
CERN

Storing data on-premise or in the cloud?

Alberto Pace (alberto.pace@cern.ch)

CERN, Geneva, Switzerland

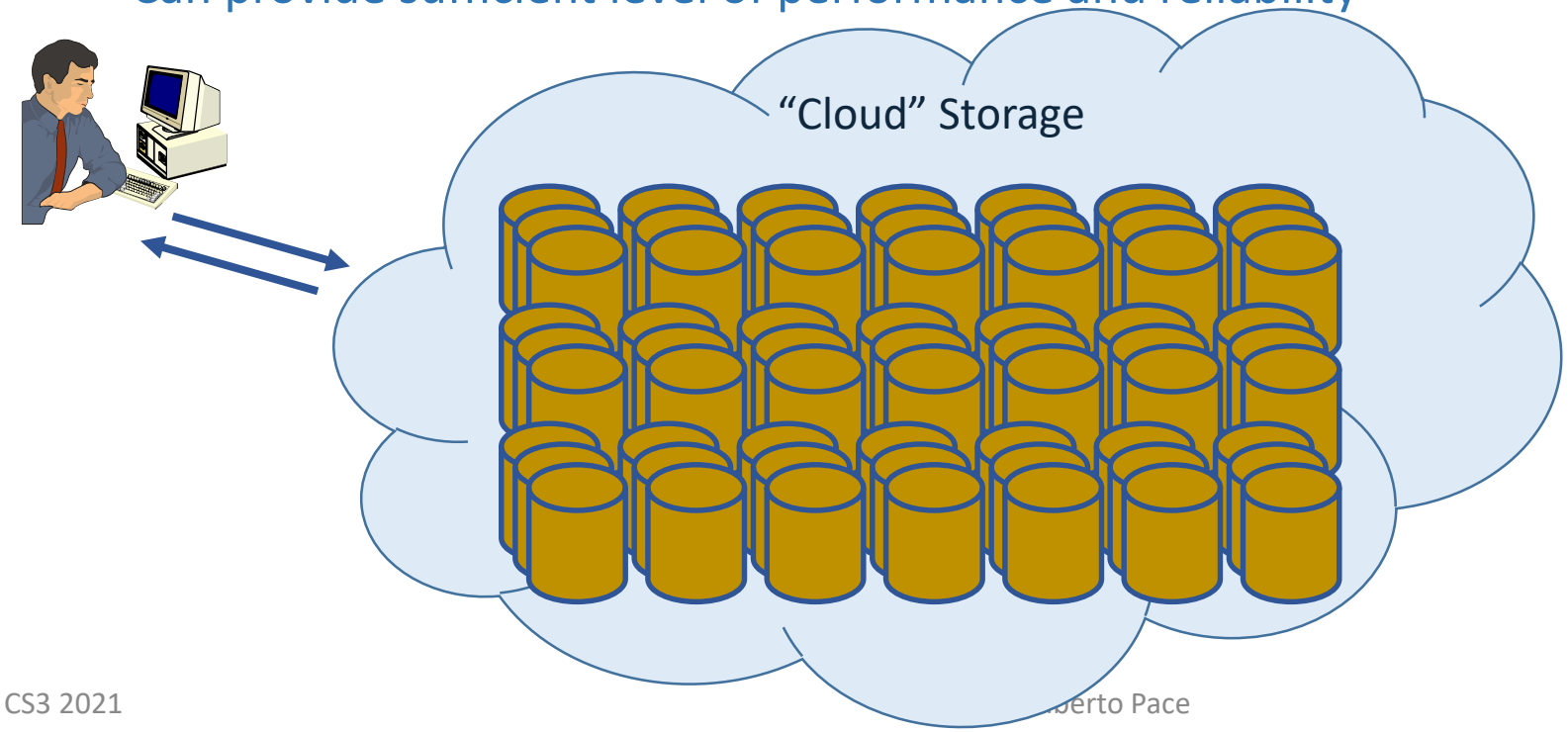
** WARNING **

- Do not confuse:

On cloud ≠ Owncloud

Why cloud storage ?

