Storage Devices

Parallel Storage Systems

2024-04-15

Prof. Dr. Michael Kuhn
michael.kuhn@ovgu.de

Parallel Computing and I/O
Institute for Intelligent Cooperating Systems
Faculty of Computer Science
Otto von Guericke University Magdeburg
https://parcio.ovgu.de

Outline

- Why are current processors increasing the core count instead of the clock rate?
  1. Higher clock rates require changing applications
  2. Increasing the clock rate also increases heat dissipation
  3. It is cheaper because cores can be interconnected more easily
  4. Additional cores increase memory throughput and graphics performance

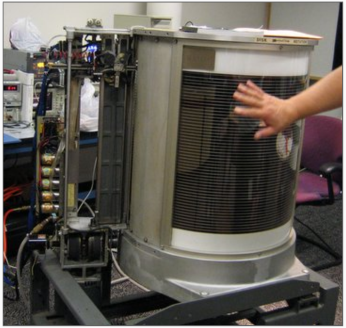
- Which architecture requires explicit message passing?
  1. Shared memory
  2. Distributed memory
  3. Shared distributed memory
  4. Non-uniform memory access

- Why are communication and I/O often responsible for performance problems?
  1. Often happen synchronously
  2. Have relatively high latency
  3. Development is relatively slow
  4. Difficult to perform efficiently

Outline

Hard Disk Drives

- The first hard disk drive in 1956
  - IBM 350 RAMAC
  - Capacity: 3.75 MB
  - Throughput: 8.8 KB/s
  - Rotational speed: 1,200 RPM
- Hard disk drive development is rather slow
  - Capacity: Factor 100 every 10 years
  - Throughput: Factor 10 every 10 years



[vnunet.com, 2006]

Hard Disk Drives...

- Different aspects improve at different rates
  - Price and density are best
- Access time has hardly improved
  - Still several milliseconds
- Throughput lags behind capacity
  - Throughput has increased by roughly 30,000-to-one

Improvement of HDD characteristics over time

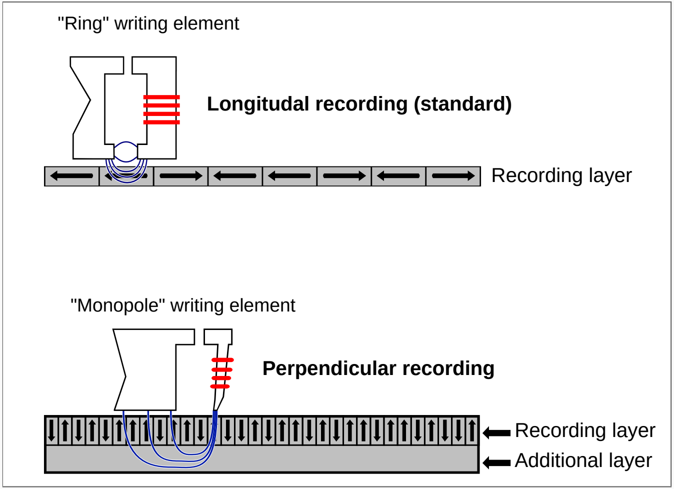
Parameter	Started with (1957)	Developed to (2019)	Improvement
Capacity (formatted)	3.75 megabytes[17]	18 terabytes (as of 2020)[18]	4.8-million-to-one[19]
Physical volume	68 cubic feet (1.9 m^3)[c][6]	2.1 cubic inches (34 cm^3)[20][d]	56,000-to-one[21]
Weight	2,000 pounds (910 kg)[6]	2.2 ounces (62 g)[20]	15,000-to-one[22]
Average access time	approx. 600 milliseconds[6]	2.5 ms to 10 ms; RW RAM dependent	about 200-to-one[23]
Price	US$9,200 per megabyte (1961)[24]	US$0.024 per gigabyte by 2020[25][26][27]	383-million-to-one[28]
Data density	2,000 bits per square inch[29]	1.3 terabits per square inch in 2015[30]	650-million-to-one[31]
Average lifespan	c. 2000 hrs MTBF[*citation needed*]	c. 2,500,000 hrs (~285 years) MTBF[32]	1250-to-one[33]

[Wikipedia, 2021a]

- Physical foundations
  - Giant magnetoresistance (GMR), tunnel magnetoresistance (TMR)
- Magnetic platters hold the data
  - Non-magnetic base material
    - Aluminium alloy, glass or ceramic
  - Coated with a magnetic layer
    - Thickness is typically 10–20 nm
  - Protective layer made of carbon
- Read-and-write heads perform accesses to data
  - Positioned above the rotating platters
  - Flying height usually tens of nanometers

Hard Disk Drives...

