# Managing Storage in a (large) Grid data center

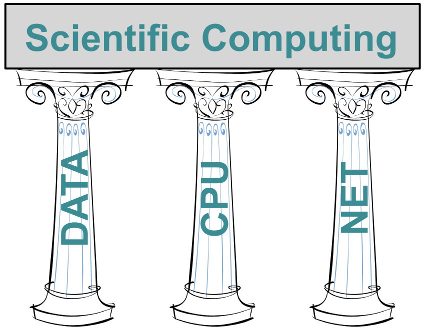
Alberto Pace
Head, Data and Storage Services group
CERN, Geneva, Switzerland

2nd symposium on Grid computing in the Americas
Instituto de Ciencias Nucleares (UNAM)
**Nov. 3-4, 2014**
**Mexico City**

# Scientific Computing

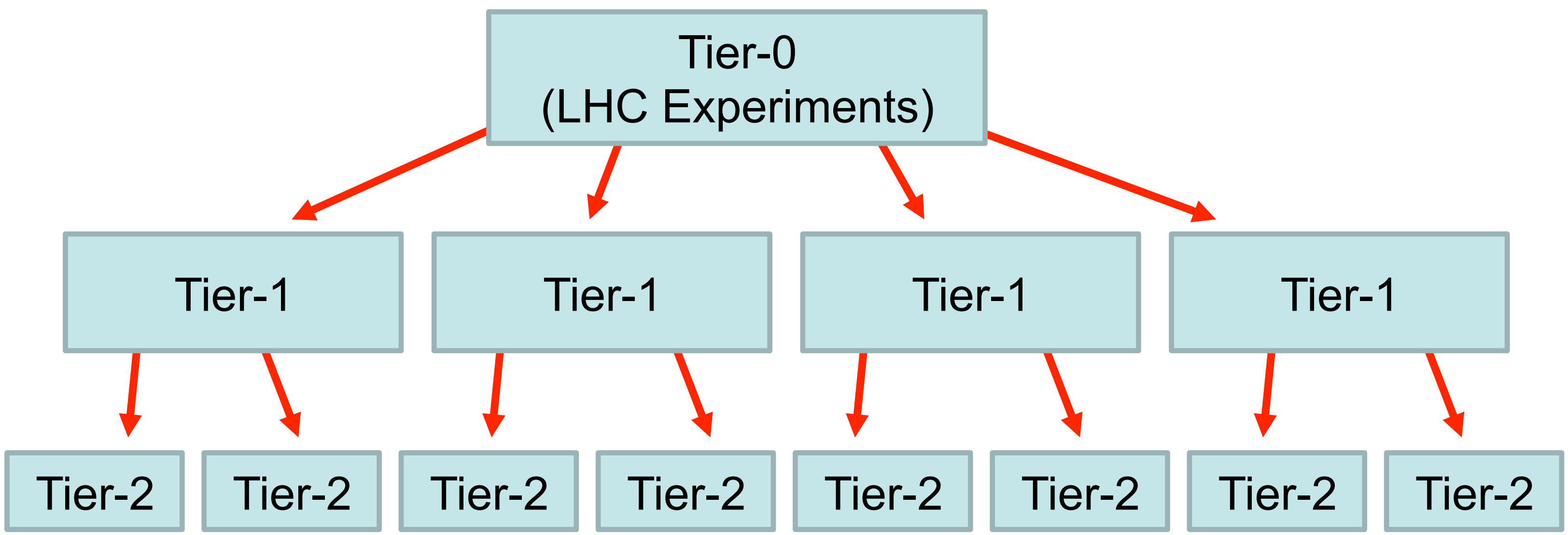


- Scientific computing for large experiments is typically based on a distributed infrastructure
- Storage is one of the main pillars
- A key parameter in distributed scientific computing is the efficiency
  - High efficiency requires to have the CPUs colocated with the data to analyze, using the network.




