From data management to storage services to the next challenges

Alberto Pace
CERN IT department

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

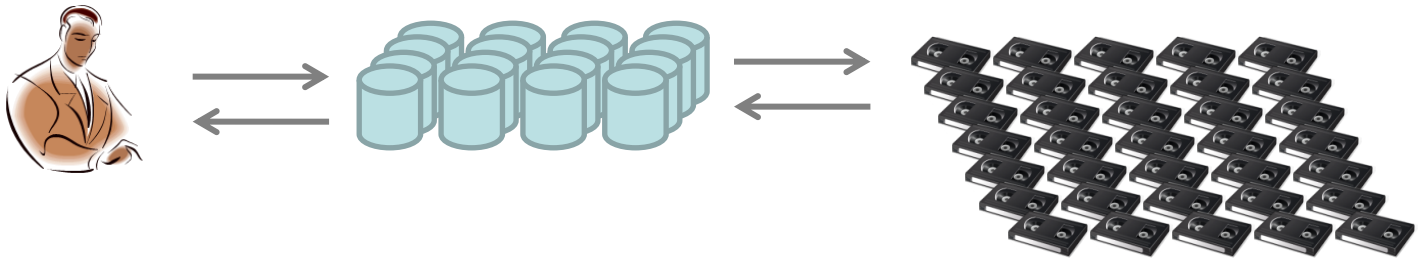
Data management

- Back few years, before the LHC starts …
- Data emphasis was on “management”

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Data management

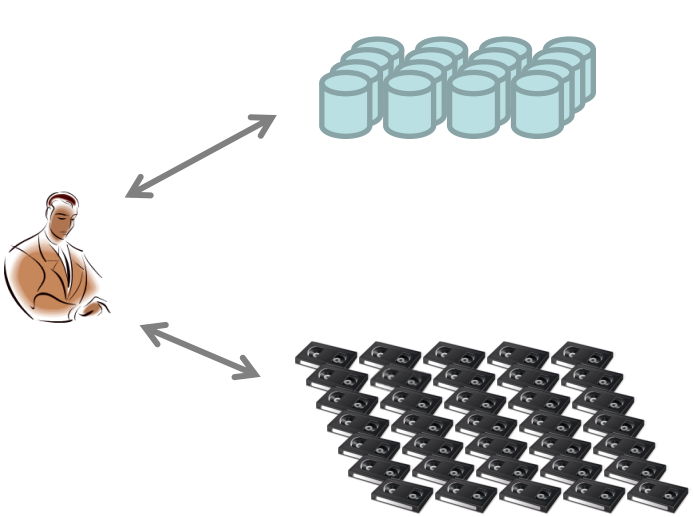
- Back few years, before the LHC starts …
- Data emphasis was on “management”

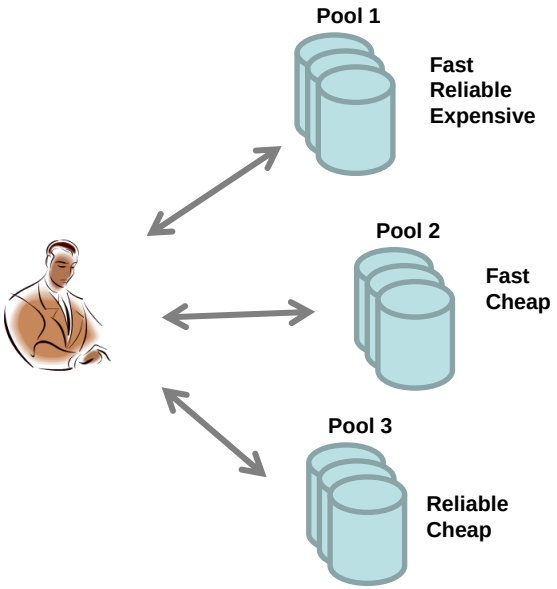


- Lot of developments to implement automation for various management strategies
  - High efficiency require a match between expected and real data access pattern
  - In depth understanding on how the system works is required for high performance
  - The system complexity made misunderstanding possible triggering, in few cases even data loss