- Easy to understand, easy to sell …
- A simple storage model: all data into the same container
  - Uniform, simple, easy to manage, no need to move data
  - Can provide sufficient level of performance and reliability

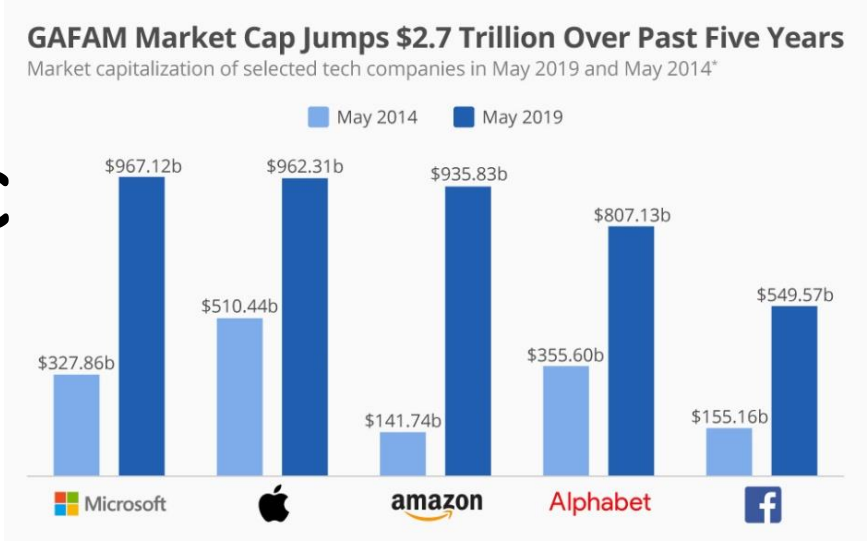


Cloud Storage is … Outsourcing !

When is it effective to outsource ?

- 3 requirements (apply to all fields, not just computing):
  - The activity is not strategic / not core business
  - The activity has clear established standard (interfaces / protocols)
  - There are multiple independent vendors implementing these standards.
- If any of these 3 requirements is not satisfied ...
  - You are exposed to problems

Data is more and more strategic



- Today … everything looks is OK
  - There are multiple (big, solid, …) vendors competing
  - There are established standard interfaces and protocol that ensure interoperability
- Ok for small amounts of data … but:
  - All major vendors offer free data ingestion, but expensive data retrieval. This breaks the model of interoperability and, in fine, data ownership.

*: source www.cnbc.com, www.statista.com

History repeats itself !