- Longitudinal recording requires a lot of space for individual bits
  - Perpendicular can store more data on the same area
- Heat-assisted magnetic recording
  - Lowers magnetic resistance
  - Allows higher capacity
- Shingled magnetic recording
  - Allows overlapping tracks
  - Updating data can be slow



[TyIzaeL, 2005]

Technology

- Energy consumption
  - Spinning the platters and moving the heads consumes energy
  - Filling drives with helium reduces friction, requiring less energy to spin them
- Capacity
  - Helium also reduces turbulence, allows reducing distance between platters
- Diagnostics
  - S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology)
  - Reports a multitude of parameters: Start/stop count, park count, spin-up time, defective sectors, operation time, temperature etc.

Solid-State Drives

- HDDs are being increasingly replaced by SSDs
    - Previously: MP3 players with small HDDs
    - Currently: Smartphones with flash storage
- Advantages
    - Read throughput: Higher by a factor of 15
    - Write throughput: Higher by a factor of 10
    - Latency: Lower by a factor of 100
    - Energy consumption: Lower by a factor of 1–10

SSDs…

- Disadvantages
  - Price: Higher by a factor of 5
  - Endurance: Only allow 10,000–100,000 write cycles
  - Complexity
    - Optimal access size differs for read and write accesses
    - Address translations is more complicated
    - Fast drives can overheat easily

Outline

RAID

- Storage arrays for higher capacity, throughput and reliability
- Proposed in 1988 at the University of California, Berkeley
  - Originally: Redundant Array of Inexpensive Disks
  - Today: Redundant Array of Independent Disks
- Historically, there have been five variants
  - Named using a so-called level: RAID 1–5
  - Nowadays, there are also RAID 0, RAID 6 etc.

RAID. . .

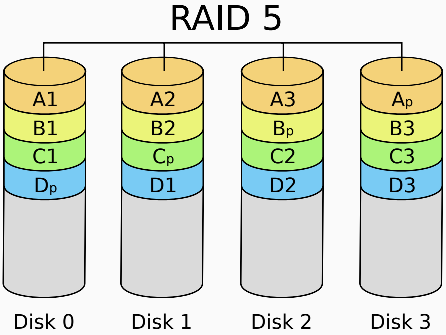
- Capacity
  - Storage arrays can be used like one big storage device
  - Problem: Organization/distribution of the data
- Throughput
  - All storage devices can contribute to the throughput
  - Problem: Devices have to be used in parallel
- Reliability
  - Data can be stored redundantly
  - Problem: Number of devices increases failure probability

RAID...

- There are multiple ways to achieve redundancy
  - Mirroring: Keeping multiple copies of all data
  - Hamming codes: All detecting and correcting errors
  - Parity: Checkums to allow correcting errors

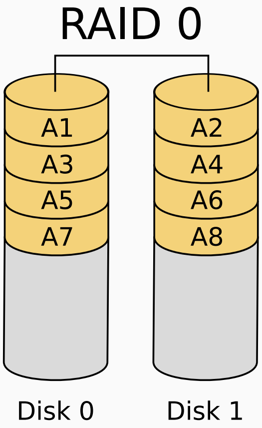
Definitions

- Strip: A single block of data (for example, A2)
- Stripe: All strips belonging together without parity (for example, A1–A3)
- Strip size: Size of a single strip (for example, 64 KB)



[Wikipedia, 2021b]

RAID 0: Striping

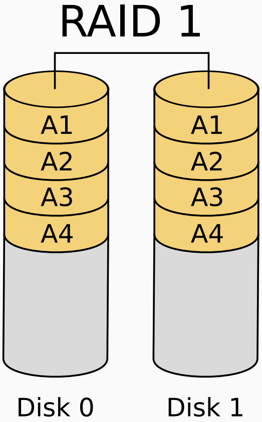


[Wikipedia, 2021b]

RAID 0...

