# Large-Scale Data Management and Distributed Systems

# I. Introduction

Vania Marangozova
Vania.Marangozova@imag.fr
https://vmarangozova.github.io/

2025-2026

# About me

- CV
  - Professor in CS
  - Research and teaching position since 2004

- Research
  - Resource Optimisation in large-scale Distributed Platformes – clouds

# Organization of the course

- 2 complementary topics
  - Advanced Data Models - S.Maniu – 18 hours
  - Distributed Data Processing  - V. Marangozova – 18 hours
- Data Processing Frameworks
  - 9 hours of lectures + 9 hours of lab
  - alternation
- Grading
  - Graded Lab (25% of the final grade)
  - Written exam (75% of the final grade)

# Covered Topics

- The challenges of Big Data and distributed data processing
- Processing large amounts of data
  - Batch Processing
    - MapReduce and Hadoop
    - In-Memory Processing and SPARK
  - Stream Processing
    - Publish-Subscribe and Kafka
- For each : API, engine (framework) for computing, storage

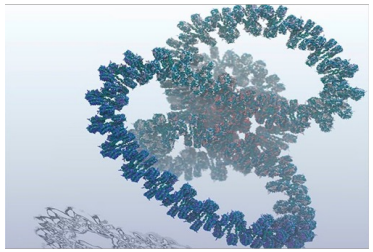
# This lecture

- Big Data Challenges
- Big Data and Distributed Computing
- The Cost of Big Data

# The Challenges of Big Data

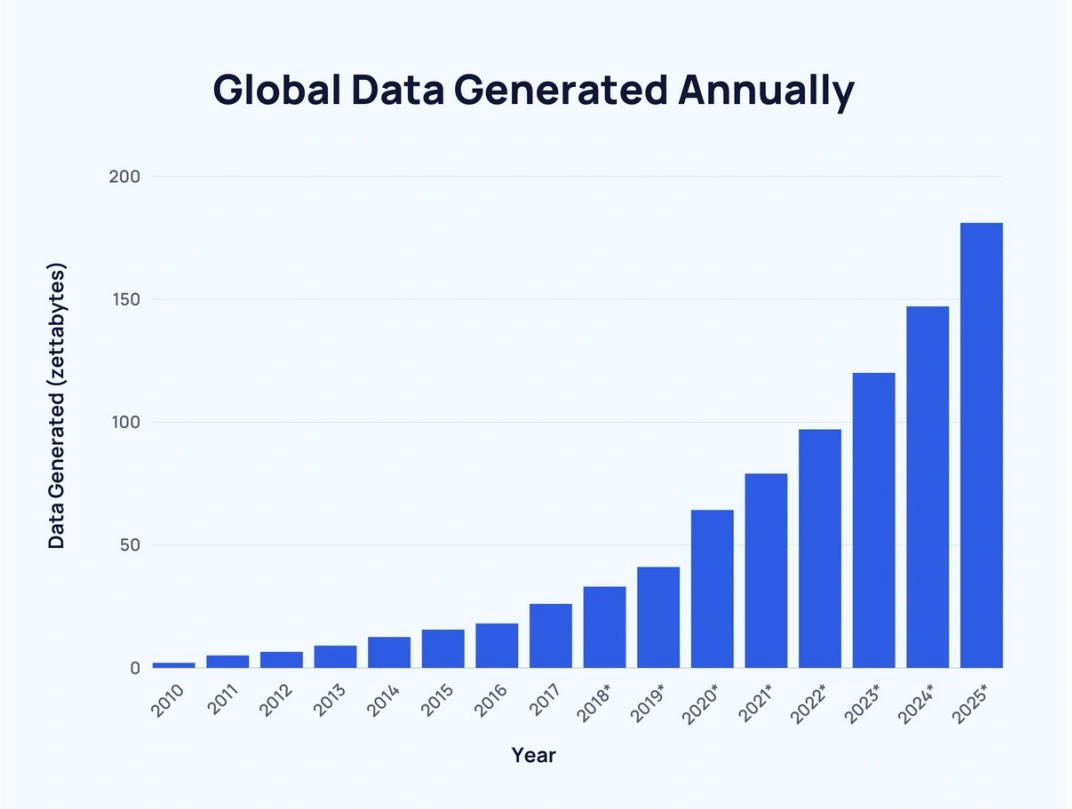
# The Data Deluge

- All activities become digitalized, data becomes value
- Various sources of data
  - Sensors
  - Social media
  - Scientific experiments
  - Industry activity
  - ...



# Some Numbers

- Every day we create
  402,74 million terabytes
  - This year we will produce
    more that 10 times more than what
    we produced in 2015
- COVID has brought +56% of
  data explosion (2020)
- ~99,000 Google search queries/second

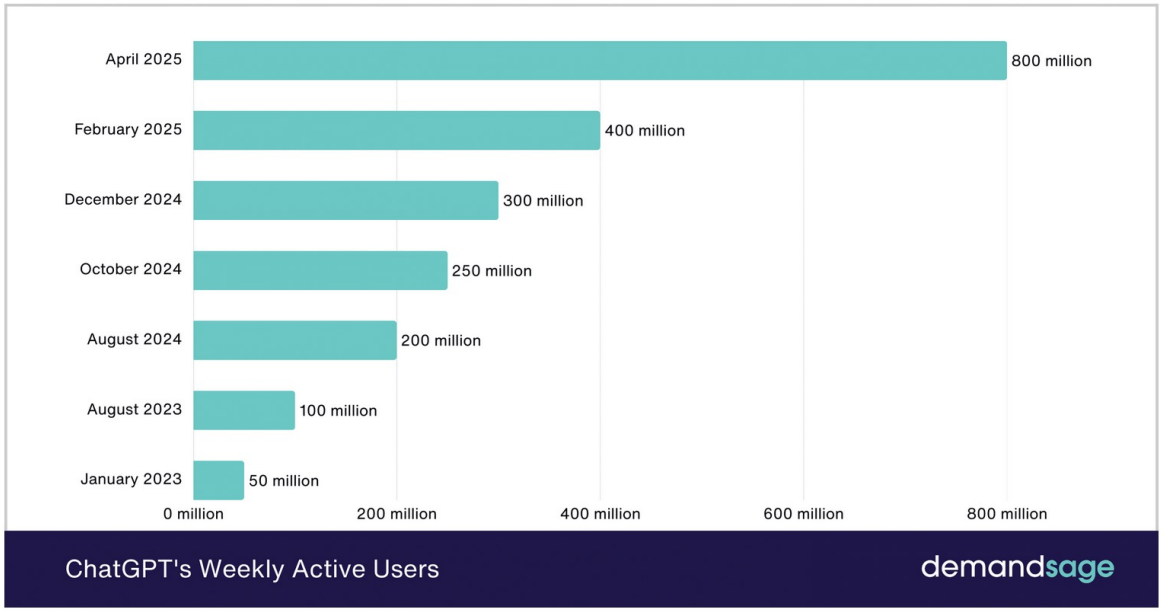



https://explodingtopics.com/blog/data-generated-per-day
Consulted 28/10/2025

# Some numbers - ChatGPT

- ChatGPT **2 billion** queries daily in 2025
  - it got x1.5 Million+ queries daily in 2024
  - 2023 - Its monthly cost is estimated at $3 million
  - 2024 – its **daily** cost is around $700,000 ($21 million)

https://www.demandsage.com/chatgpt-statistics/



# How big is xxxByte?

- https://highscalability.com/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte/ (14/11/2024)
  - $x10^0$ - 10 bytes: A single word
  - $x10^3$ - 2 Kilobytes: A Typewritten page
  - $x10^6$ - 5 Megabytes: The complete works of Shakespeare / 30 sec of TV-quality video
  - $x10^9$ - 1 Gigabyte: A pickup truck filled with paper
  - $x10^{12}$ - 2 Terabytes: An academic research library
  - $x10^{15}$ - 1 Petabyte: 5 years of EOS data (at 46 mbps)
  - $x10^{18}$ - 5 Exabytes: All words ever spoken by human beings.
  - $x10^{21}$ - Research from the University of California, San Diego reports that in 2008, Americans consumed 3.6 zettabytes of information.