Lesson learned

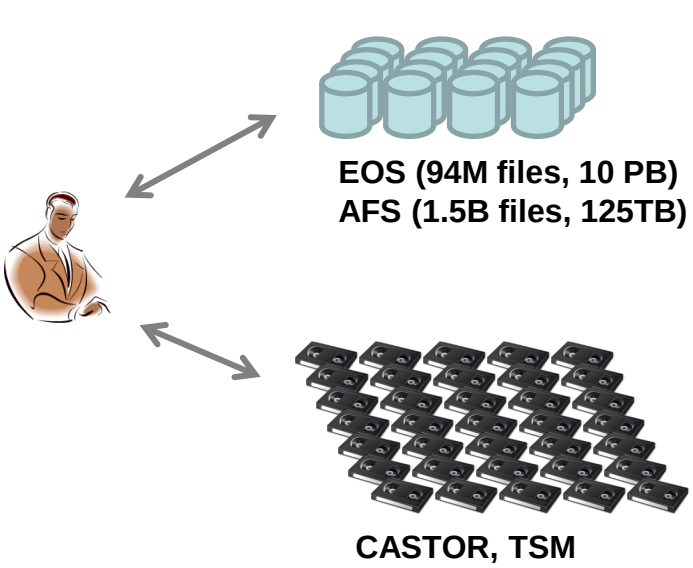
- Data “management” is better done by the data owner (experiment) who has the knowledge about data and access pattern
- We (IT) should focus in data and storage services (DSS) and offer tools for efficient data management
- Two building blocks to empower data management
  - Data pools with different quality of services
  - Tools for data transfer between pools and its automation

