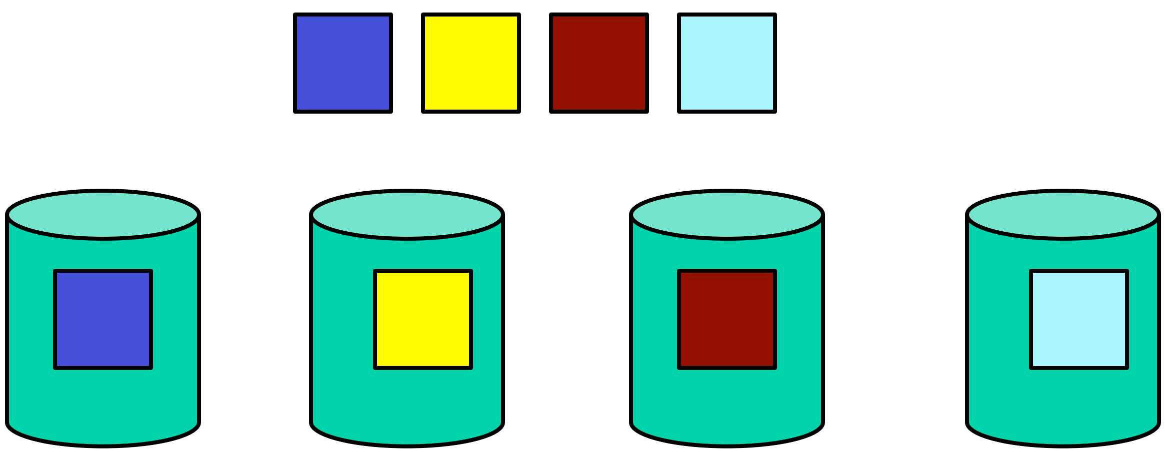
- MTTF(group) = mean time to two failed disks before first gets repaired in one group
- MTTF(raid) = mean time to failure over entire array
- MTTF(raid) = MTTF(group) / Num. groups

Computing reliability (cont'd)

- Assume single-error tolerance in one group
  - If another error comes before repair, group fails
- MTTF(group) = MTTF(1 disk) / Prob[Another failure within MTTR]
  - If Prob[...] ≈ 1, MTTF(group) same as MTTF(1 disk) - no benefit of RAID
  - If Prob[...] ≈ 0, MTTF(group) approaches ∞.
- MTTF(1 disk) = MTTF(disk)/(D+C)
- MTTF(another disk) = MTTF(disk)/(D+C-1)
- Prob[Another failure within MTTR] = MTTR/(MTTF(disk)/(D+C-1))
- MTTF(group) = MTTF(1 disk)/Prob[Another failure within MTTR] = (MTTF(disk))2/((D+C)*(D+C-1)*MTTR)
- Num groups G = N / (D+C)
- MTTF(raid) = MTTF(group) / G = MTTF(group) / (N/(D+C))
- Thus: MTTF(raid) = (MTTF(disk))2 / (N * (D+C-1) * MTTR)
- But: are the assumptions valid?

RAID 0: non-redundant striping

- Structure
  - Data striped across all disks in an array
  - No parity
- Advantages:
  - Good performance: with N disks, roughly N times speedup
- Disadvantages:
  - Poor reliability: one disk failure → data loss
  - MTTF(raid)=MTTF(disk)/N