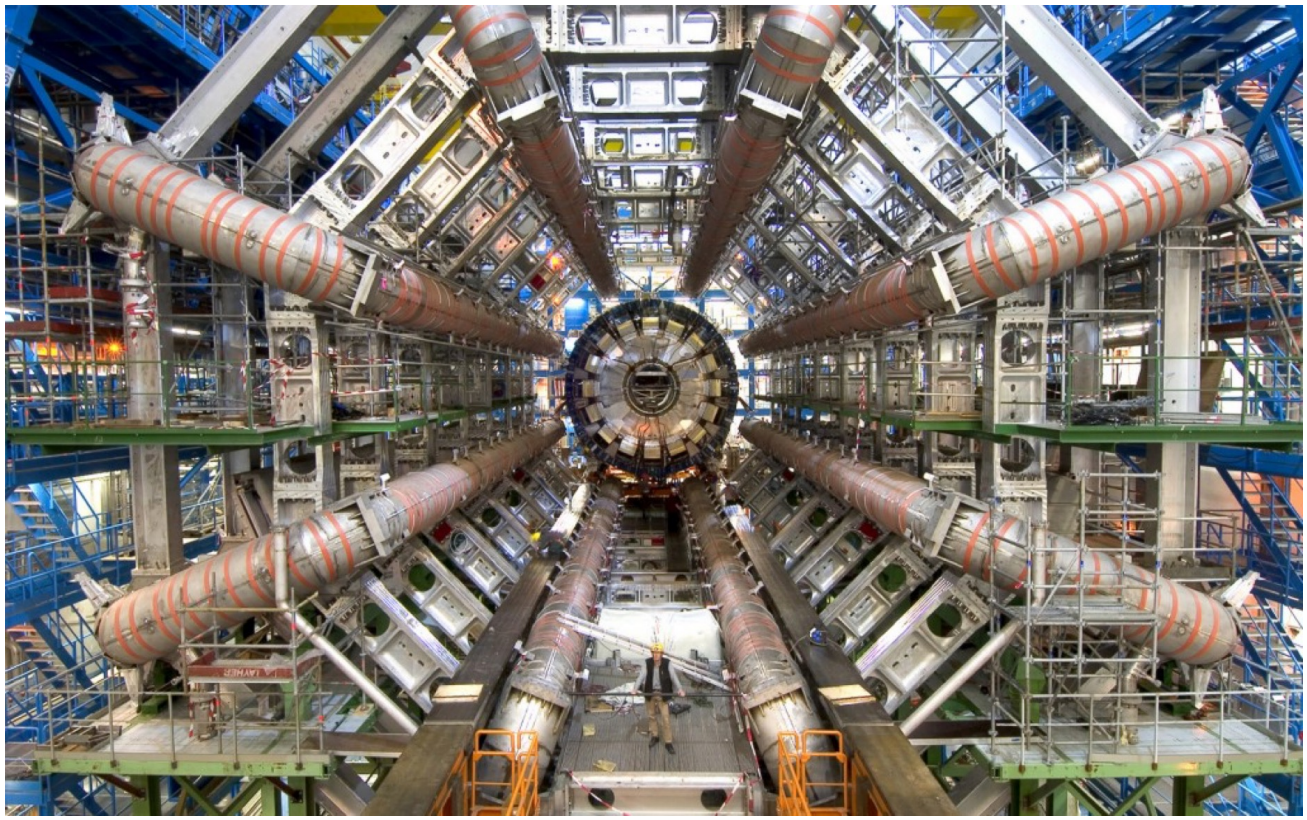
# Numbers continued…

- Video streaming accounted for nearly half of all downstream internet traffic
  - https://bloggingwizard.com/live-streaming-statistics/
- The number of devices connected to IP networks is more than three times the global population
  - Statista & CISCO report
- 18 billion connected IoT devices (https://www.demandsage.com)
  - 14 billion in 2024
  - https://explodingtopics.com/blog/iot-stats
- A new website is built every 3 seconds
  - https://www.forbes.com/advisor/business/software/website-statistics/

GLOBAL APPLICATION CATEGORY TRAFFIC SHARE

| | Rank Change | Category | Downstream | Upstream |
|---|---|---|---|---|
| 1 | - | Video Streaming | 48.9% | 19.4% |
| 2 | - | Social Networking | 19.3% | 16.6% |
| 3 | 2 | Web | 13.1% | 23.1% |
| 4 | -1 | Messaging | 6.7% | 20.4% |
| 5 | - | Gaming | 4.3% | 1.9% |
| 6 | -2 | Marketplace | 4.1% | 1.2% |
| 7 | 2 | File Sharing | 1.3% | 6.6% |
| 8 | -1 | Cloud | 1.1% | 6.7% |
| 9 | -3 | VPN and Security | 0.9% | 3.9% |
| 10 | - | Audio | 0.2% | 0.2% |

# 40 TB of data every second during an experiment at the Large Hadron Collider



# Need for Sufficient Hardware Capacity

- **To produce data**
- **To store data**
  - All the music of the world stored for $~ 500
  - Amazon EC2 instances: RAM up to 24TB, disk x10x16TB
- **To compute on data**
  - Google data-centers: more than 2.5M servers (2016), secret, "x10x1000" servers
  - Amazon capacity increases each day, the offer changes each year
- Not to forget generative AI which is a disruptive technology

**Huge opportunities for "data science"**

# Hardware Capacity

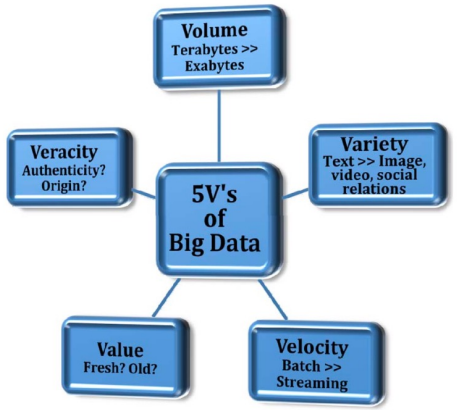
- **To produce data**
- **To store data**
  - All the music of the world stored for $~ 500
  - Amazon EC2 instances: RAM up to 24TB, disk x10x16TB
- **To compute on data**
  - Google data-centers: more than 2.5M servers (2016), secret, "x10x1000" servers
  - Amazon capacity increases each day, the offer changes each year
- Not to forget generative AI which is a disruptive technology

**Huge opportunities for "data science"**

**Huge opportunities for rethinking the cost and the usage of data**

# Big Data challenges: The V's

Source : Big Data for Modern Industry: Challenges and Trends
https://ieeexplore.ieee.org/document/7067026




- Volume: Amount of data generated
- Variety: all kinds of data are generated (text, image, voice, time series, etc.)
- Velocity: Rate at which data are produced, stored, processed
- Veracity: Noise/anomalies in data, truthfulness, bias
- Value: How do we extract/learn valuable knowledge from the data

# In this course

- Volume, Velocity
  - will see Variety

Questions to be answered:

- How to build a system and algorithms that can process huge amount of data?
- How to build a system and algorithms that can process data in a timely manner?
- (Bonus questions) How to build software that can deal with the variety of data?

# Big Data needs Distributed and Parallel Systems

# General Context

## Big data = big storage + big computation

- Needs quite a lot of resources !

## Make it possible

- **Distribution**:
  when it does not fit in a single machine
- **Paralellization**:
  when the computation takes a lot of time

## Performance criteria

- **Fast**
  - Increase the amount of data that can be processed (weak scaling)
  - Decrease the time needed to process a given amount of data (strong scaling)
- **Correct**
  - Quality of data ? Quality of computation ?
  - Failures ?
- **Economic**
  - Cost of resources
  - Environment impact

# Increase the processing power and the storage capacity

## **Vertical scaling** (scaling up)

add resources to existing nodes

- Upgrade the processor (more cores, higher frequency)
- Increase memory volume
- Increase storage volume

### Pros and Cons

☺ Performance improvement without modifying the application

☹ Limited scalability (capabilities of the hardware, cf The end of Moore's law)

☹ Expensive (non linear costs)

## **Horizontal scaling** (scaling out)

add more nodes to the system

- Cluster of commodity servers

### Pros and Cons

☹ Often requires modifying applications

☺ Less expensive (nodes can be turned off when not needed)

☺ Infinite scalability

**The solution studied in this course**

# BigData needs Large Scale Infrastructures

## which are distributed and parallel

Figure: Google Data-center

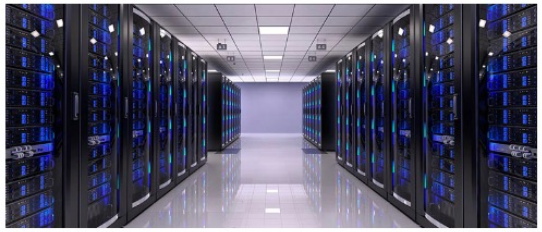


Figure: Amazon Data-center

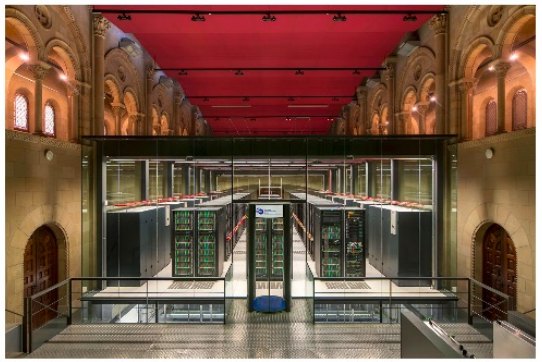


Figure: Barcelona Supercomputing Center

### Properties

- Network: latency & failures
- No global memory =
  no global state,
  each entity has its own local memory
- Coordination via message passing