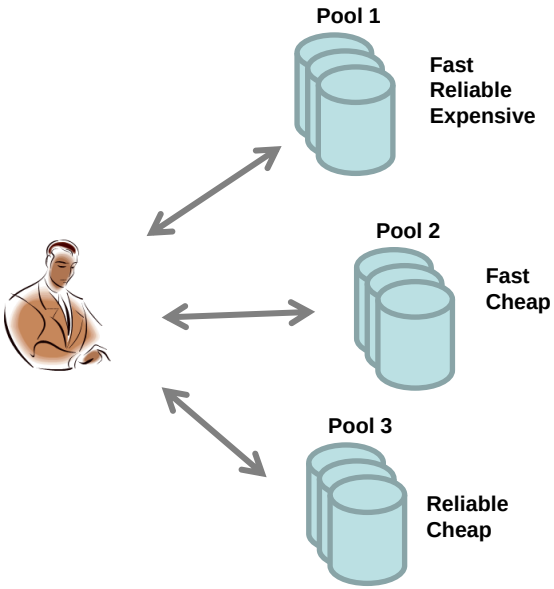




Present Architecture in IT

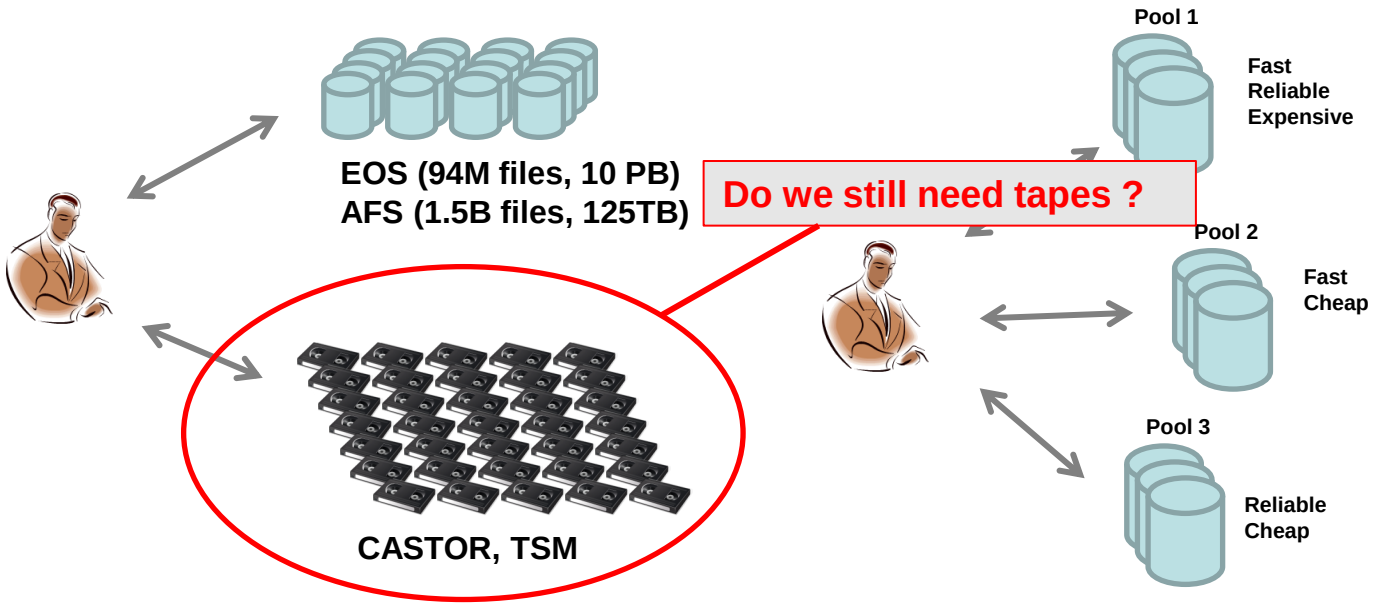
- 4 services to cover (nearly) all use cases
  - Disk and SSD based: EOS and AFS
  - Tape and Robotics: CASTOR and TSM
- Key requirements: Simple, Scalable, Consistent, Reliable, Available, Manageable, Flexible, Performing, Cheap, Secure.
- Aiming for “à la carte” services (storage pools) with on-demand “quality of service”





Present Architecture in IT

- 4 services to cover (nearly) all use cases
  - Disk and SSD based: EOS and AFS
  - Tape and Robotics: CASTOR and TSM
- Key requirements: Simple, Scalable, Reliable, Available, Manageable, Flexible, Performing, Cheap.
- Aiming for “à la carte” services (storage pools) with on-demand “quality of service”



Why we have tapes

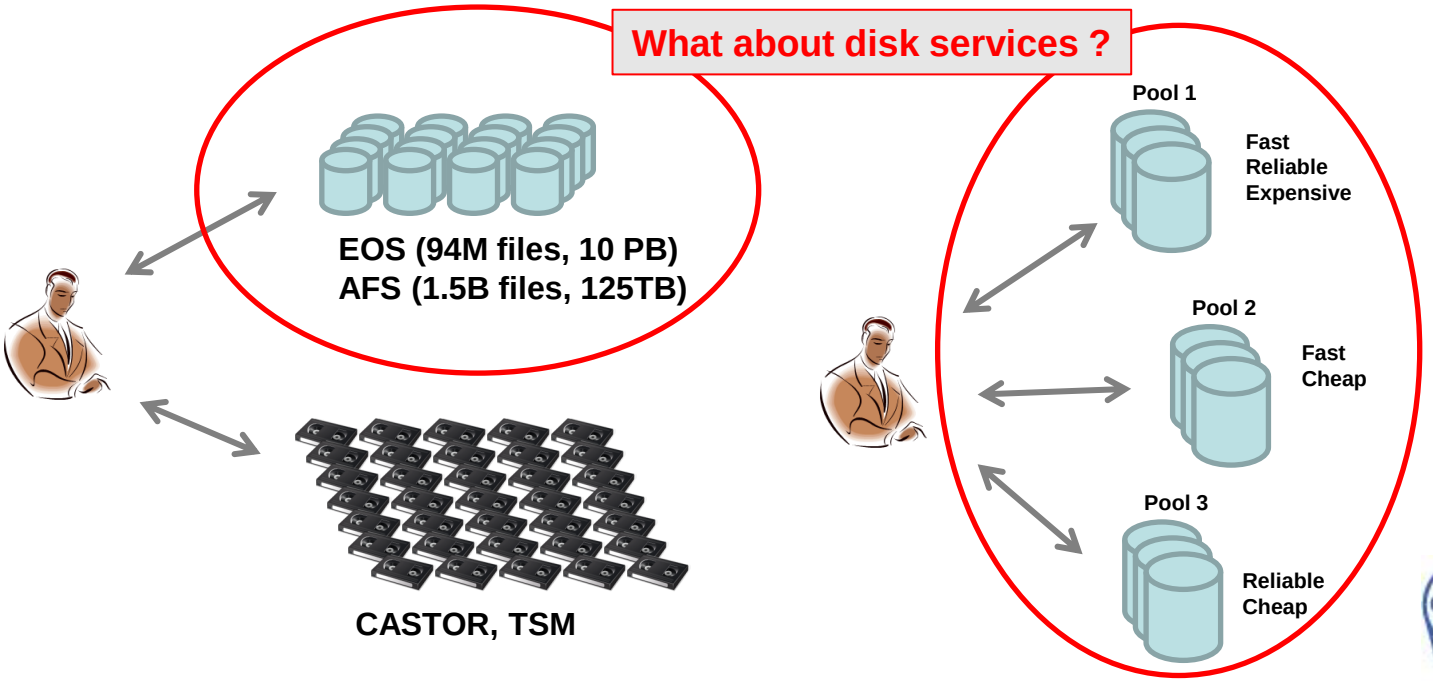
- Tapes have a bad reputation in some use case
  - Extremely slow in random access mode
    - high latency in mounting process and when seeking data (F-FWD, REW)
  - Inefficient for small files (in some cases)
  - Comparable cost per (peta)byte as hard disks

Why we have tapes

- Tapes have a bad reputation in some use case
  - Extremely slow in random access mode
    - high latency in mounting process and when seeking data (F-FWD, REW)
  - Inefficient for small files (in some cases)
  - Comparable cost per (peta)byte as hard disks
- Tapes have also some advantages
  - Extremely fast in sequential access mode
    - 2x faster than disk, with physical read after write verification
  - Several orders of magnitude more reliable than disks
    - Few hundreds GB loss per year on 80 PB tape repository
    - Few hundreds TB loss per year on 50 PB disk repository
  - No power required to preserve the data
  - Less physical volume required per (peta)byte
  - Inefficiency for small files issue resolved by recent developments
  - Nobody can delete hundreds of PB in minutes
- Bottom line: if not used for random access, tapes have a clear role in the architecture

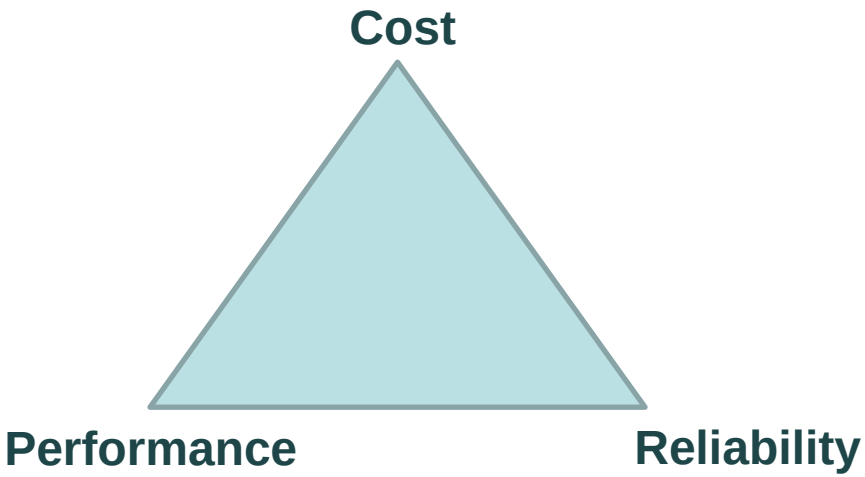
Present Architecture in IT

- 4 services to cover (nearly) all use cases
  - Disk and SSD based: EOS and AFS
  - Tape and Robotics: CASTOR and TSM
- Key requirements: Simple, Scalable, Reliable, Available, Manageable, Flexible, Performing, Cheap.
- Aiming for “à la carte” services (storage pools) with on-demand “quality of service”