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

- Storage in scientific computing is distributed across multiple data centres
- Data flows from the experiments to all datacenters where there is CPU available to process the data

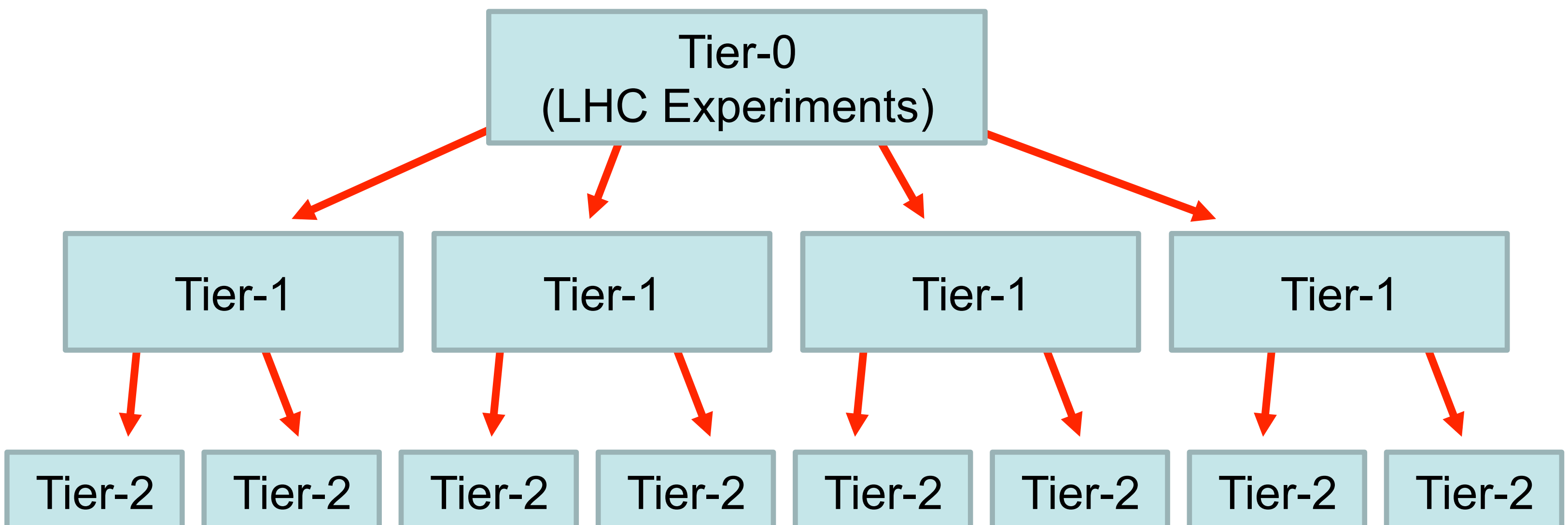


- Whenever a site has
  - idle CPUs (because no Data is available to process)
  - or excess of Data (because there is no CPU left for analysis)
- the efficiency drops