# Distributed computing: Goals & Challenges (1)

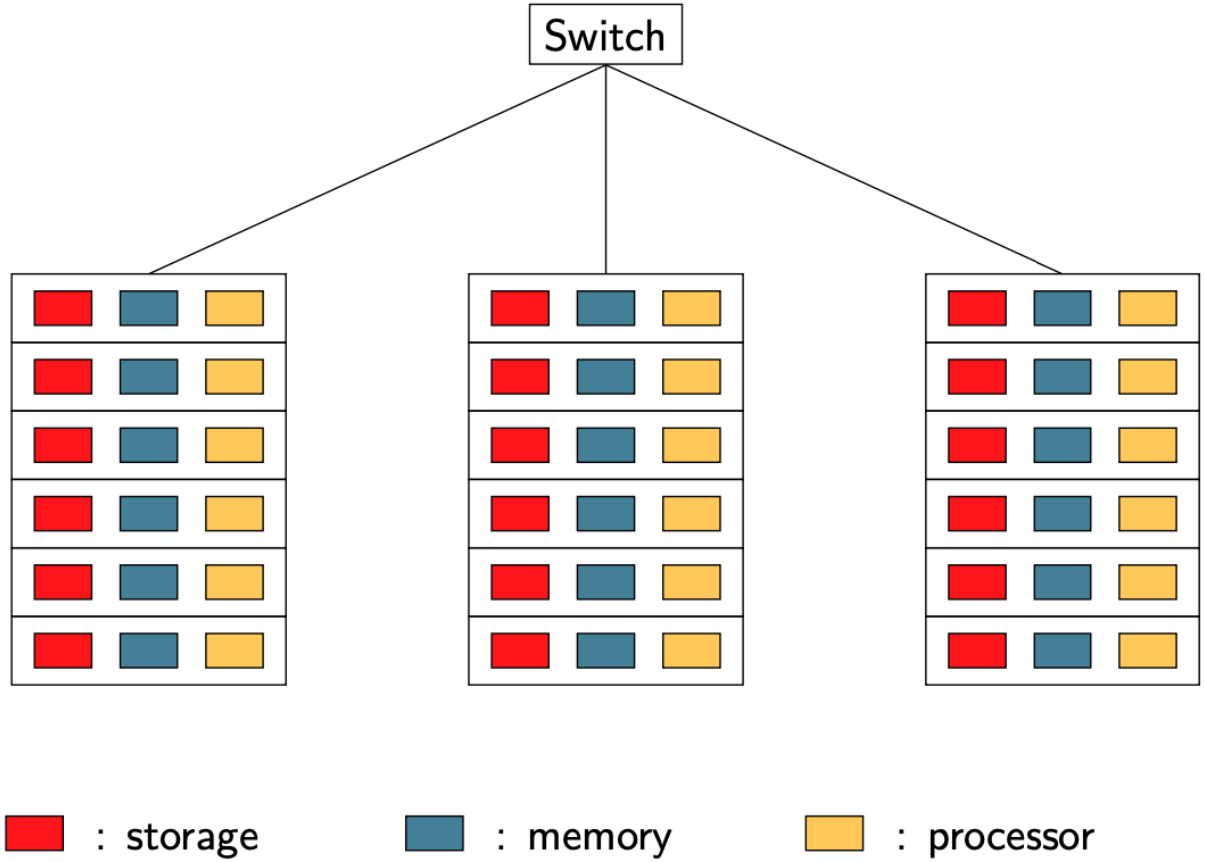
- Performance
  - How to take full advantage of the available resources?
  - Moving data is costly: how to maximize the ratio between computation and communication?
  - How to optimize te latency/throughput of requests?

# Distributed computing: Goals & Challenges (2)

- Fault tolerance
  - The more resources, the higher the probability of failure
  - MTBF (Mean Time Between Failures)
    - MTBF of one server = 3 years
    - MTBF of 1000 servers ≃ 19 hours (beware: over-simplified computation)
  - How to ensure computation completion?
  - How to ensure that results are correct?
- Programmability
  - How to provide programming models that hide the complexity of distributed computing? (while remaining efficient)
  - What high level services should be made available to ease life of programmers?

# Architecture of a Data Center

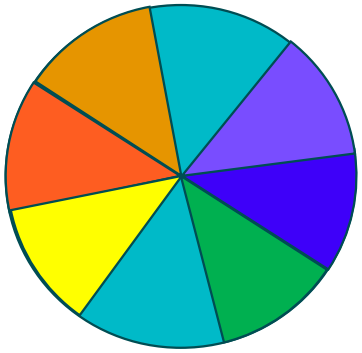
Simplified



# How to Distribute Data ?

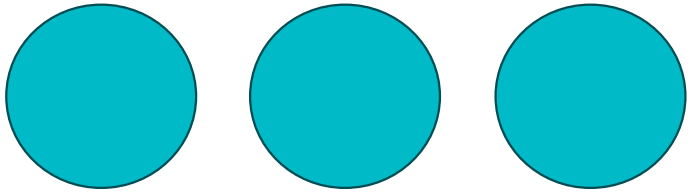
## Partitioning

- Splitting the data into partitions
- Partitions are assigned to different nodes
- Main goal: Performance
    - Partitions can be processed in parallel



## Replication

- Several nodes host a copy of the data
- Main goal: Fault tolerance
    - No data lost if one node crashes

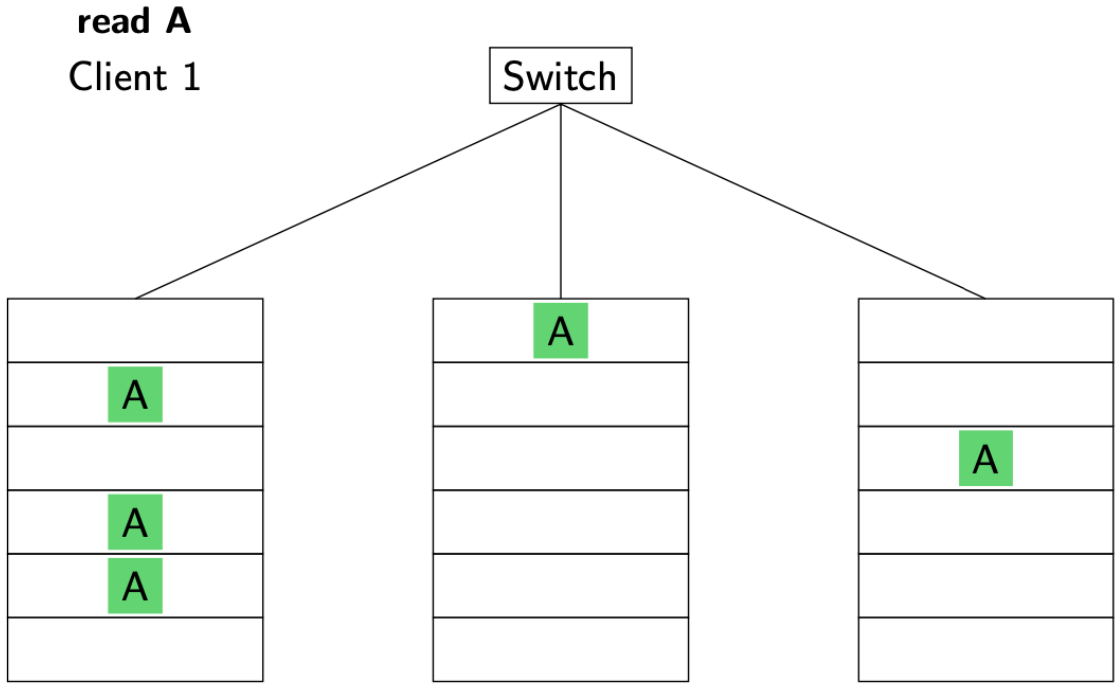


# Replication

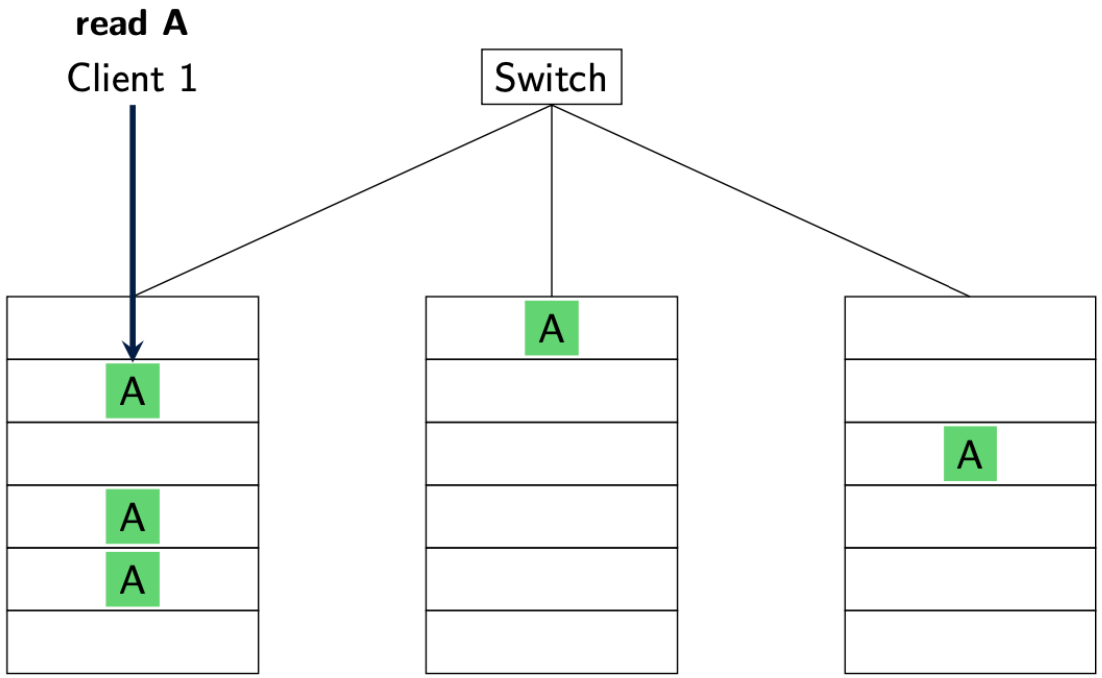
- Purposes
  - Continuing to serve requests when parts of the system fail
  - Keep data close to the users
  - Having multiple servers able to answer read requests

- Challenges
  - How to handle operations that modify data? (write operations)
    - Consistency (Consensus in a distributed system is a very difficult problem)
    - Performance

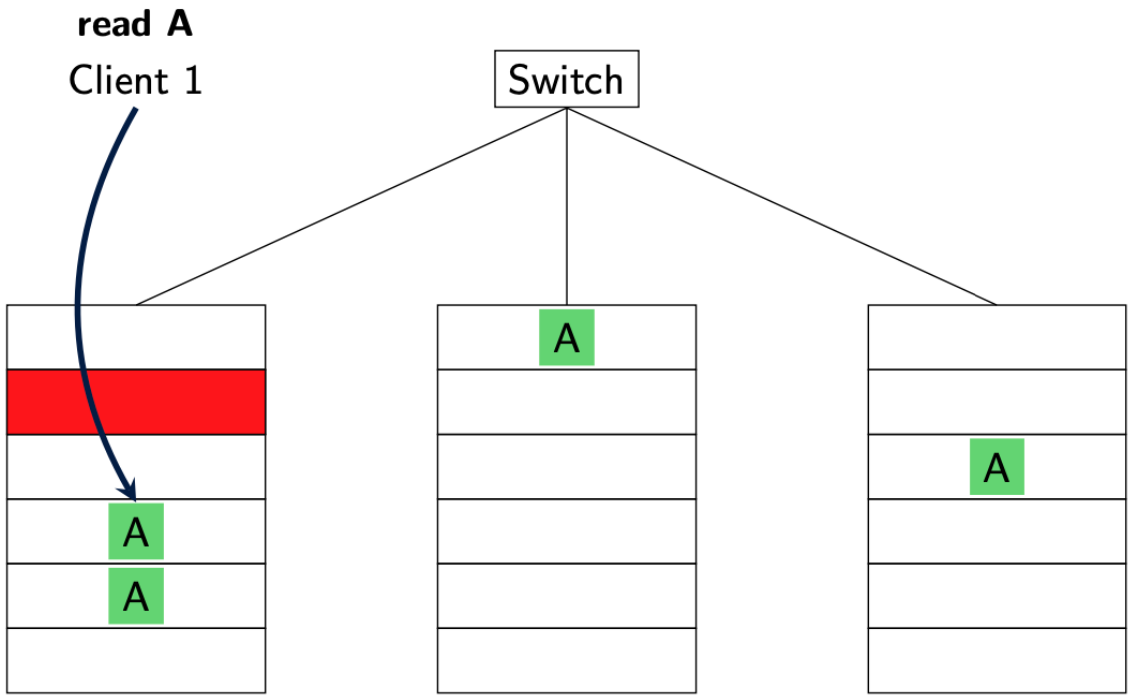
# Replication: read



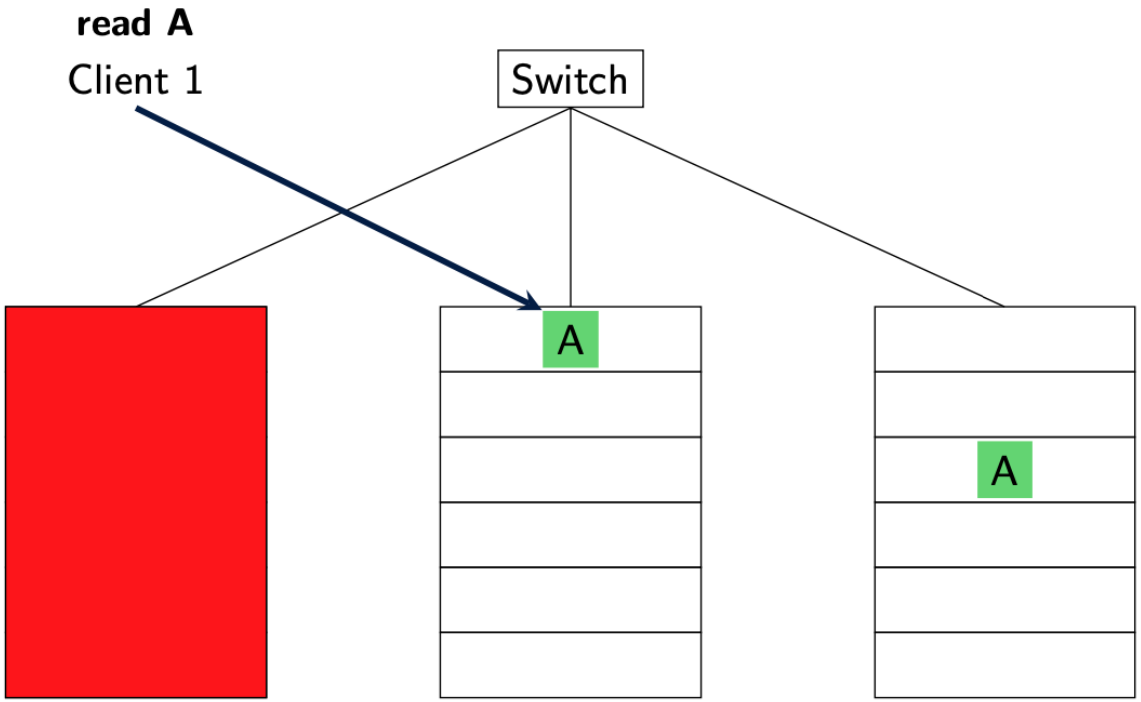
# Replication: read the closest



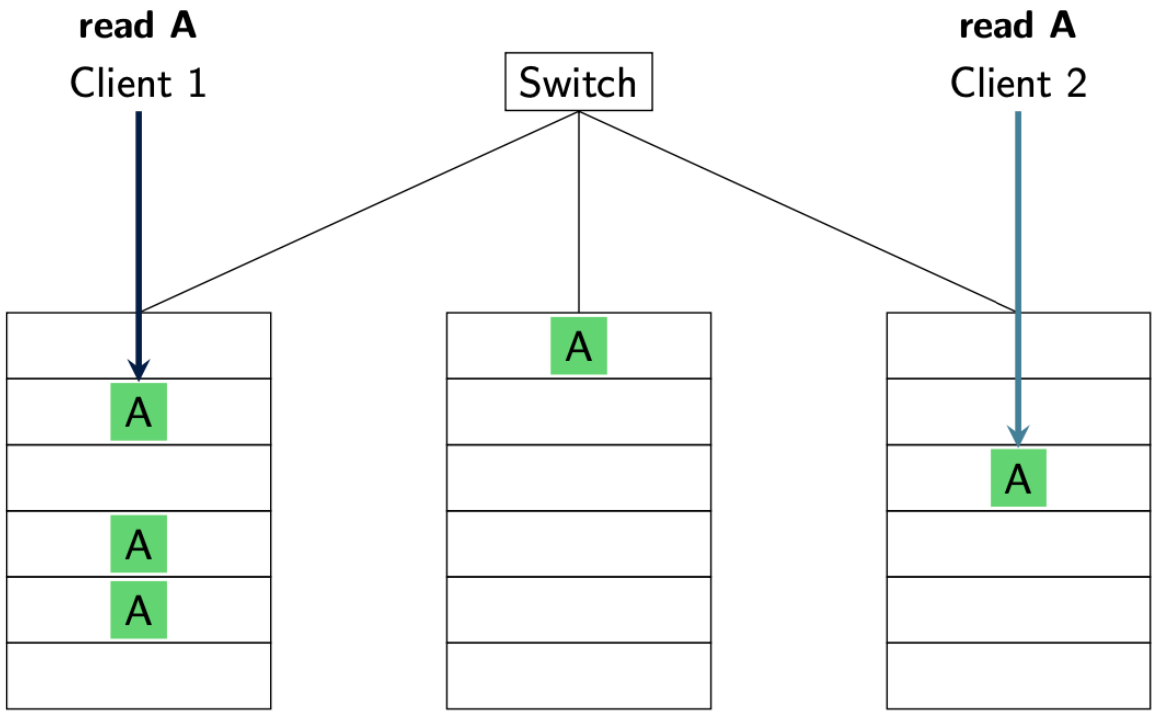
# Replication: read the closest replica



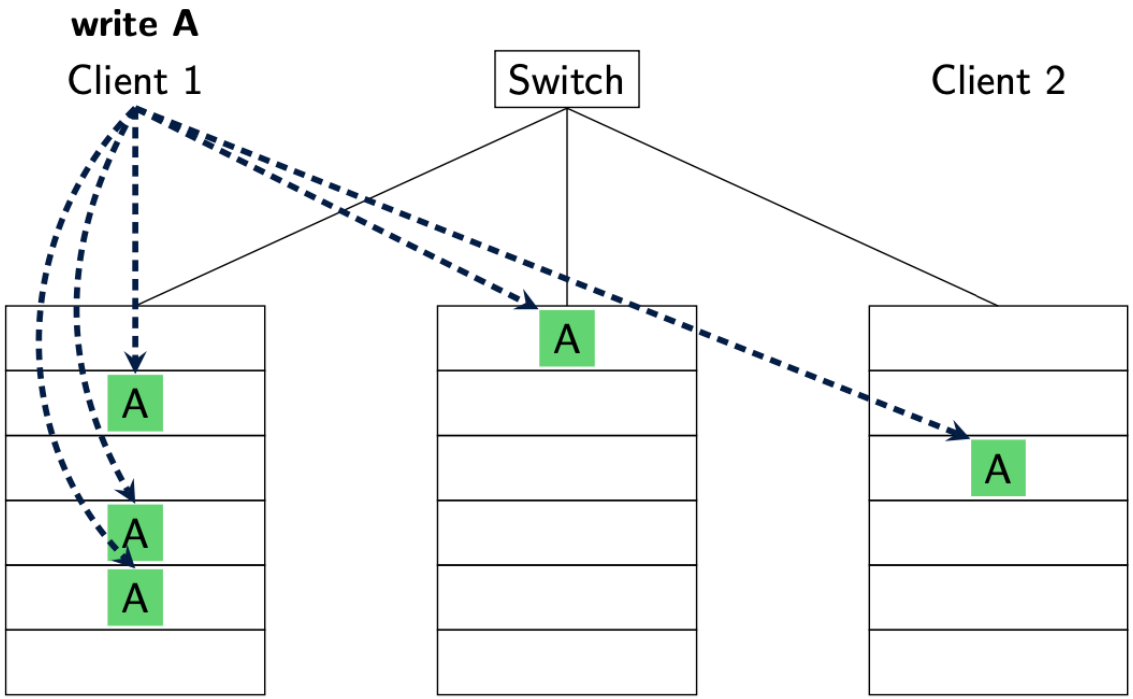
# Replication: if the closest crashes



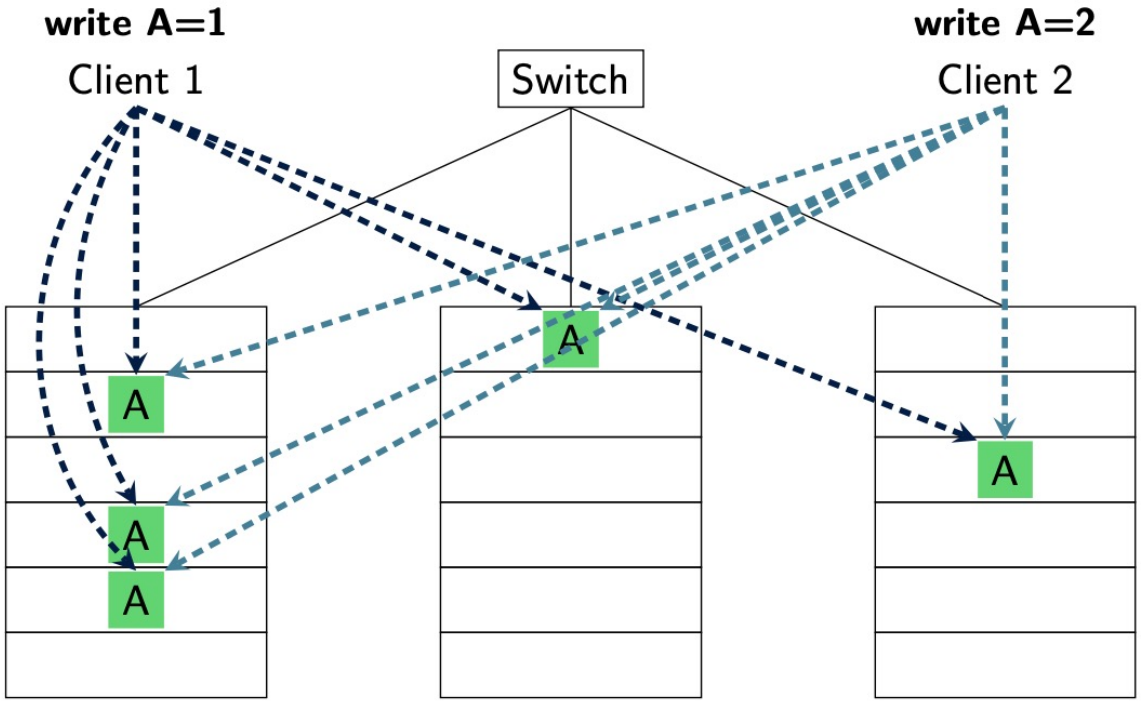