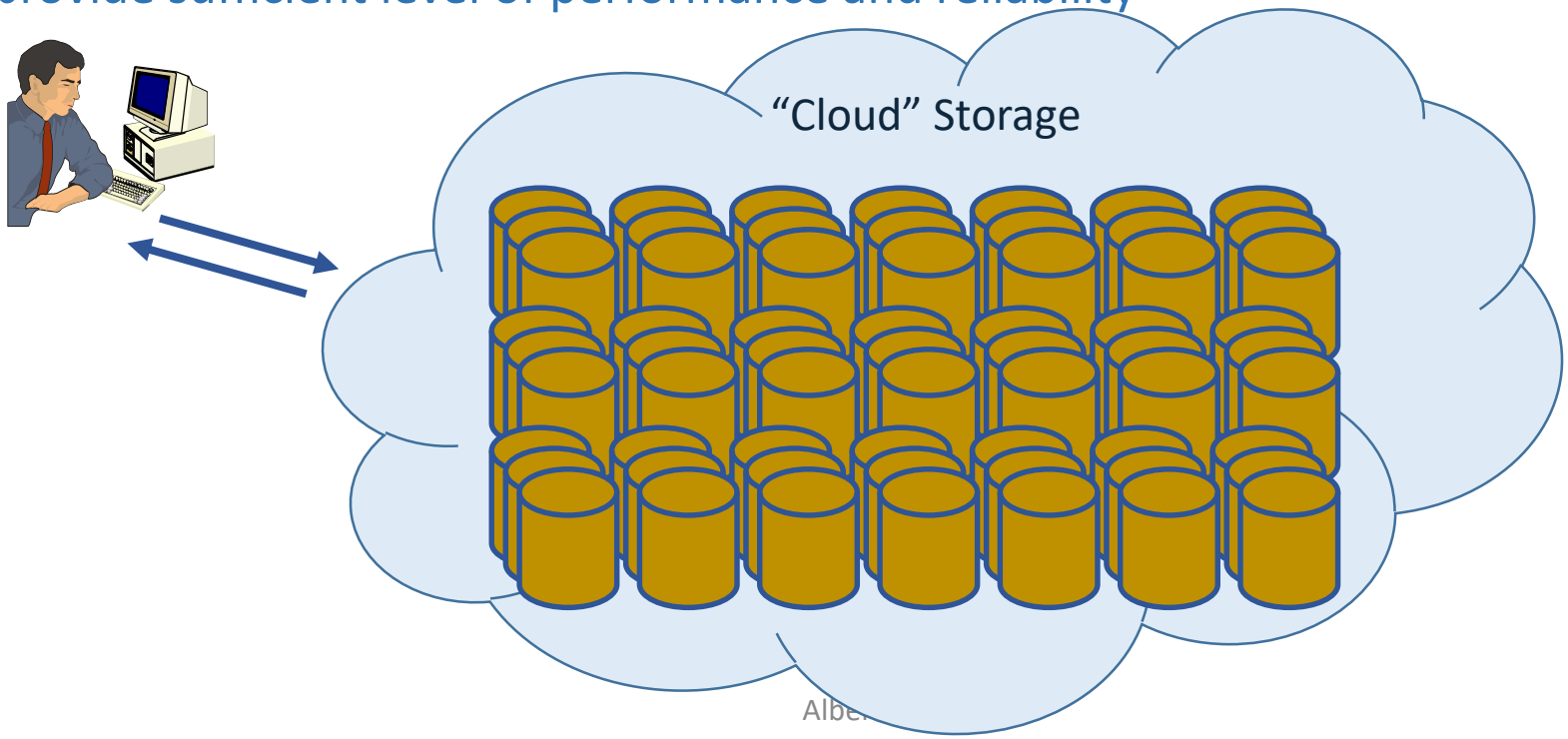
- Same story already heard:
  - 1990's Software: "is so flexible that can be done in the last minute …"
  - 2010's Data: "Why care about data? Just put It in the cloud (for free)"
- Vendor lock-in has a huge impact: access to your data is at stake
  - Companies fails
  - Contract fails
  - Law changes
  - Subject to remote jurisdictions
  - Sudden loss of access to our data
  - Loss of your own data

What are the options ?

- Two extremes:
  - Build your full software stack
    - Important fixed cost, minimal variable cost
    - Below critical mass, it is expensive (due to high fixed cost)
    - You have no external dependencies.
      - You can audit and assess your risks
      - You can only blame yourself if things fail
  - Move all your data in the cloud
    - Only variable costs. For small volumes, big savings possible
    - No data management necessary ...
    - But assessing risk is difficult as you cannot audit !