- RAID 0: Increasing throughput by using striping
- Advantages
  - Throughput can be increased due to parallelism
  - Multiple devices can be aggregated
- Disadvantages
  - No redundancy in case of errors

RAID 1: Mirroring

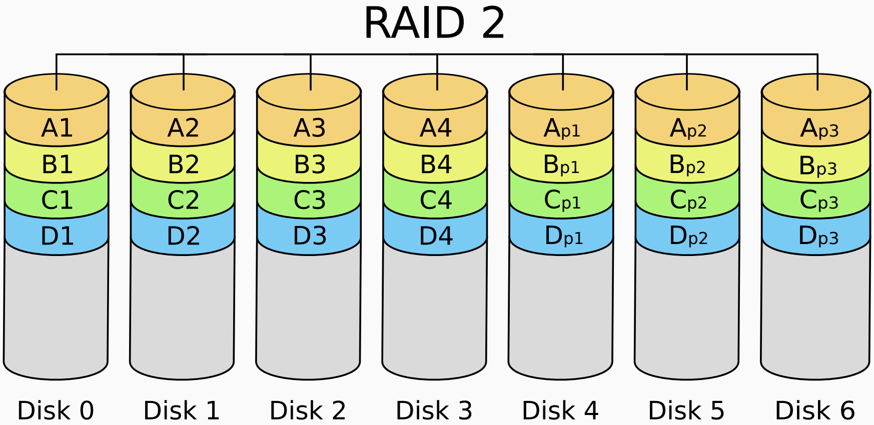


[Wikipedia, 2021b]

RAID 1...

- RAID 1: Increasing reliability using mirroring
- Advantages
    - One device can fail without losing data
    - Read throughput can be increased by reading from both devices
- Disadvantages
    - Capacity stays the same
    - Costs are doubled
    - Write throughput corresponds to that of a single device

RAID 2: Bit Striping with Hamming Codes

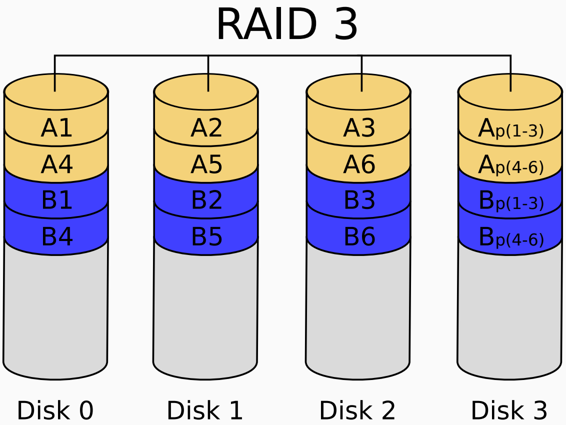


[Wikipedia, 2021b]

RAID 2...

- RAID 2: Increasing reliability by using Hamming codes
  - Four effective bits, three control bits
  - Can correct single-bit and detect single-bit and two-bit errors
- Advantages
  - Throughput can be increased due to parallelism
- Disadvantages
  - All devices active for each access due to bit-based striping
  - Spindles have to be synchronized to reduce latency
  - Overhead almost as high as with RAID 1
- RAID 2 is irrelevant in real life
  - Multi-bit errors happen very seldom
  - HDDs implement Hamming codes internally

RAID 3: Byte Striping



[Wikipedia, 2021b]

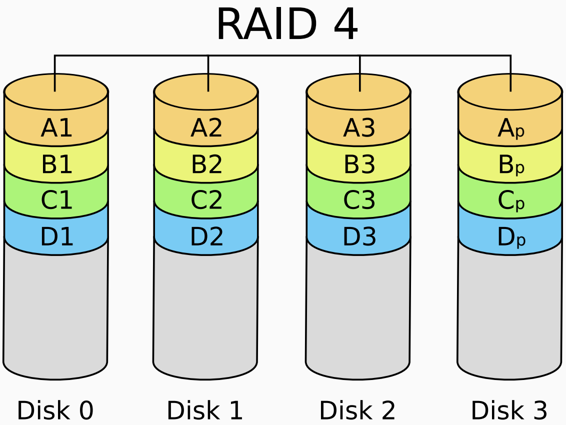
RAID 3…

- RAID 3: Increasing reliability by using parity
- Advantages
  - Throughput can be increased due to parallelism
- Disadvantages
  - All devices are active for each access due to byte-based striping
  - Spindles have to be synchronized to reduce latency

Parity Computation

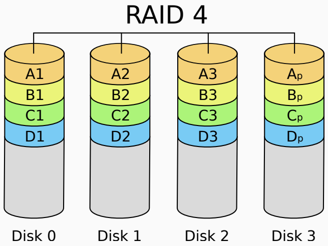
- Parity can be computed easily using XOR ($\oplus$)
  - $A \oplus B = 1 \Leftrightarrow A \neq B$
- Important property for reconstruction
  - $A \oplus B = P \Rightarrow A \oplus P = B$
- This also works for multiple inputs
  - $A \oplus B \oplus C \oplus D \oplus E = P$