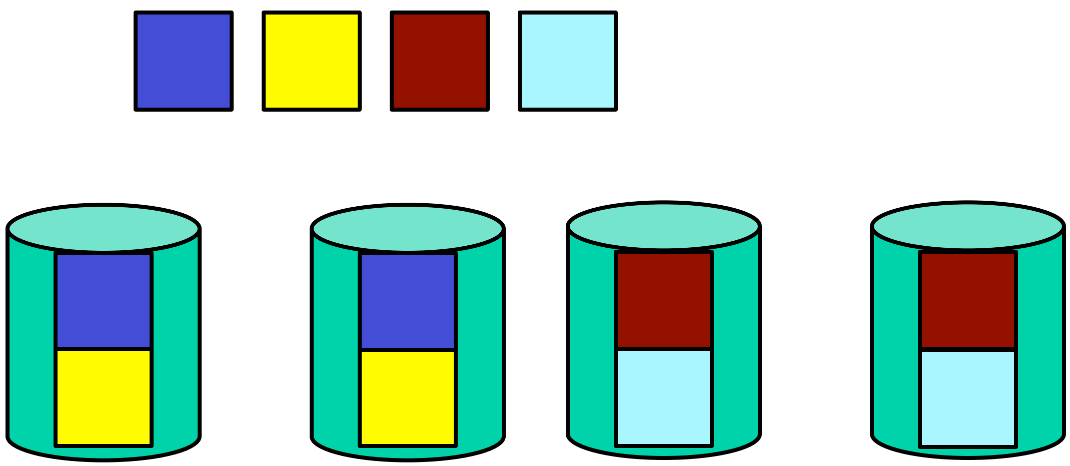


RAID 0 performance

- Large read of 100 blocks.
  - One disk: 100 * t,
  - Raid0: 100/N * t * S
  - S: slowdown. Need to wait for slowest disk to complete before return.
- Performance:
  - Large read: N/S
  - Large write: N/S
  - Large R-M-W: N/S
  - Small read: N
  - Small write: N
  - Small R-M-W: N

RAID 1: mirroring

- Structure
  - Keep a *mirrored* (shadow) copy of data
- Advantages
  - Good reliability: one disk failure OK
  - Good read performance
- Disadvantage
  - High cost: one data disk requires one parity disk



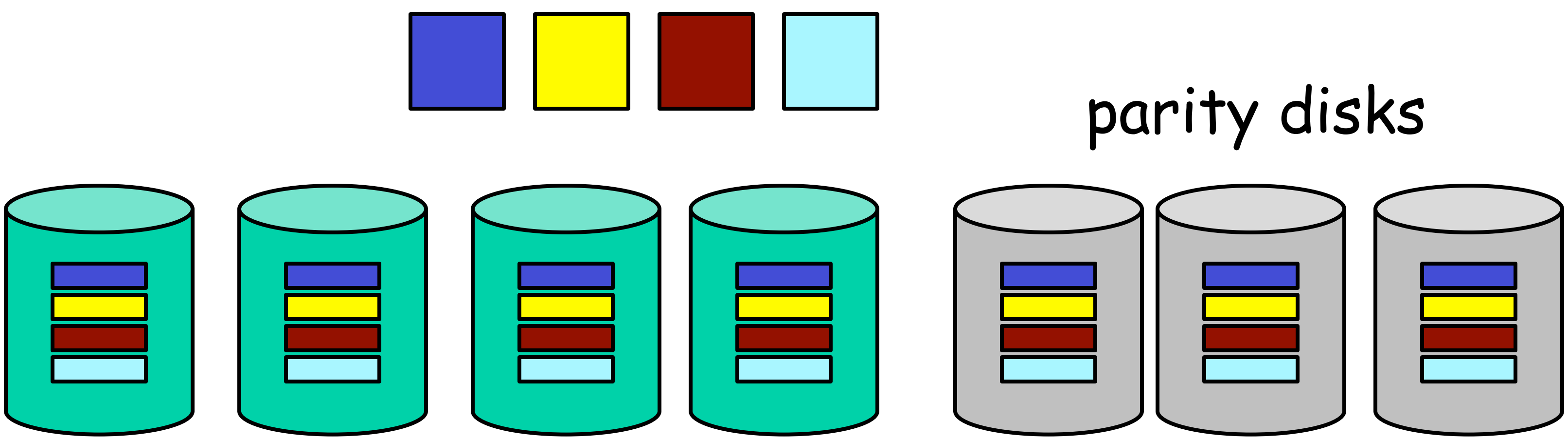
RAID 1 performance

- Cost = C/(D+C) = 1/(1+1) = 50%
- MTTF(raid) = MTTF(disk)2/(N*MTTR)