(Disk) Data pools

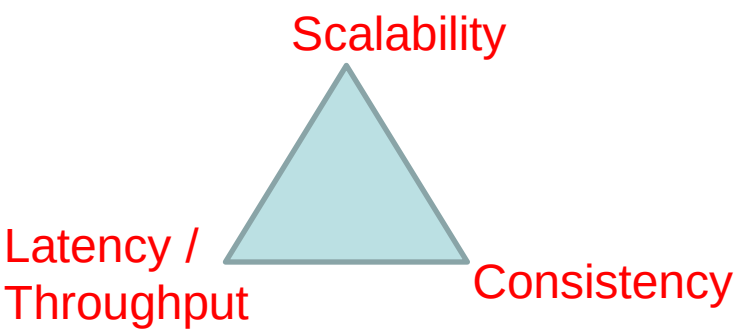
- Required to cover the entire space of requirements
- Different quality of services
  - Three parameters:
    (Performance, Reliability, Cost)
  - You can have two but not three

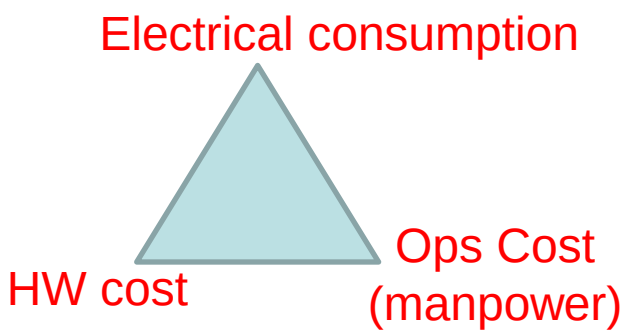


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

But the balance is not simple

- Many ways to split (performance, reliability, cost)
- Performance has many sub-parameters
- Cost has many sub-parameters
- Reliability has many sub-parameters





CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Next challenges

- Several areas of R & D
  - With close links with the experiments, with PH-SFT, with the open source community and storage vendors
- Storage for Analysis (EOS)
  - Scalability vs Consistency
  - Arbitrary Performances / Arbitrary Reliability
- Evolution of “File System” Service
  - Cope with unprecedented growth
  - Support offline work and sync using cloud protocols
- Hadoop, virtualization, no-sql / DB applications, and cloud storage, require storage support for “block level” access

EOS Challenges

- Scalability vs Consistency
  - The EOS nameserver performance is outstanding but requires to remain ahead of growing requirements
  - Identify ideal architectures to cope with WAN (Geneva – Budapest) distributed services
- Arbitrary Performances / Arbitrary Reliability
  - À la carte Reliability & Performance
    - Arbitrary number of replicas, RAIN (Redundant Array of Inexpensive Nodes), pluggable error-corrections algorithms
  - Read vs Write performance
  - Jbod, Block storage, Variable % of SSDs

File system services challenges

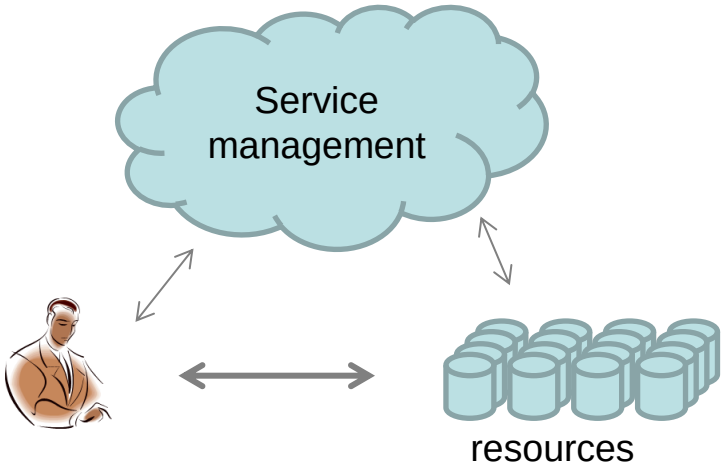
- Recognize the need of a distributed file system for the community
- Evolve the present AFS service
  - Additional client protocols ?
    - Both for file and cloud access
  - Offline / disconnected mode
  - Journaling / Versioning ? Self-Service backup / restore
  - “unlimited” space

Conclusion

- Ensure that IT service management does not get between the user and the resources

Conclusion

- Ensure that IT service management does not get between the user and the resources



- Experiments and end-users should be able to get the full performance of the resources available
  - Beware of the associated risks …