# Why storage is complex ?



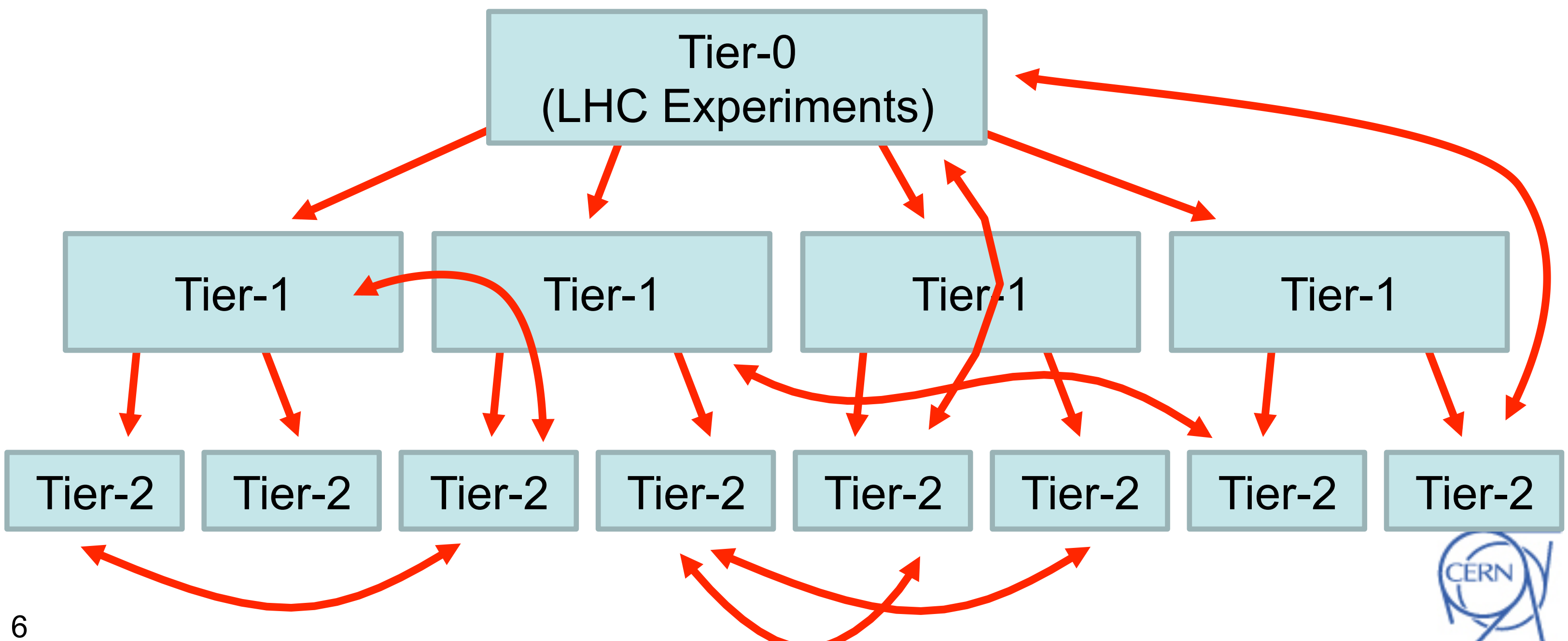
- Analysis made with high efficiency requires the data to be pre-located to where the CPUs are available




# Why storage is complex ?



- Analysis made with high efficiency requires the data to be pre-located to where the CPUs are available
- Or to allow peer-to peer data transfer
  - This allows sits with excess of CPU, to schedule the pre-fetching of data when missing locally or to access it remotely if the analysis application has been designed to cope with high latency