RAID 4: Block Striping



[Wikipedia, 2021b]

- Does RAID 4 provide improved throughput for both read and write operations?
    1. Yes, both
    2. No, only read
    3. No, only write
    4. No, neither

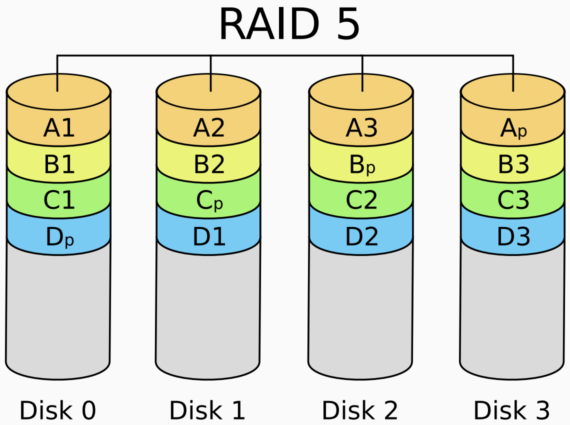


[Wikipedia, 2021b]

RAID 4...

- RAID 4: Increasing reliability by using parity
- Advantages
  - Throughput can be increased due to parallelism
- Disadvantages
  - Parity device is stressed by many accesses
  - Write throughput is limited by the parity device

RAID 5: Block Striping with Distributed Parity



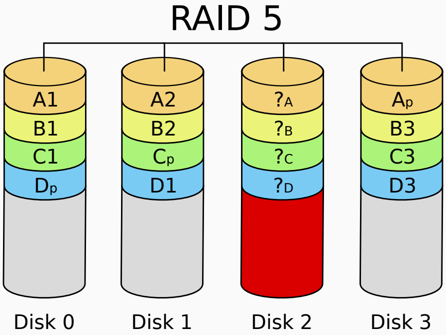
[Wikipedia, 2021b]

RAID 5...

- RAID 5: Increasing reliability by using parity
- Advantages
    - Throughput can be increased due to parallelism
    - Accesses can be processed in parallel due to block-based striping
    - Parity accesses are distributed across multiple devices

Reconstruction

- $?_A = A1 \oplus A2 \oplus A_p$
- $?_B = B1 \oplus B2 \oplus B3$
- $?_C = C1 \oplus C_p \oplus C3$
- $?_D = D_p \oplus D1 \oplus D3$

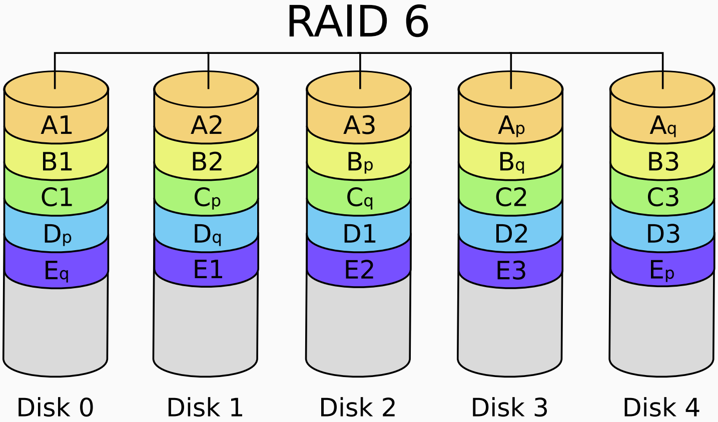


[Wikipedia, 2021b]

Reconstruction...

- Requests can be processed during array reconstruction
  - Hot spare: Spare device is connected and will be used automatically in case of failure
  - Hot swap: Spare device can be swapped at runtime
  - Cold swap: Spare device can only be swapped after system has been shut down
- Performance will be decreased during reconstruction
  - Reconstruction time can also increase significantly

RAID 6: Block Striping with Distributed Double Parity



[Wikipedia, 2021b]