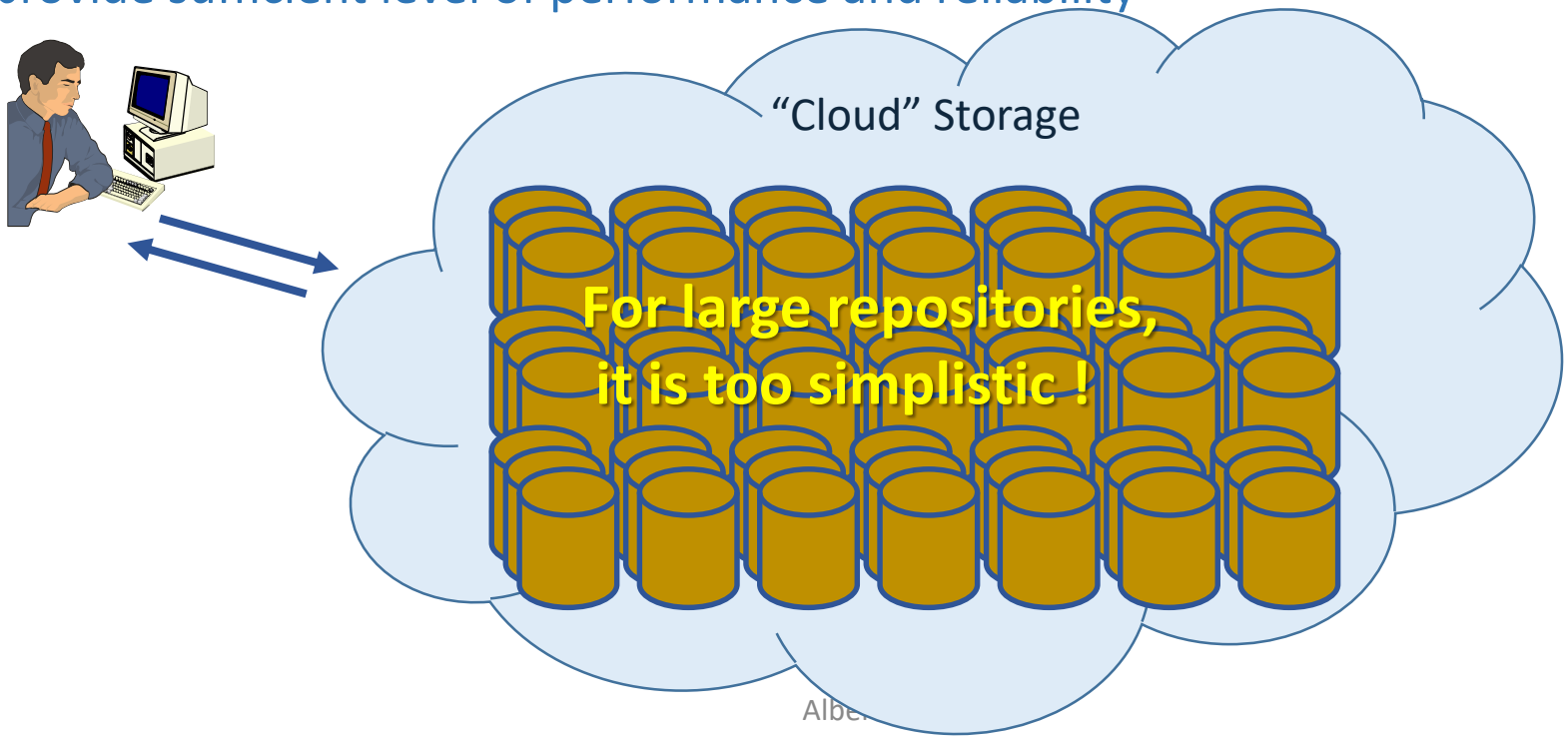
Why cloud storage ? (Back to slide #1)

- Easy to understand, easy to sell …
- A simple storage model: all data into the same container
  - Uniform, simple, easy to manage, no need to move data
  - Can provide sufficient level of performance and reliability