# Why storage is complex ?

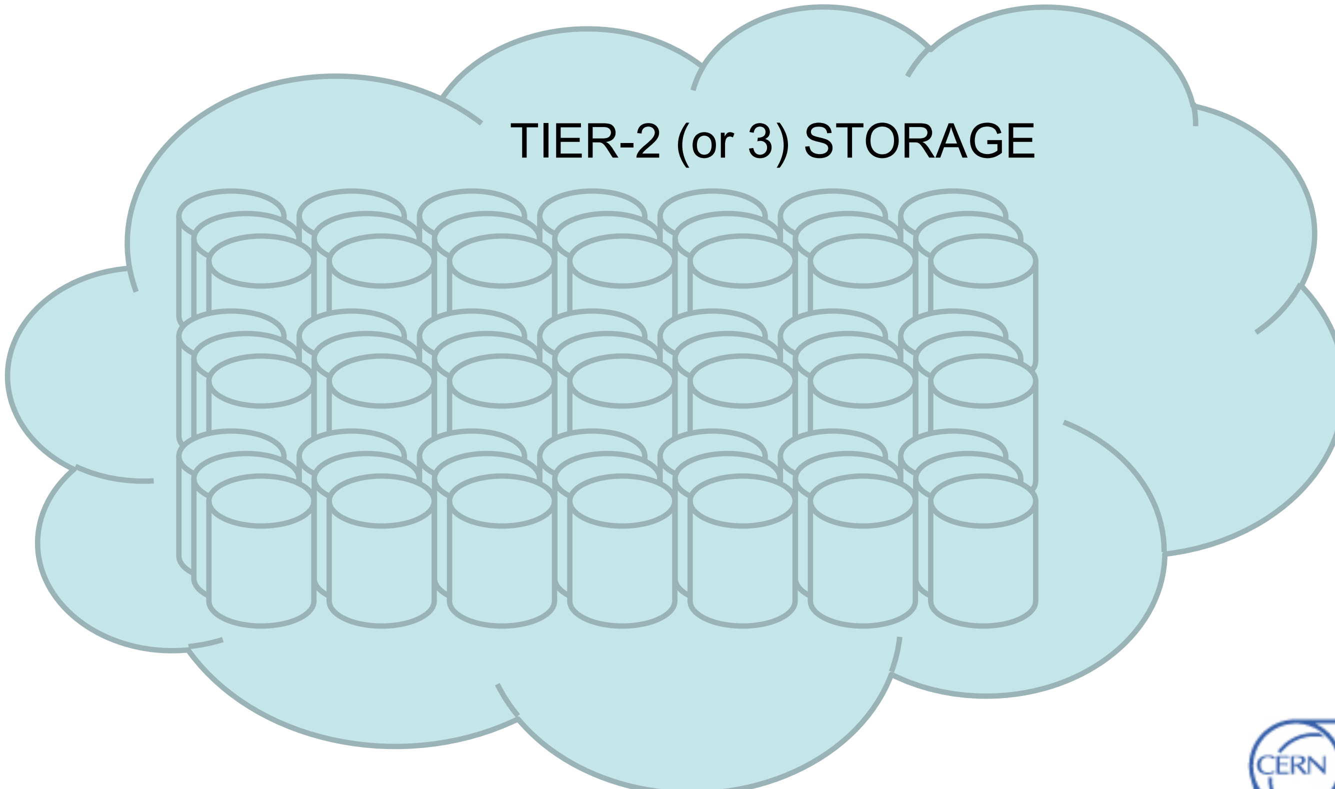
CERN IT Department

- Both approaches coexists in WLCG
- Data is pre-placed
  - This is the role of the experiments that plans the analysis
- Data is globally accessible and federated in a global namespace
  - The middleware always attempt to take the local data and uses an access protocol that redirects to the nearest remote copy when the local data is not available
  - All middleware and jobs are designed to minimize the impact of the additional latency that the redirection requires
- Using access protocols that allows global data federation is essential
  - http, xroot