- Performance
  - Large read: N/S
  - Large write: N/2S
  - Large R-M-W: 2N/3S
    - X sectors, 2X events (X reads, X writes)
    - Speedup (w.r.t. to 1 disk) = 2X / (X/(N/S) + X/(N/2S)) = 2N/3S
  - Small read: N (no S here since only two disks)
  - Small write: N/2
  - Small R-M-W: 2N/3

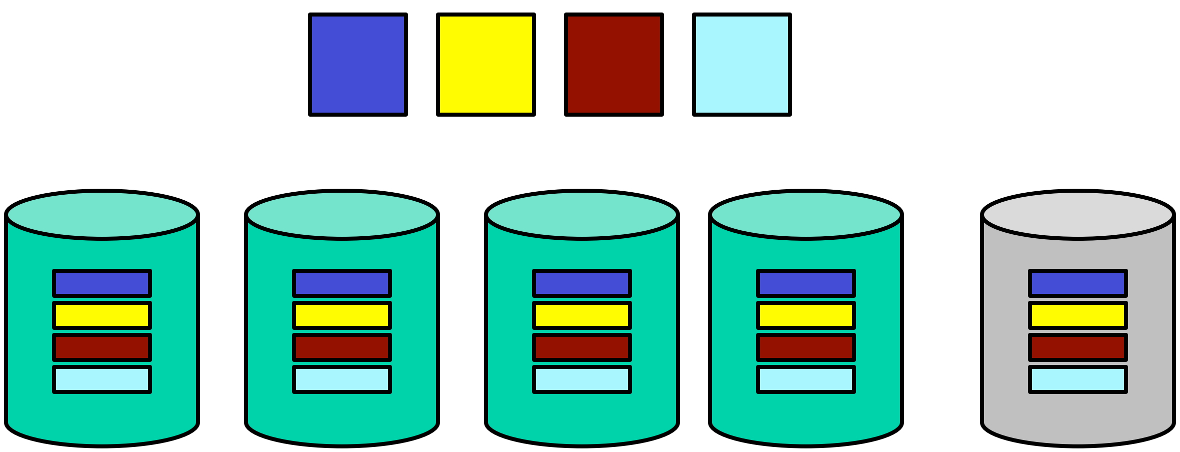
RAID 2: error-correction parity

- Structure
  - A data sector striped across data disks
  - Compute *error-correcting parity* and store in parity disks
- Advantages
  - Good reliability with higher storage utilization than mirroring
- Disadvantages
  - Unnecessary cost: disk can already detect failure
  - Poor random performance