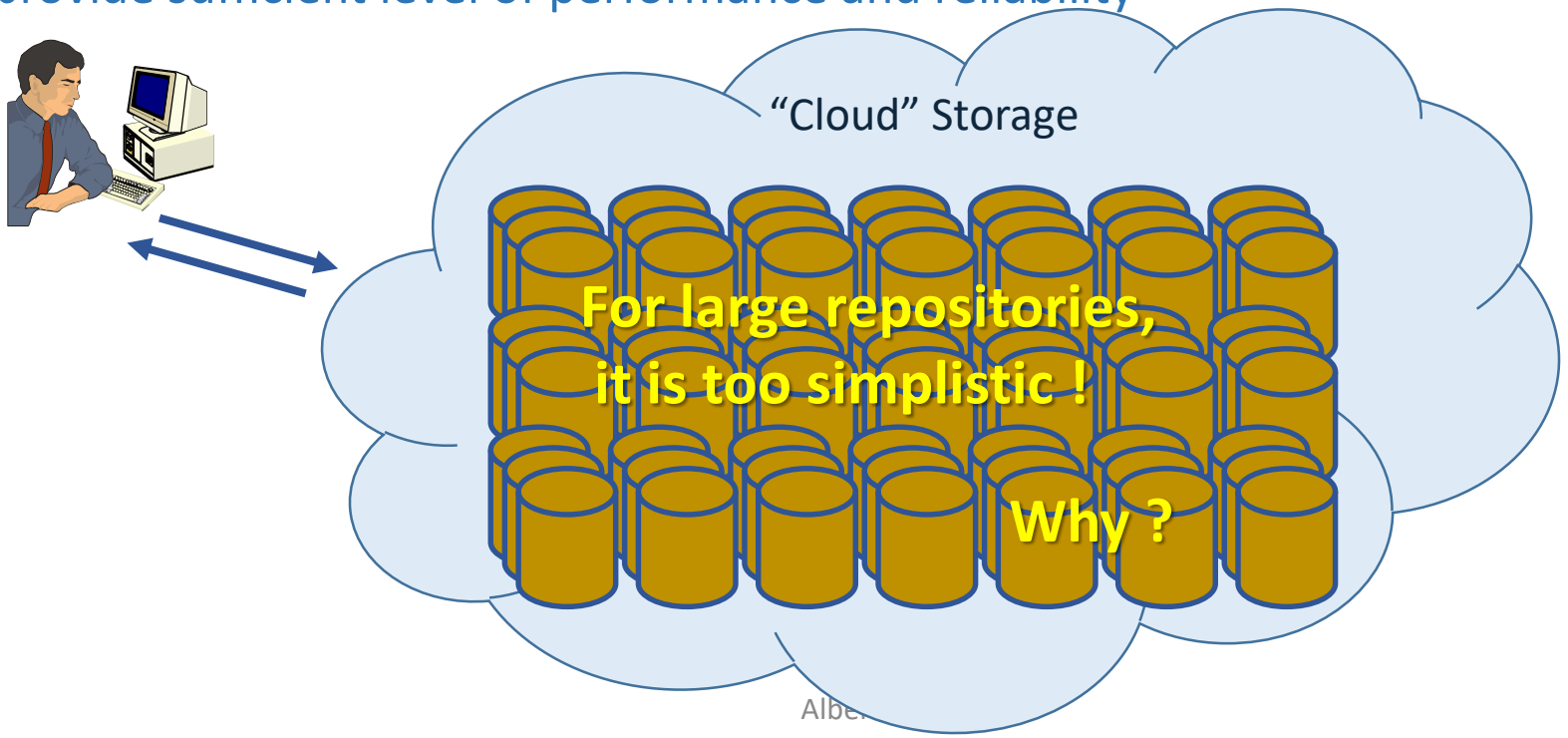
Why cloud storage ? (Back to slide #1)

- Easy to understand, easy to sell …
- A simple storage model: all data into the same container
  - Uniform, simple, easy to manage, no need to move data
  - Can provide sufficient level of performance and reliability



Why cloud storage ? (Back to slide #1)

- Easy to understand, easy to sell …
- A simple storage model: all data into the same container
  - Uniform, simple, easy to manage, no need to move data
  - Can provide sufficient level of performance and reliability

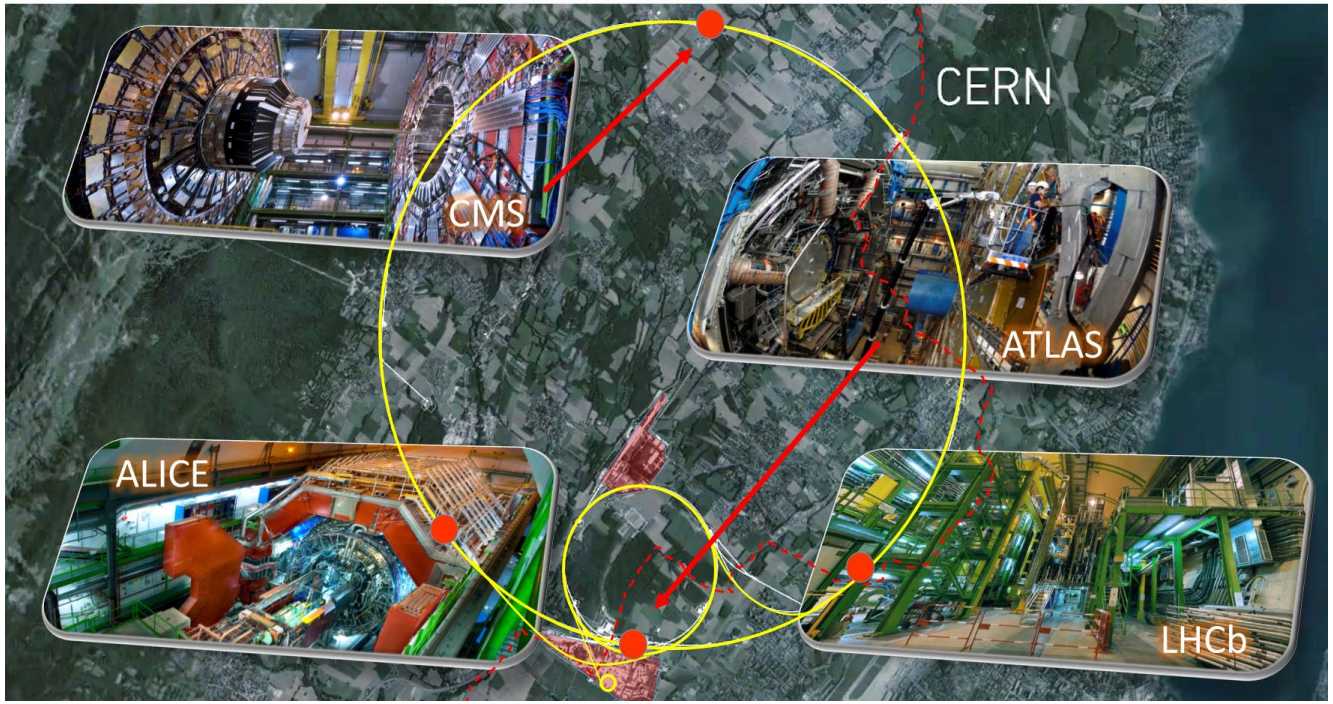


The need for multiple pools and quality

- Raw data that need to be analyzed
  - Need high performance, High reliability, can be expensive (small sizes)
- Raw data that has been analyzed and archived
  - Must be low cost (huge volumes), High reliability (must be preserved), performance not necessary
- Derived data used for analysis and accessed by thousands of nodes
  - Need high performance, Low cost, minimal reliability (derived data can be recalculated)

How Large is “Large Quantities of Data” ?

- Today, some sciences are dealing with hundreds of petabytes, reaching Exabyte scale in few years
  - Weather forecast, Earth observations, biology, medicine, astronomy, high energy physics, …

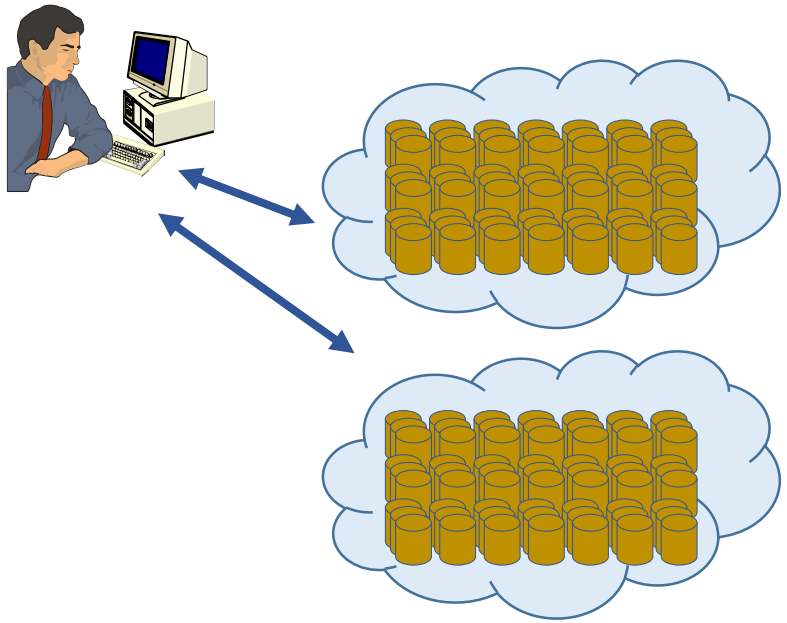


Why on premise can be cheaper ?