# Within a Storage site



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



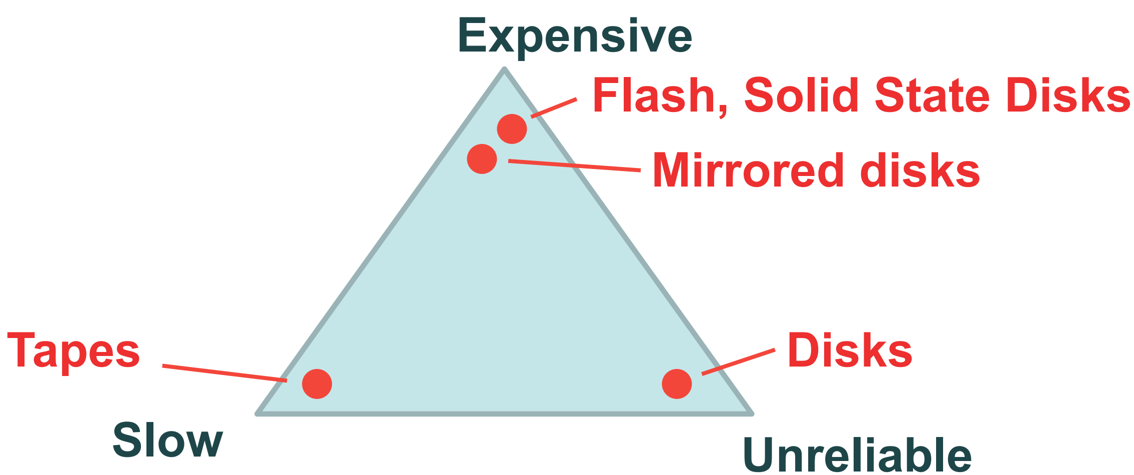

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

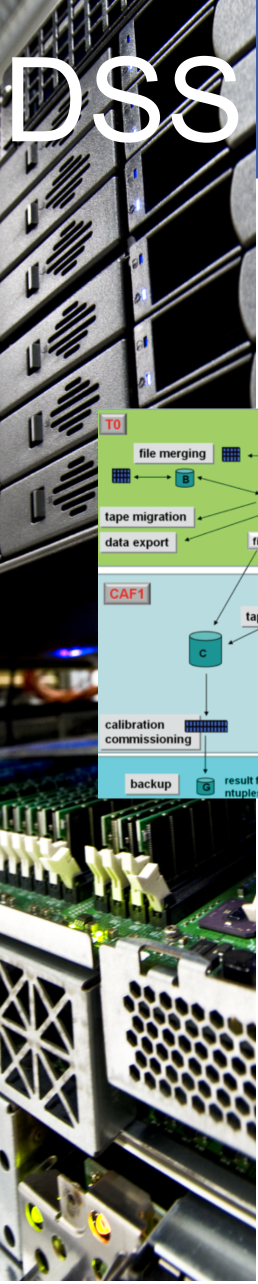


# Limitations



- Different storage solutions can offer different quality of services
  - Three parameters: Performance, Reliability, Cost
  - You can have two but not three
- To deliver both performance and reliability you must deploy expensive solutions
  - Ok for small sites (you save because you have a simple infrastructure)
  - Difficult to justify for large sites



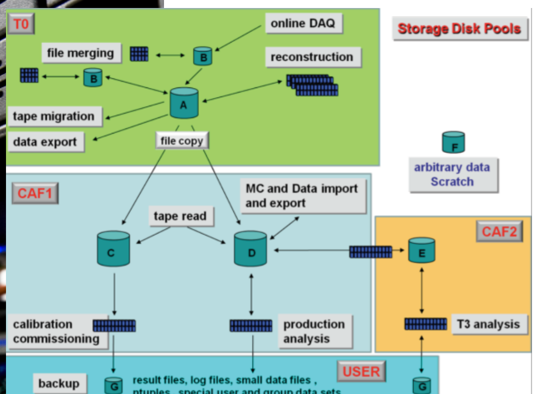


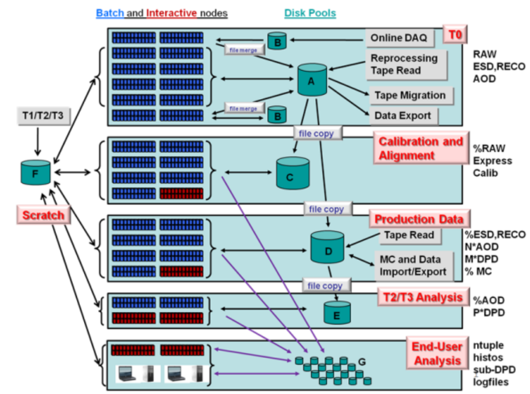


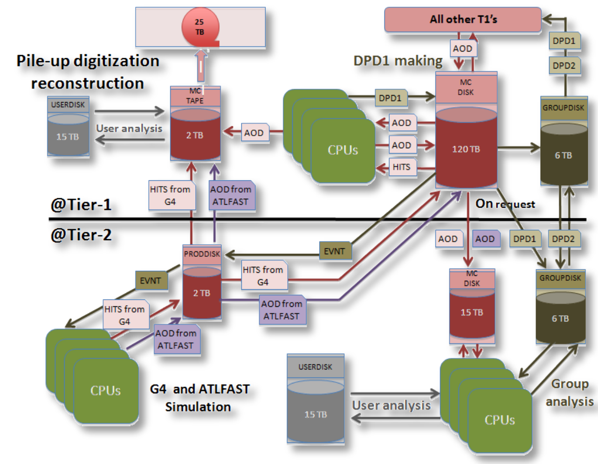
# What is data management ?



- Examples from LHC experiment data models



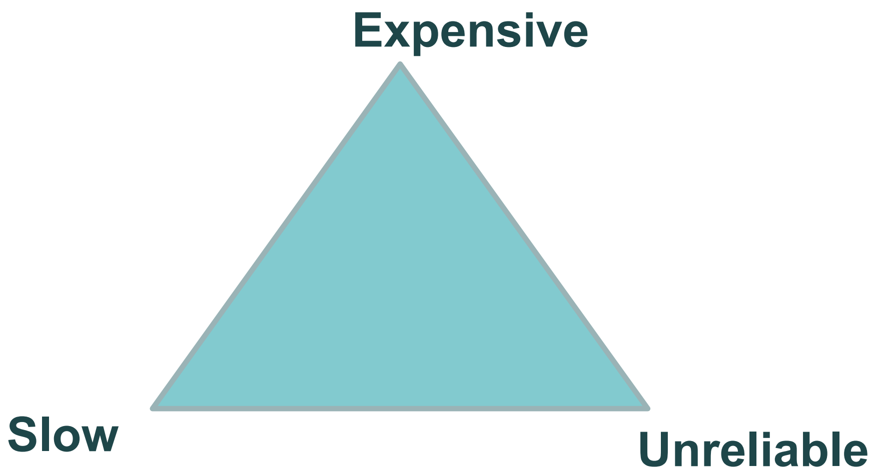




- Two building blocks to empower data analysis
  - Data pools with different quality of services
  - Tools for data transfer between pools

# Storage services



- Storage need to be able to adapt to the changing requirement of the experiments
  - Cover the whole area of the triangle and beyond



# Examples

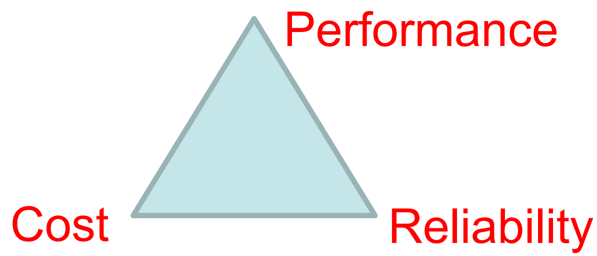


- **Why multiple pools with different quality of services are needed ?**
- Derived data used for analysis and accessed by thousands of nodes
  - Need high performance, Low cost, minimal reliability (derived data can be recalculated)
- Raw data that need to be analyzed
  - Need high performance, High 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

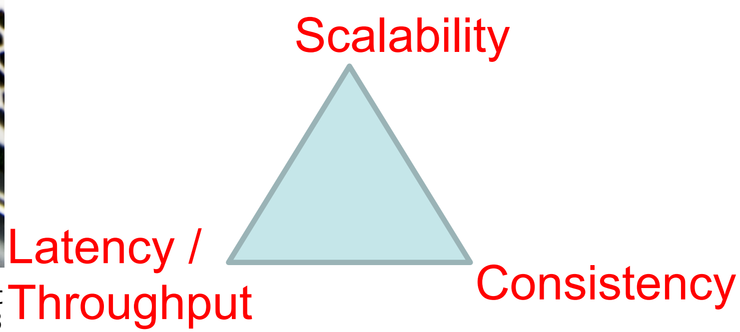
# DSS But the balance is not simple

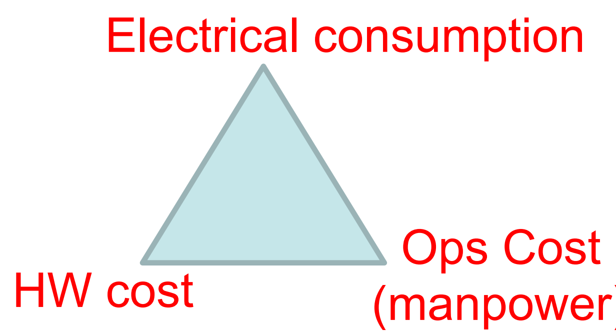


- Many ways to split (performance, reliability, cost)




- Performance has many sub-parameters
- Cost has many sub-parameters
- