Reliable storage

Problems

Storage devices develop faults

- It should be minimized the failures in storage devices and loss of data
- Failure is certain and cannot be ignored

Access to mechanical disks is slow (hard disks)

- Access Time = Translation time + Rotation Time
- More information → higher impact of storage media

Problems

Solid State Devices (SSDs) have a limited number of write operations

- 2000-3000 writes per sector for MLC (2 bits per cell)

Specific events may result in total data loss

- Fire, robbery, “energy peaks”, floods, user mistakes, attacks

May be required to distribute data in an intelligent manner

- To maximize performance
- To reduce costs

Solutions

Data backups

- Local
- Remote

Redundant Storage

- RAID
- Other: ZFS

Better storage devices, environments with higher control

- SLED (Single Large Expensive Disks)
- Enterprise Grade devices
- Temperature and Humidity Control

Infrastructures dedicated for storage

- Single policy control point

Backups

Periodic copy of data
- Snapshot of the storage state in a specific moment
- Copies will allow to set files to a previous version
- May be encrypted

Full: Complete snapshot of the data volume
- Fast recovery
- Requires a large amount of space

Differential: Differences since the last full backup
- Slower recovery, but also lower storage requirements
- Daily differential backups will grow as changes increase

Incremental: Differences since the last backup
- Even slower recovery
- Requires reconstruction of all intermediate backups since the last full
- Higher storage space efficiency

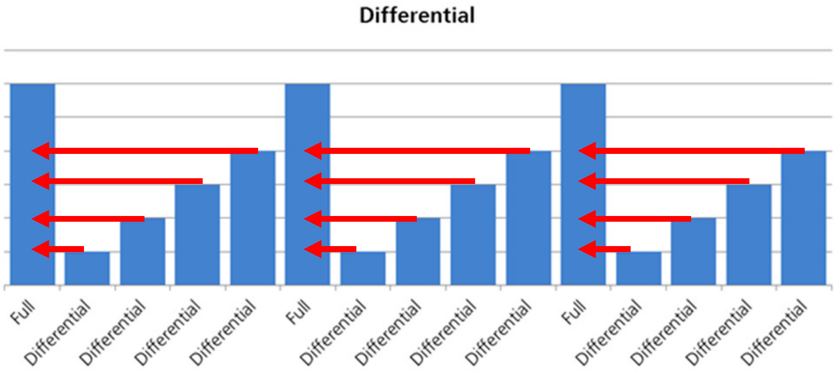
Backups

A backup is not an additional disk with data
- External or remote

It considers policies, mechanisms and processes to make, maintain and recover copies of the same data
- Should resist specific situations
- Should be used only in emergency situations
- Important to consider both the copy, storage and recovery!

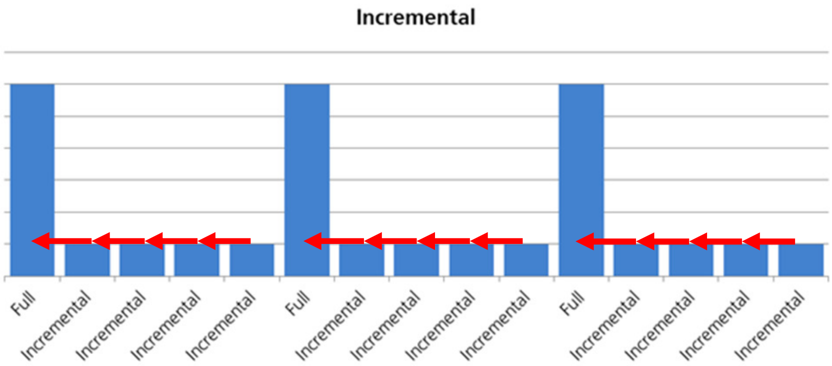
Legal framework implies a special care
- When dealing with personal data
- Frequently impose a retention policy
  - Backups should expire after some time

Backups types: Differential



http://www.teammead.co.uk/

Backups types: Incremental



http://www.teammead.co.uk/

Backups: Compression

Uses lossless compression algorithms and solutions
- Ex: ZIP

Copy only some parts of the information
- Only modified files

Deduplication
- Only store unique files/blocks
- Usually using full copy with offline deduplication
  - Of disk blocks using specific image formats
  - Of files using hard links

Backups: Levels

Applications
- Extract data from applications (e.g. mysqldump)
- Represent a consistent view of the application
  - May be required to block the application state (e.g., database changes)
- May be repeated for each individual application

Files
- Copy of individual files
- May backup any application in a filesystem
- State may be inconsistent
  - e.g., open files without data written, or applications change many files at once

Backups: Levels

Filesystem

- Internal features provided by each individual filesystem
- Creation of periodic snapshots with records of all changes or current state
- May allow the recovery of individual files, or the entire filesystem

Device Blocks

- Copy of all blocks of a storage medium
- Independent of the filesystem or operation system in use
- May be implemented by the storage infrastructure
  - Transparent and without any impact to applications

Backups: Location of data

In the same volume or in the same server

- Allow users to rapidly recover information
- Protects against changes/deletions made by users
- May not protect against hardware malfunction
  - e.g., macOS Timemachine

In a system location in the same infrastructure

- Also, with fast access time
- Protects against isolated storage failures
- Doesn't protect data against events with broader reach
  - Floods, fire, robbery
- Examples: Most enterprise storage solutions, backuppc, TimeCapsule, Borg, Kopia

Backups: Location of data

Remote (off-site)

- Implemented to a system outside the local datacenter
  - Dedicated service or through the internet
    - e.g., Amazon S3, or to servers in a dedicated datacenter
    - Encryption if recommended (or mandatory) in the case of external services!
- Implemented with specialized secure transport
  - Armored car transporting backups to a secure place
- Allow recovery even if far reaching events occur
  - Terrorism, Earthquake
- Recovery will be slower
  - Limited by the speed of a network link or the physical transport

Selecting Storage Devices

Different device grades: Enterprise vs Desktop

- Different construction quality and recovery features
- Different MTBF: Mean Time Between Failures
  - Enterprise HDD: 1.2M hours, at 45°C, working 24/7, 100% use rate (1)
  - Desktop HDD: 700K hours, at 25°C, working 8/5, 10-20% use rate(1)

Adjusted to each use case

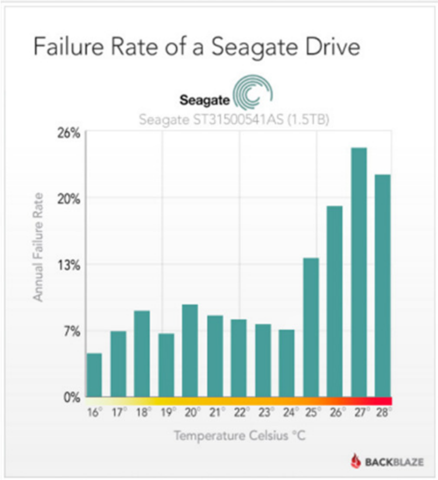
- Write intensive vs Read Intensive
- NAS vs Video vs Desktop vs Cold Storage vs Data Center
  - Differences in power consumption, reliability and performance

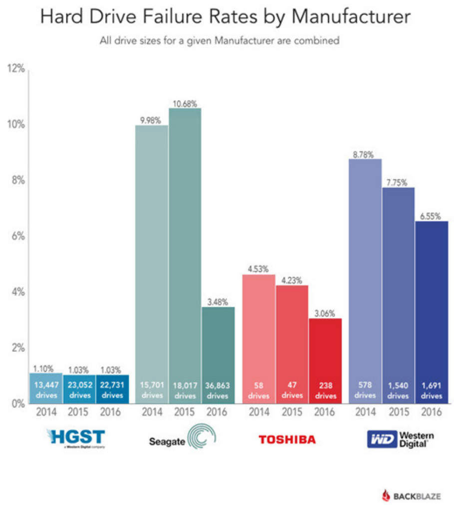
Adjusted to a specific performance level

- Tier 0: Highest performance, low capacity (PCIe NVME SLC SSD)
- Tier 1: Some performance, high capacity and availability (M2 SATA SSD)
- Tier 3: Low performance, high capacity, low price (SATA HDD)

1) Enterprise-class versus Desktop-class Hard Drives, rev 1.0, Intel, 2008

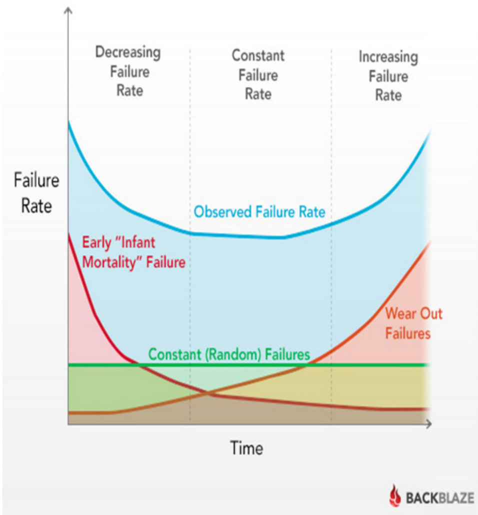
Controlled Environment and Equipment





https://www.backblaze.com/b2/hard-drive-test-data.html

Controlled Environment and Equipment



RAID: Redundant Array of Inexpensive Drives

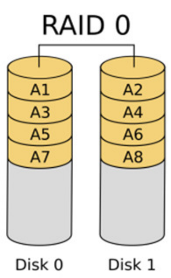
Improves the survivability of information
- Data is only lost after several devices are lost
- The number of lost devices is configurable

Low cost and efficient solution
- Can use cheap, lower quality hardware
- Can improve read and write performance

RAID doesn't replace backups
- Only tolerates the failure of a limited number of devices
- Cannot cope with user mistakes (file modification/deletion)

RAID can even increase the failure probability
- As it can be tweaked towards performance

RAID 0 (Striping)



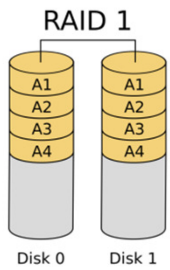
Objectives
- Speedup data access

Approach
- Access disks in parallel
- Striping
  - Data is split in small chunks (stripes)
  - Stripes are stored among all disks in a distributed manner

Advantages
- May speedup performance as a factor of the number of disks

Disadvantages
- Increases the probability of loosing data
  - If Pf is the probability of failure of a single disk, an N-disk RAID 0 volume will have a $1-(1-Pf)^N$ failure probability
- Increases the number of devices
  - At least it will double the number

RAID 1 (Mirroring)



Objectives
- Tolerate disk failures

Approach
- Data duplication (mirroring)
  - Synchronized writing
  - Distributed read from any disk with or without comparison from another disk

Advantages
- Decreases the probability of data loss
  - If Pf is the probability of failure of a single disk, the probability of failure with N disks is Pf^N

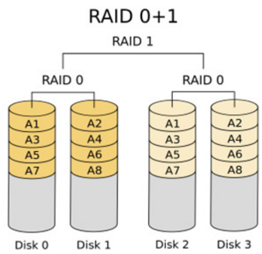
Disadvantages
- Storage inefficiency
  - Will lose at lease 50% of the total capacity
  - For 3 disks it will lose 66%... Loss is (N-1)/N
- Increase the number of devices
  - At least to the double

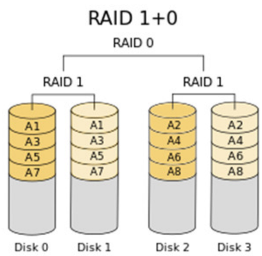
RAID 0+1 and 1+0 (Nested)

Objectives
- Benefits of RAID 0 (performance)
- Benefits of RAID 1 (resilience)

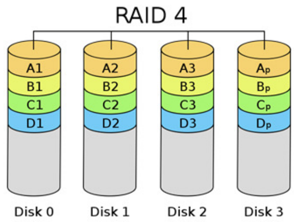
Approach
- 0+1: A RAID 1 volume using RAID 0 volumes
  - Mirroring of striped volumes
- 1+0: RAID 0 over RAID 1 volumes
  - Striping over mirrored volumes

Disadvantages
- Storage capacity waste
  - At least 50%
- Increase the number of devices





RAID 4



Objectives

- Have some resilience as RAID 1
- With a performance close to RAID 0

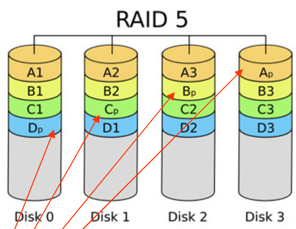
Approach

- Store data in N-1 disks
- Store parity data in an additional disk
  - Total waste is dependent on the capacity and number of disks
  - Data from any N-1 disk can be used to recreate another one

Disadvantages

- Requires at least 3 disks
  - Updating parity data is complex and will require specific hardware
  - Imposes the need to read before any write
    - Read data from existing block (e.g., C1) and from the corresponding parity disk (Cp)
    - Compare old data block with new, and change the parity block (Cp')
    - Write the new data block (C1') and the new parity block (Cp')
  - Writes must be serialized due to the existence of a parity disk
- Recovery is way more complex than with RAID 1

RAID 5



Objectives

- Similar to RAID 4
- But with higher write efficiency

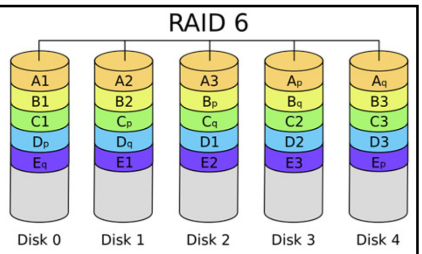
Approach

- Distribute the parity blocks among all disks
- Waste is similar to RAID 4
- Write concurrency is improved

Disadvantages

- More complex to be implemented

RAID 6



Objectives
- Improve the reliability of RAID 5

Approach
- Use 2 parity blocks, distributed among all disks
- Capacity waste will be higher than in RAID 5 (equal to 2 disks)
- Concurrency is slightly worse than with RAID 5

Advantages
- Allows the failure of two disks without data loss

Disadvantages
- Even more complex than RAID 5

NAS and SAN

NAS: Network Attached Storage
- Storage system available in the network
- Frequently created with RAID disks
- Cost: Hundreds to Thousands of Euro

SAN: Storage Area Network
- Set of systems available in a network
- Implemented distributed storage with redundancy
- Cost: Hundreds of Thousands to Millions of Euro

Advantages
- Allow centralizing the storage policies
- Provide a normalized interface, independent of the real storage
- May be used to distributed backups