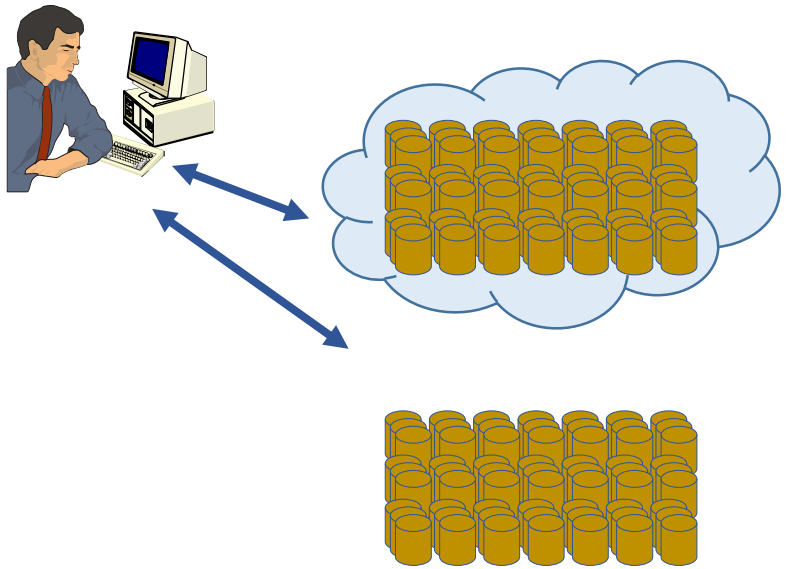
- If data is strategic, the software for data management is also strategic
  - … but you must have the skills
  - Managing your software stack has a fixed-cost only
  - On-premise open source has no variable cost. Maintaining the skills is a fixed costs
- Storage Hardware
  - If the "software" has fixed cost, the hardware + energy is where variable-costs are concentrated – scale out is possible at minimal ‘marginal’ cost.
- With this approach …
  - the cost of adding a PB of storage is limited to the cost of a PB of HW
  - the cost of operating an additional PB of storage is limited to the cost of the required energy and hardware amortisation
  - If you do not have the critical mass …. cooperation is the solution

Mitigating risks in cloud solutions

- Multiple cloud providers
  - You have to do your own data management

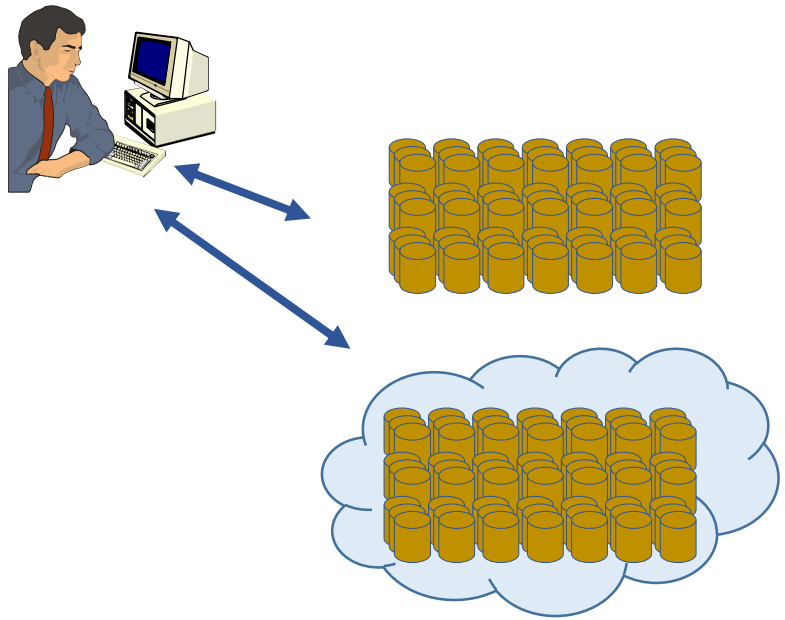


- Keep a copy on premise
  - You need an infrastructure
  - You have to do your own data management



Mitigating risks when data in on premises

- Keep a copy in the cloud
  - You can simplify your infrastructure (you have a safety net)
  - It costs more ...