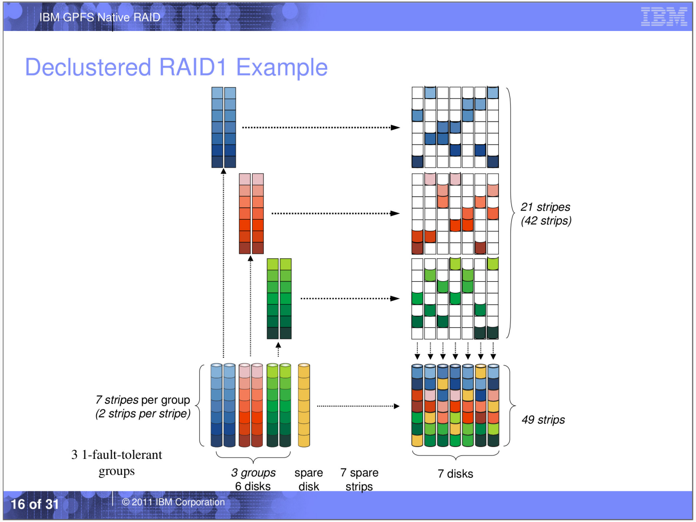
RAID 6...

- RAID 6: Increasing reliability by using parity
- Advantages
  - Reliability is increased in contrast to RAID 5
- Disadvantages
  - Additional overhead caused by second parity
  - Different implementations, some with increased computational overhead

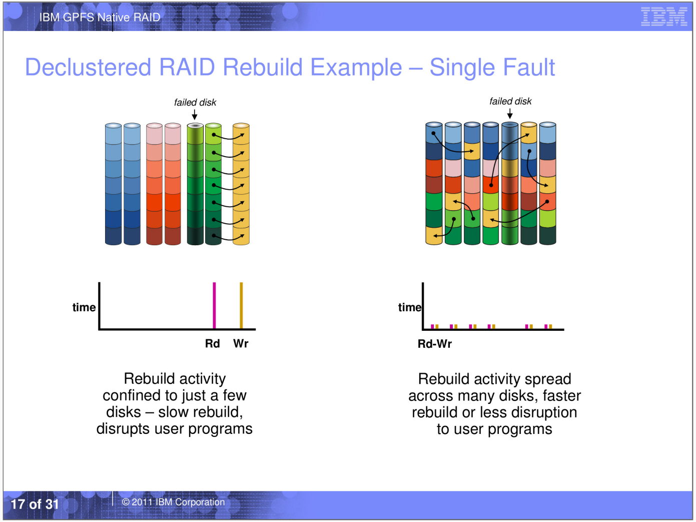
- Which RAID level would you choose for a server with 10 HDDs?
    1. RAID 0
    2. RAID 1
    3. RAID 2
    4. RAID 3
    5. RAID 4
    6. RAID 5
    7. RAID 6

Problems

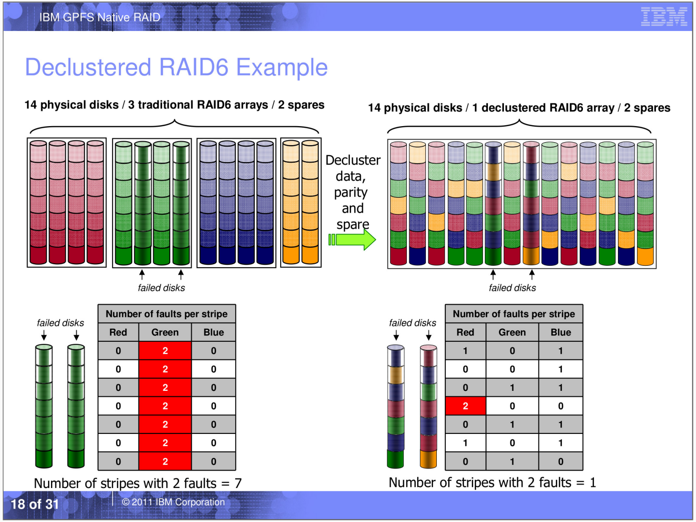
- Devices can fail partially or completely
  - Storage devices typically have roughly the same age
  - Storage devices are often from the same manufacturing batch
- Reconstruction stresses the array and takes a long time
  - Read errors can happen on the other devices
  - **Duration (30 min in 2004, 22 h in 2023)**
- Reliability suffers from inconsistencies
  - **Write Hole**



[Deenadhayalan, 2011]



[Deenadhayalan, 2011]



Number of faults per stripe		
Red	Green	Blue
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0

Number of faults per stripe		
Red	Green	Blue
1	0	1
0	0	1
0	1	1
2	0	0
0	1	1
1	0	1
0	1	0

[Deenadhayalan, 2011]

Problems...

- Write operations in a RAID
  1. Read old data
  2. Read old parity
  3. XOR old data and old parity
  4. XOR new data and result of previous step (= new parity)
  5. **Write new data**
  6. **Write new parity**

Problems…

- Write hole can occur in multiple RAID levels
  - Most popular in RAID 5 and RAID 6
- Writing of new data and new parity must happen atomically
  - Data and parity can be inconsistent otherwise
- Inconsistency will only be noticed during reconstruction
  - Cannot determine whether data or parity is correct
- Potential solutions are costly
  - Uninterruptible power supply
  - Regular synchronization of the array

Outline

Considerations

- Different performance criteria are important
  - Important to consider actual use cases and workloads
- Data throughput
  - Large amounts of data are read or written sequentially
  - Examples: Photo/video editing, numerical applications
- Request throughput
  - Small amounts of data are read or written in many small requests
  - Examples: Databases, metadata management

- Data throughput varies depending on actual hardware
  - HDDs: 150–250 MB/s
  - SSDs: 0.5–3.5 GB/s
- Request throughput can differ tremendously
  - HDDs
    - 75–100 IOPS (7,200 RPM)
    - 175–210 IOPS (15,000 RPM)
  - SSDs
    - 90,000–600,000 IOPS
- Access to partial blocks/pages can reduce performance significantly
  - Blocks and pages typically have a size of 4 KiB

- Storage arrays also need to be tuned according to workload
  - The most important parameter is the strip size
- Data throughput
  - All devices should process a single request
  - Total performance should be the sum of all devices
  - Requires small strip sizes, so all devices can contribute
- Request throughput
  - Each device should be able to process a request independently
  - High number of requests can be processed via parallelism
  - Requires larger strip sizes, so one device can handle a request

RAID…

- RAID 0
    - Pure striping
    - High data and request throughput
- RAID 1
    - Pure mirroring
    - High read performance, lower write performance
- RAID 2/3
    - Bit/byte striping
    - High data throughput, lower request throughput

- RAID 4
    - Block striping
    - High data throughput
    - High read request throughput, lower write request throughput
- RAID 5/6
    - Block striping
    - High data throughput, high request throughput

- Write operations in a RAID
    1. **Read old data**
    2. **Read old parity**
    3. XOR old data and old parity
    4. XOR new data and result of previous step (= new parity)
    5. **Write new data**
    6. **Write new parity**

- Can we circumvent steps 1 to 4?
  1. It is not possible
  2. Always write full strips
  3. Always write full stripes
  4. Compute parity from scratch

- Write operations in a RAID
  1. **Read old data**
  2. **Read old parity**
  3. XOR old data and old parity
  4. XOR new data and result of previous step (= new parity)
  5. **Write new data**
  6. **Write new parity**

- Reading requires only one device
  - Data block can be read without parity
- Writing requires at least two devices
  - Read-modify-write for both data and parity
  - Results in lower data throughput
- Performance can be improved by caching
  - Hardware RAID controllers typically have large battery-backed caches
  - Caches can help avoid partial updates

Outline

Summary

- Storage capacity and throughput increase at different rates
  - RAID improves capacity, throughput and reliability
- Performance assessment can become complex
  - Data vs. request throughput, read vs. write operations
- Historically, there were only RAID 1–5
  - Nowadays, there are also RAID 0, 6 and mixed variants
  - RAID performed by hardware controllers, software layers or within the file system
- RAID approach is used on multiple levels of abstractions
  - Storage devices, file systems, parallel distributed file systems

References

[Deenadhayalan, 2011] Deenadhayalan, V. (2011). **General Parallel File System (GPFS) Native RAID.** https://www.usenix.org/legacy/events/lisa11/tech/slides/deenadhayalan.pdf.

[TylzaeL, 2005] TylzaeL (2005). **Perpendicular Recording Diagram.** https://en.wikipedia.org/wiki/File:Perpendicular_Recording_Diagram.svg.

[vnunet.com, 2006] vnunet.com (2006). **IBM 350 RAMAC.** https://en.wikipedia.org/wiki/File:IBM_350_RAMAC.jpg. License: CC BY-SA 2.5.

[Wikipedia, 2021a] Wikipedia (2021a). **Hard disk drive.** https://en.wikipedia.org/wiki/Hard_disk_drive.

[Wikipedia, 2021b] Wikipedia (2021b). **Standard RAID levels.** https://en.wikipedia.org/wiki/Standard_RAID_levels.