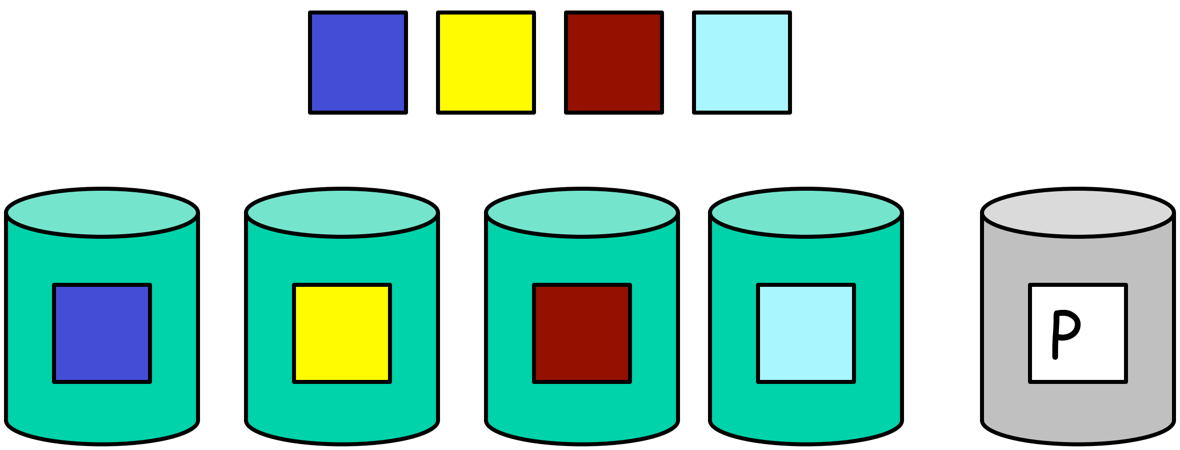
RAID 3: bit-interleaved parity

- Structure
  - Single parity disk (XOR of each stripe of a data sector)
- Advantages
  - Same reliability with one disk failure as RAID2 since disk controller can determine what disk fails
  - Higher storage utilization
- Disadvantages
  - Poor random performance



RAID 4: block-interleaved parity

- Structure
  - A set of data sectors (*parity group*) striped across data disks
- Advantages
  - Same reliability as RAID3
  - Good random read performance
- Disadvantages
  - Poor random write and read-modify-write performance

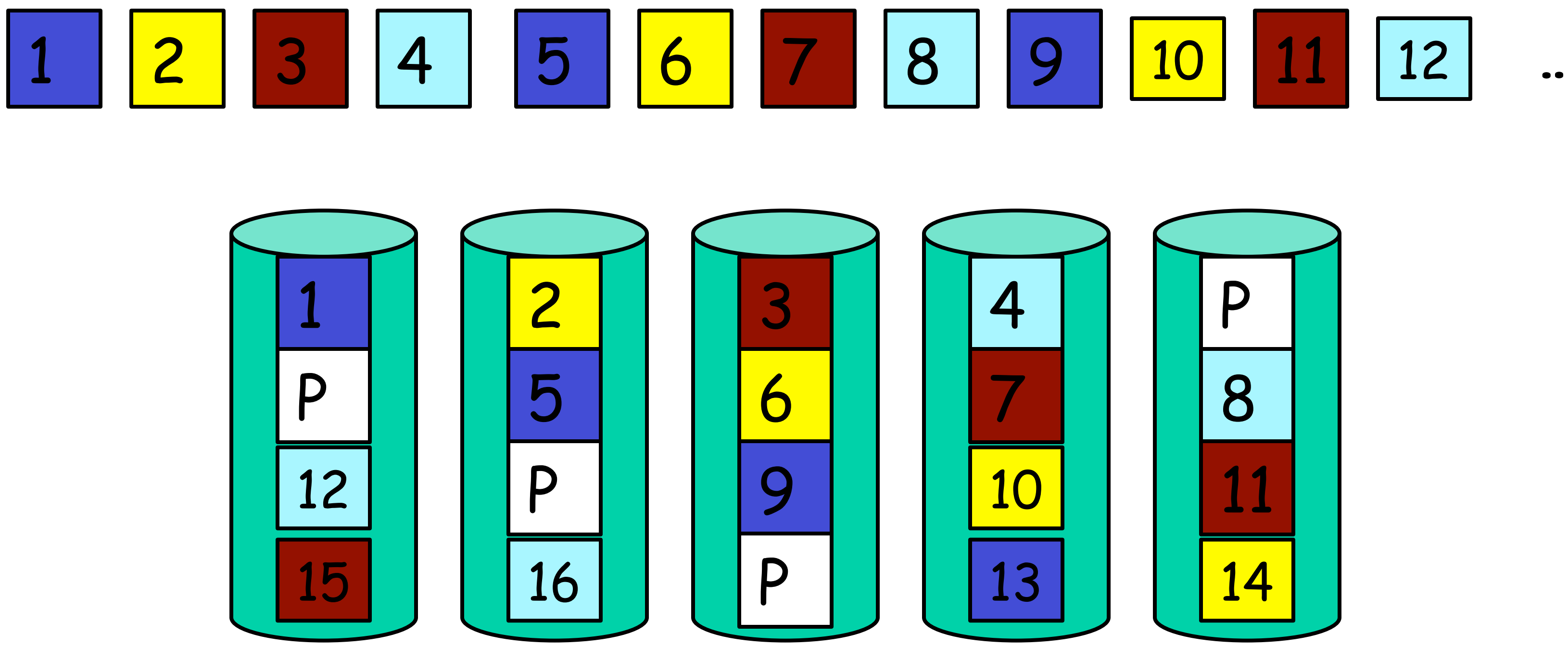


RAID 4 performance

- One parity disk (XOR of data sectors)
  - Write data disk + parity disk
  - To update parity, don't have to read all disk sectors
  - Parity = oldParity xor (changed bits) = oldParity xor newData xor oldData
- Number of groups: G = N/(D+1) = number of check disks
- Performance
  - Large read: (N-G)/S
  - Large write: (N-G)/S
  - Large R-M-W: (N-G)/S
  - Small read: N-G
  - Small write: ½*G (for each block, need a read and a write to parity disk)
    - RAID: X sectors. X/((X/1) + (X/1)) = ½
  - Small R-M-W: 1*G
    - RAID: X sectors. 2X/((X/1) + (X/1)) = 1

RAID 5: block-interleaved distributed parity

- Structure
  - Parity sectors distributed across all disks
- Advantages
  - Good performance



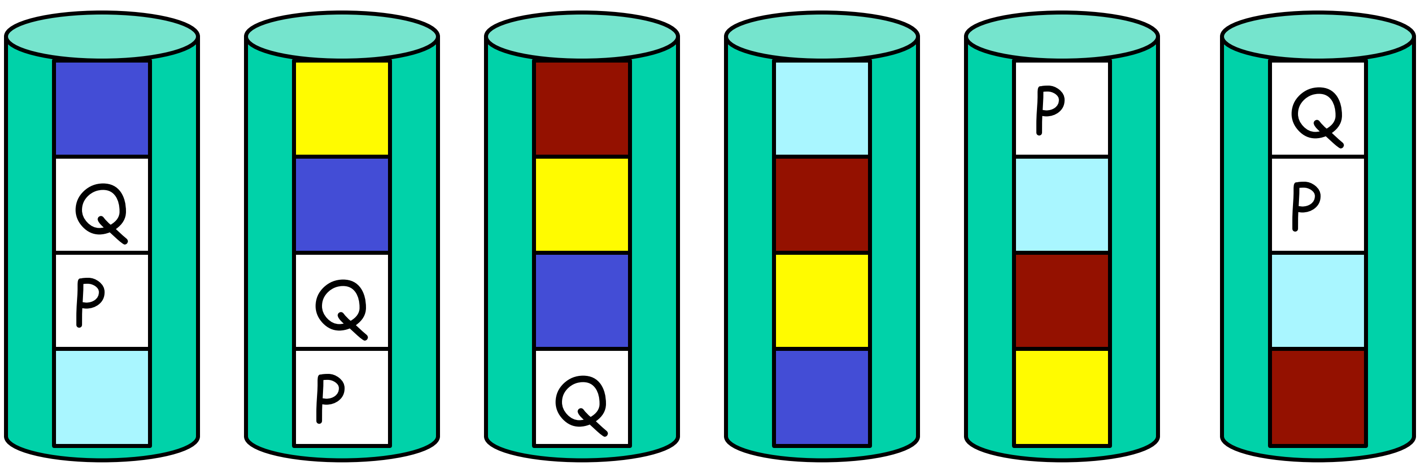
RAID 5 performance

- Same as RAID4 except no single parity disk
  - Good small write and read-modify-write performance
- Performance
  - Large read: (N-G)/S
  - Large write: (N-G)/S
  - Large R-M-W: (N-G)/S
  - Small read: N
  - Small write: N/4
    - One disk: X sectors * t.
    - Raid 5:   (X (read original) + X (read parity) + X (write original) + X (write parity)) / N * t
    - Raid5 can do 4X over all N disks
  - Small R-M-W: N/2
    - Same as small write, except read-original is not wasted.

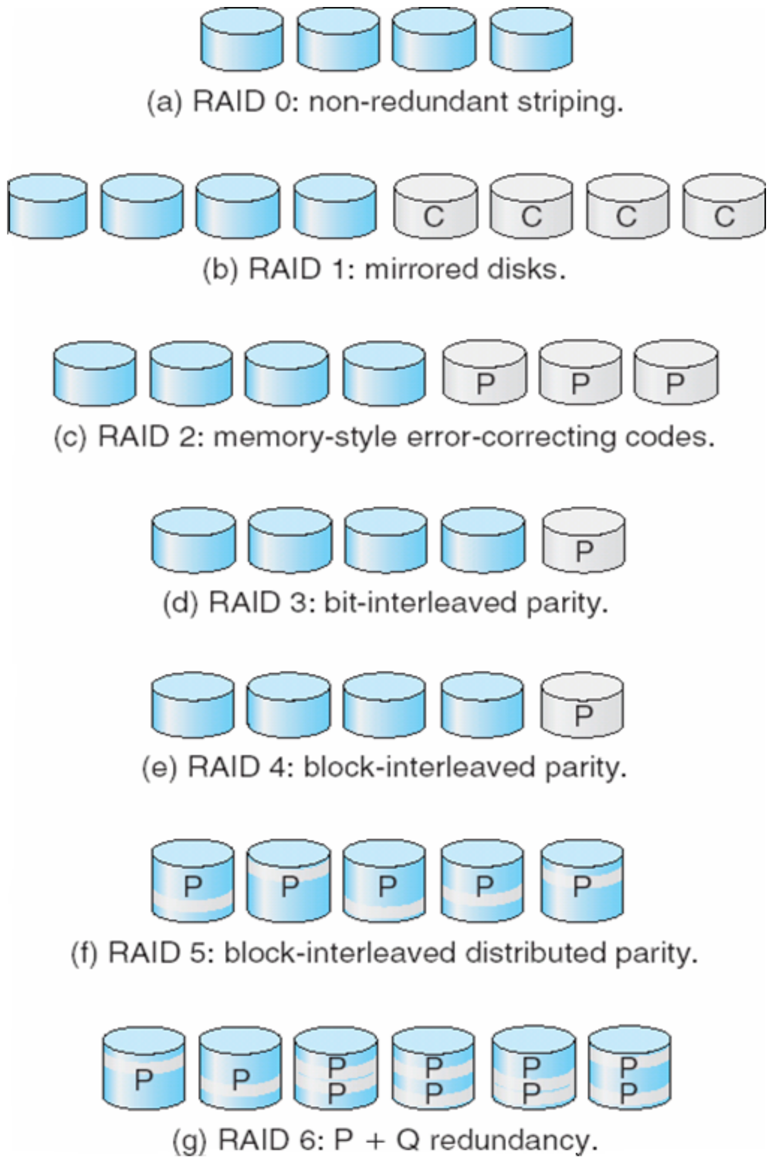
RAID6: P+Q redundancy

- Structure
  - Same as RAID 5 except using two parity sectors per parity group

- Advantages
  - Can tolerate two disk failures



RAID levels



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

(c) RAID 2: memory-style error-correcting codes.

(d) RAID 3: bit-interleaved parity.

(e) RAID 4: block-interleaved parity.

(f) RAID 5: block-interleaved distributed parity.

(g) RAID 6: P + Q redundancy.