Data at CERN

- CERN has the critical mass to build its own data centre
  - planning to reach exabyte scale in 2 years
- For storage, CERN has its own software stack (CERNBOX, EOS, FTS, …) built on top of established and interoperable open source solutions (Linux, Virtualization, Containers, Block Storage, Sync and Share, …)
  - The stack ensures Storage, Backup/Archival, Data Transfers, Sync and Share
- This approach has required important investments in infrastructure
  - but the marginal cost of storage is limited to the media and energy cost, which is one order of magnitude lower than cloud storage

Findings so far …

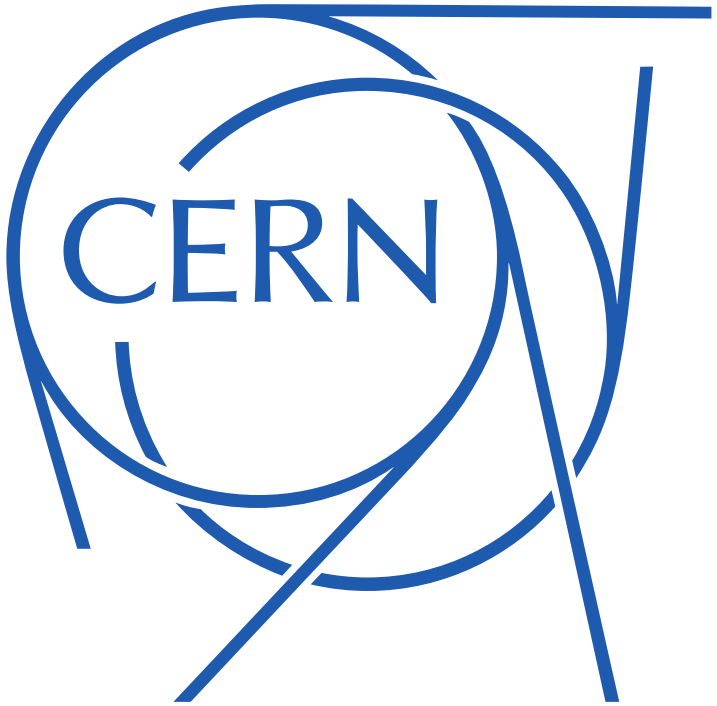
- For Storage, the open source approach is working very well since several years
- The CERNBOX Sync & Share portal to access storage is becoming also an access point to corporate recommended applications in the MALT project
- The critical mass to build a successful infrastructure is beyond what a small institute can afford:
  - The **consortium** / open source approach is the best practice to collaborate on well focused projects that guarantee maintaining ownership of your critical activities at a minimum cost

When funded with public money …

- The public sector is particularly risk-exposed as contracts are granted on the lowest bid …
  - Interoperability is required in the initial contract, but as the product evolves, it can be deliberately removed
- If interoperability is not guaranteed, it is easy to win public contracts by bidding at very low prices (big organization are even offered free services)
  - Vendors' expectation is to achieve a lock-in, and increase prices at contract renewal.
- Costs of change can be huge, and are not provisioned in contracts.
  - They represent a debt
  - Debt removes the freedom of choice, it removes sovereignty

Conclusions

- The more critical a component is to your business, the more marketing pressure you will receive to outsource it (i.e. use proprietary sw)
- Outsource standard activities, well defined and interoperable
  - Do not outsource your own business
- Insource what is specific to you, and your critical activities
- The on-premise / open source approach is the best practice to insource your critical activities at a minimum cost
  - This will guarantee fixed cost for software.
  - No license cost proportional to data volumes (or number of nodes, or cores, or disk, or data transferred).


CERN