# Replication: parallel reads



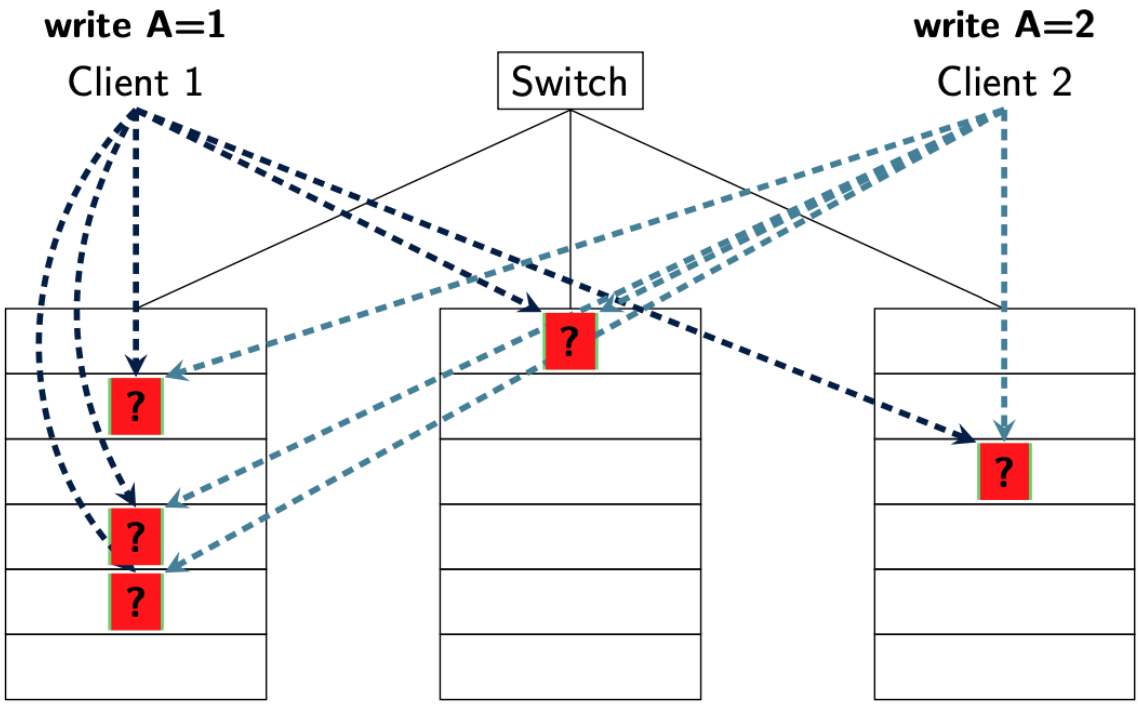
# Replication: write



# Replication: parallel writes



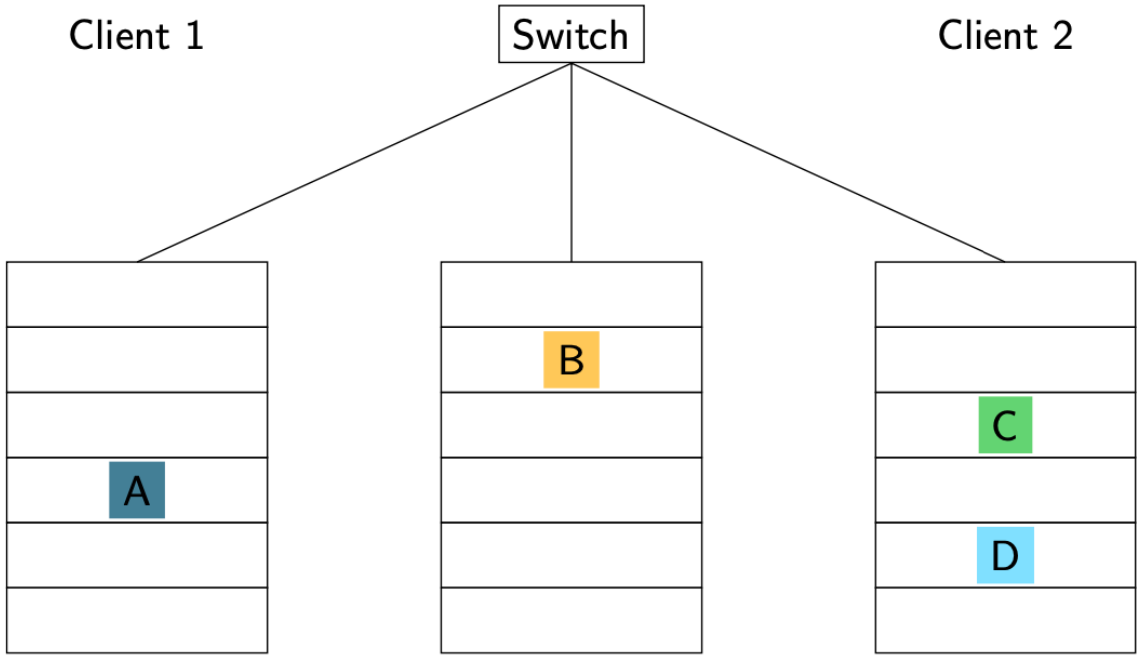
# Replication: parallel writes



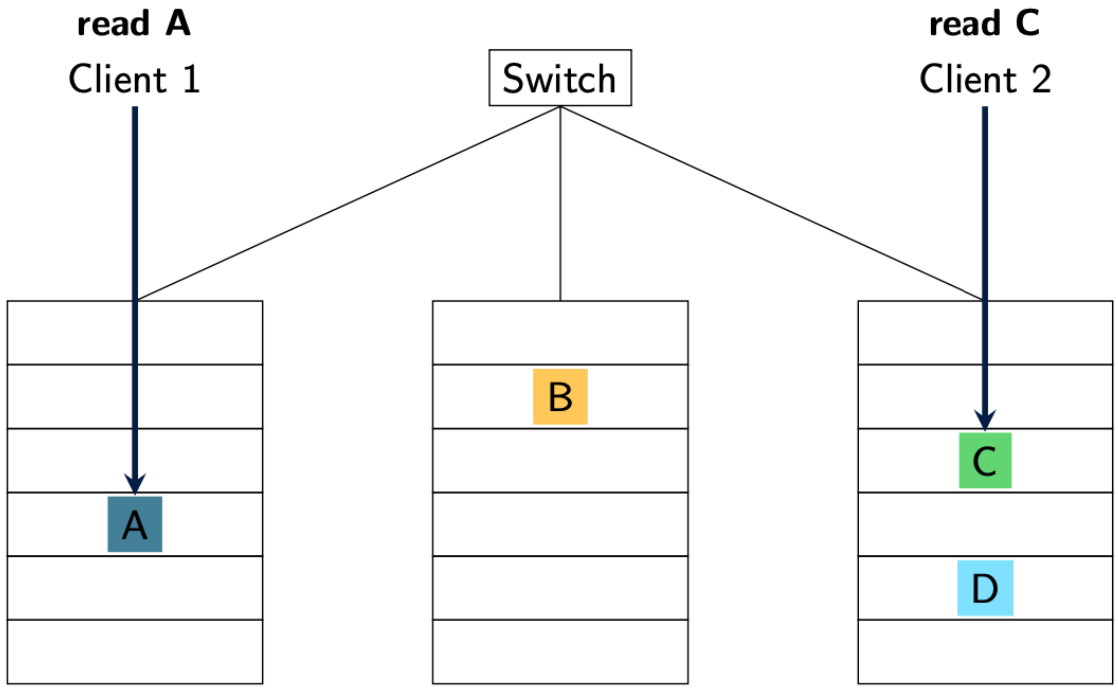
# Partitioning

- Purposes
  - Performance
    - Distributing the load over several nodes
- Challenges
  - How to partition the data?
  - Evenly distributed load (even for skewed workloads)
  - Range queries

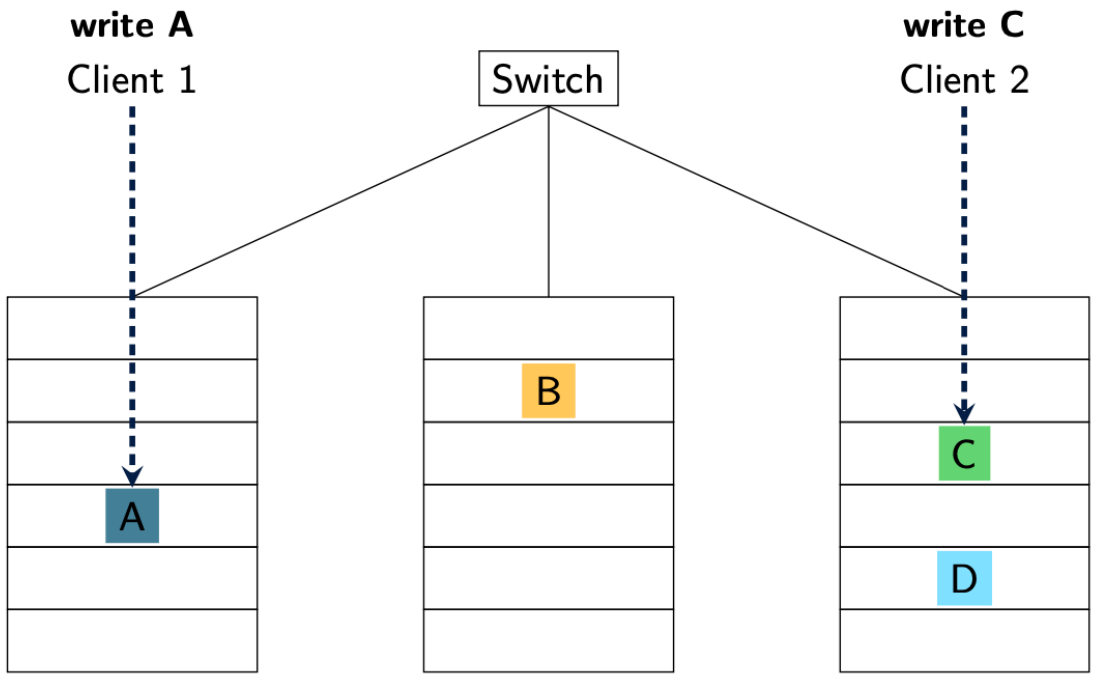
# Partitioning



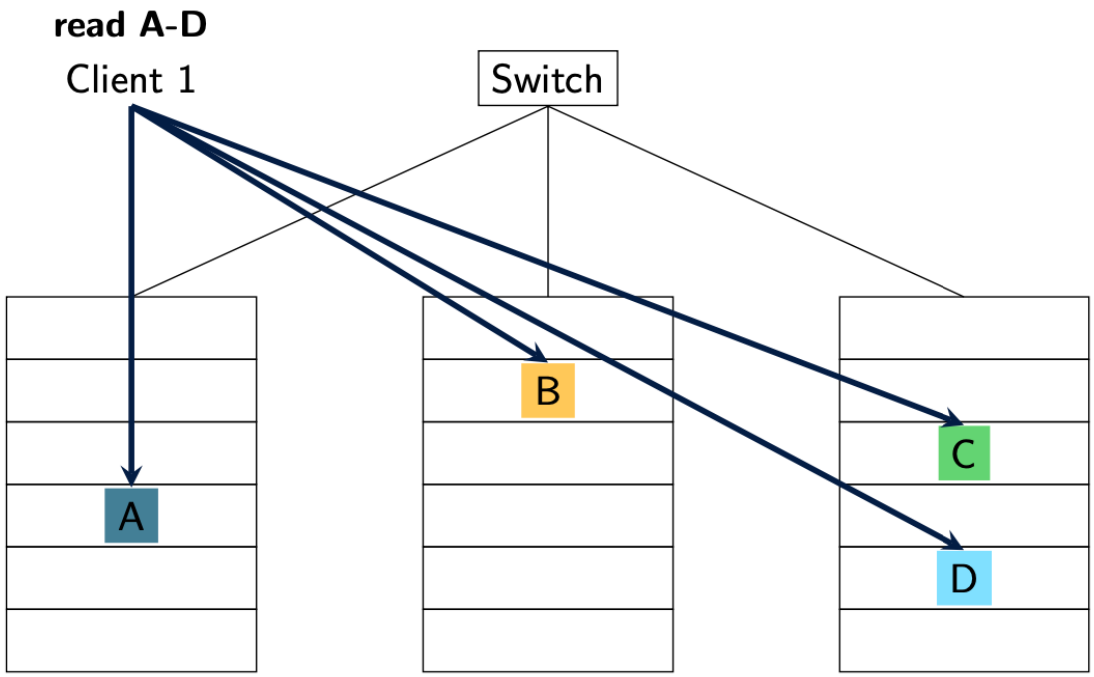
# Partitioning: parallel reads



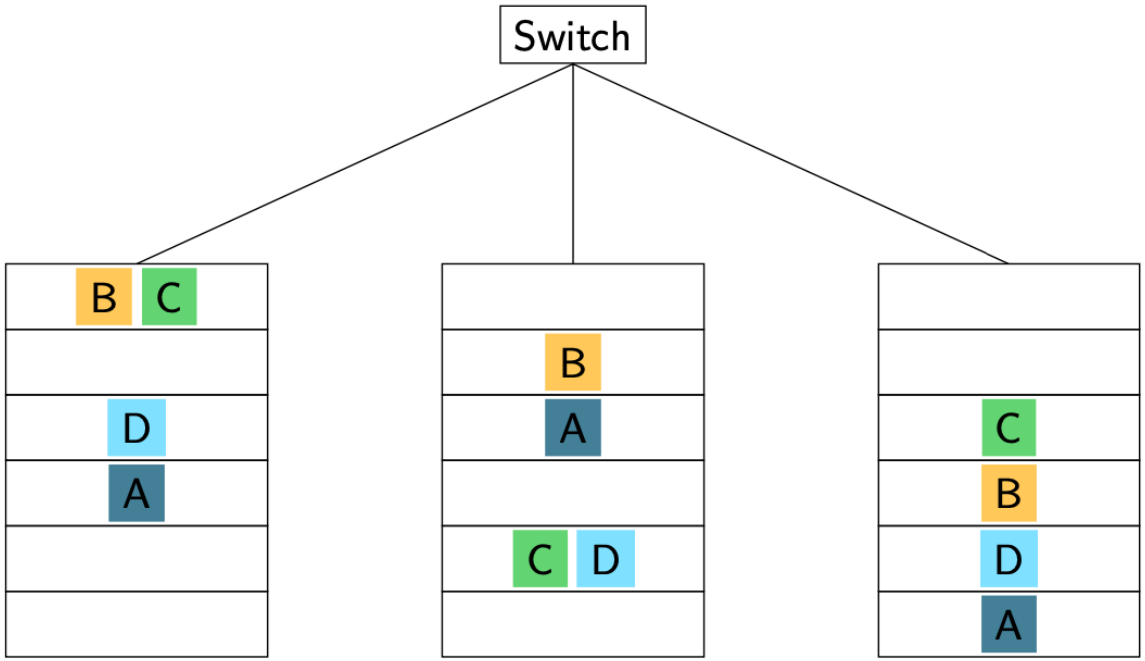
# Partitioning: parallel writes



# Partitioning: reading the whole data



# Partitioning+Replication



# BigData Runtime

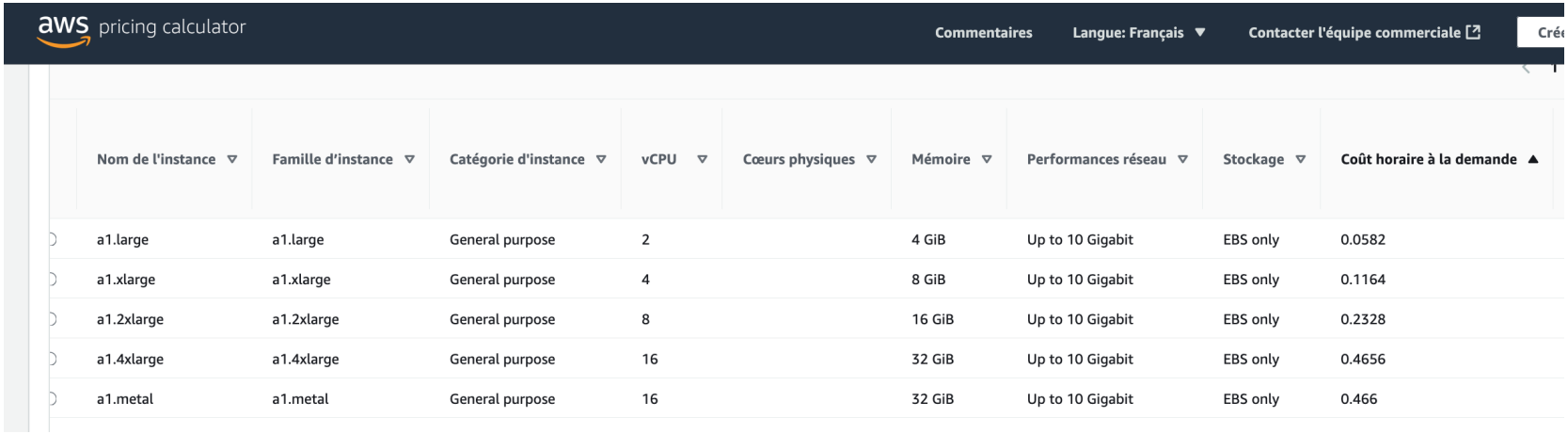
- Middleware



Replication
Partitioning
Distribution
Fault Tolerance
Monitoring
Cluster Management
Job Management
Scaling
...

# The Cost of Big Data

- User Perspective
  - Price in $ of the resources used for the BigData platform
  - Amazon, Google, … instances, storage, configuration, geographical location …

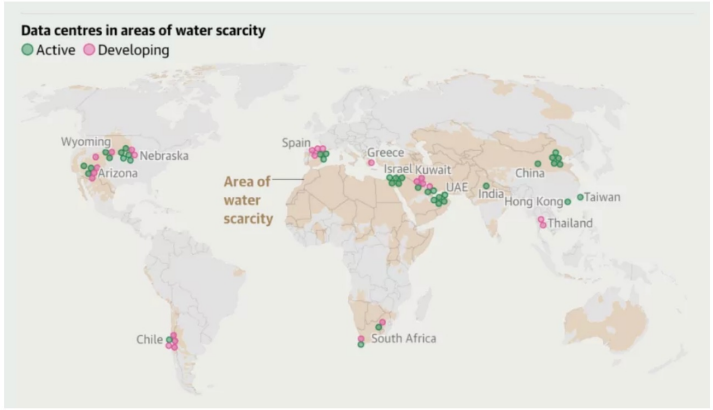



| Nom de l'instance | Famille d'instance | Catégorie d'instance | vCPU | Cœurs physiques | Mémoire | Performances réseau | Stockage | Coût horaire à la demande |
|---|---|---|---|---|---|---|---|---|
| a1.large | a1.large | General purpose | 2 | | 4 GiB | Up to 10 Gigabit | EBS only | 0.0582 |
| a1.xlarge | a1.xlarge | General purpose | 4 | | 8 GiB | Up to 10 Gigabit | EBS only | 0.1164 |
| a1.2xlarge | a1.2xlarge | General purpose | 8 | | 16 GiB | Up to 10 Gigabit | EBS only | 0.2328 |
| a1.4xlarge | a1.4xlarge | General purpose | 16 | | 32 GiB | Up to 10 Gigabit | EBS only | 0.4656 |
| a1.metal | a1.metal | General purpose | 16 | | 32 GiB | Up to 10 Gigabit | EBS only | 0.466 |

# The Cost of BigData (2)

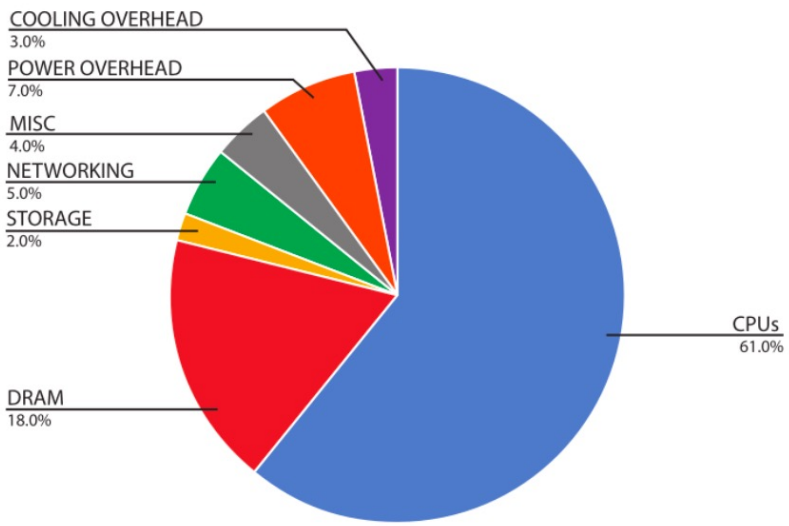
- Provider Perspective
  - Buy the HW, Operate te HW = Electricity, maintenance, cooling ...
- Environmental & societal cost (impact)

  (details in the lecture of T. Ropars on the carbon footprint of datacenters)
  - Electricity to operate the datacenter - in France 5,15 MWh/m2
    - 10,000m2 datacenter consumes as much as a 50,000 people town
  - Water to cool the datacenter
    - WEF (World Economic Forum)
      1MW datacenter may consume 25,5 million liters of water/year
      = daily consumption of 300000 people
    - A datecenter = x10MW
    - Datacenters are being developed in already desertical areas
  - Rare metals to manifacture the hardware
    - China and geo-political issues
  - Greenhouse gases emitted, equivalent in tons of CO2
  - ...




**The more data, the higher the impact !**

# Storage Cost Estimation




Peak power usage for a server

## 2017 – Storage cost in a server = 2%

- **relative share, what about the absolute value ?**
- the data explosion has lead to a significant increase in storage and data transfers
- the cloud adoption has lead to an increase in the data lakes for all industrials

## 2020 estimation

- **storage responsible for 1% of total CO2 emissions** = 330 Mt CO2/year
  - 2 tCO2/year 
  - approximatively the emissions of all cars in France for 4 years

## SSD vs HDD

- SSD pollutes more during manufacturing according to a study … by a HDD company (*Seagate)*

# Data Production Going Up

**https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/**

- IDC estimates that there will be 41.6 billion connected Internet of Things (IoT) devices, or "things," generating 79.4 zettabytes (ZB) of data in 2025
  - Cisco estimates that by the end of 2019, the IoT will generate more than 500 zettabytes per year...
- Industrial enterprises worldwide will generate a total of 4.4 Zettabytes (ZB) of data by 2030...

https://www.abiresearch.com/news-resources/chart-data/manufacturing-industry-amount-of-data-generated

- Video is responsible for over half **(53.72%)** of all global data traffic

https://explodingtopics.com/blog/data-generated-per-day

# AI Consumption

- Google paper " Measuring the environmental impact of delivering AI at Google Scale"
  https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference/?hl=en

**"*median Gemini Apps text prompt uses less energy than watching nine seconds of television (0.24 Wh) and consumes the equivalent of five drops of water (0.26 mL)"***

- 5 prompts/day limit/user

=>1 billion prompts/day ($10^9$)

  - Energy 240MWh = 8,000 homes consumption /day)
  - Water 260000 liters = 7 months of water for a family of 4

| Platform | Monthly Active Users | Daily Active Users |
|---|---|---|
| ChatGPT | 600–700 million | 160–190 million |
| Gemini | **400–450 million** | 35 million |

24 sept. 2025

# What can be done?

- The 5C's (https://www.oreilly.com/radar/the-five-cs/)

***a good data product or service:***
***useful, commercially viable, with ethical and responsible usage of data***

- **consent:** establish trust between the people who are providing data and the people who are using it with a clear agreement about what data is being collected and how that data will be used
  - **The case of cookies:**

> The death of cookies refers to the phase-out of third-party cookies. Once Google Chrome completes its phase-out of third-party cookies by the end of 2024, the death of cookies will be a reality since Chrome holds 65 percent of the browser market. Most other major browsers have already done this.

2024

> For years, digital marketing teams braced for the end of third-party cookies in the Chrome browser. Deadlines came and went. Industry strategies shifted, investments were made, and new technologies were tested. And then in April 2025, Google finally confirmed: cookies aren't going anywhere.

2025

# What can be done? (2)

- The 5C's (https://www.oreilly.com/radar/the-five-cs/)

***a good data product or service:***
***useful, commercially viable, with ethical and responsible usage of data***

- **clarity:** You can't really consent to anything unless you're told clearly what you're consenting to.
  - access by whom ?
  - for what ?
  - how (free, for sale) ?
- **consistency:** Trust requires consistency **over time**.
  You can't trust someone who is unpredictable.
  - An organization may expose user data intentionally or unintentionally
  - Frustration, anger, and surprise when users don't realize what they've agreed to
  - **The simplest usage: show you what you like/advertize what you may like**

# What can be done? (3)

- The 5C's (https://www.oreilly.com/radar/the-five-cs/)

***a good data product or service:***
***useful, commercially viable, with ethical and responsible usage of data***

- **control (and transparency):** understand and control what is happening to your data
  - Europe's General Data Protection Regulation (GDPR) requires a user's data to be provided to them at their request and removed from the system if they so desire.
- **consequences (and harm):** it is essential to ask whether the data that is being collected could cause harm to an individual or a group
  - IA bias
  - Facebook hiring policy
  - genetic research
  - ...

# What can we do? Company Level

## Company, Provider, Private Usage Levels

- **Company**: clear strategy, process, minimizing data quantity, maximize data value
- **Provider**: clear cost estimation, how much data vs how much computation
- **Private usage**:
  streaming Netflix, TikTok, voice messaging, being connected all the time has a price and it is not only in **$**, but also in terms of **planet** balance, and **health**

# What can we do? Company Level

## Datalake Management Strategy

https://www.littlefish.co.uk/news-insights/death-of-the-data-lake-data-strategy-management/

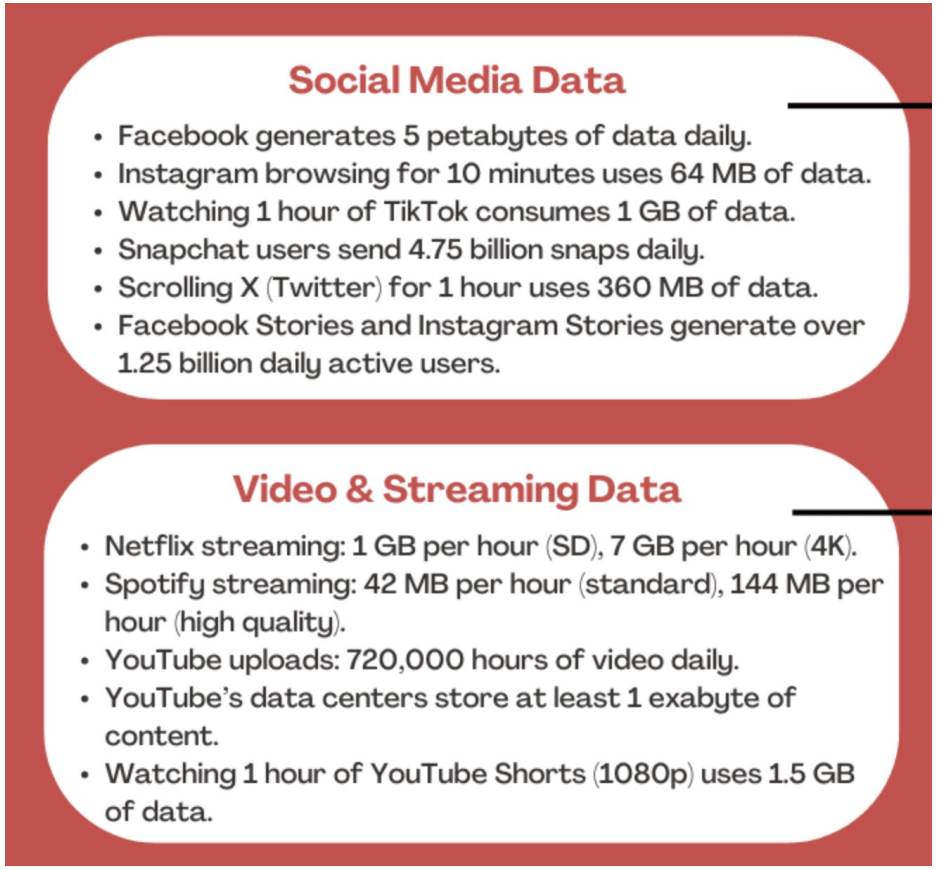
- **Less data hoarding, more data offboarding**
  - Data storage is cheap, so keep the data.
  - " The uncomfortable reality was, however, that most of these lakes had quietly become swamps; they were murky, disorganised, and hard to navigate."
- **The costs have shifted**
  - **Compliance and regulation:** Every extra dataset becomes a liability under GDPR, CCPA, and new AI governance laws
  - **Security and risk:** Every forgotten dataset is another attack surface.
  - **Operational complexity:** Engineers spend disproportionate time cleaning data.
- **2. Noise drowns out signal**
  - the sheer mass of irrelevant data slows data-led decision-making rather than enabling it
- **3. AI raises the stakes on quality**
  - AI makes bad data bigger. Feed an LLM incomplete, biased, or low-quality data, and it will confidently amplify errors at scale.

# What can we do? Provider Level

- Reliable metrics on
  - data quantities,
  - data storage and
  - data usage
- Examples
  - How much of the data has been accessed for a given computation?
  - Which part of the data has been extensively used? Used only once?
  - ...

# What can we do? Private Level

- Be aware
  - streaming Netflix,
    TikTok,
    voice messaging,
    being connected all the time
    has a price and
    it is not only in **$**,
    but also in terms of **planet**
    balance,
    and **health**




1000 users, with 1H Netflix/day =>
for 1 year , 9 round-trip flights Paris-New York

According to Netflix, the average carbon footprint of one hour of streaming in Europe is approximately 55 gCO2e (grams of carbon dioxide equivalents) - the same as driving about 300 metres in a car.

https://impactco2.fr/outils/comparateur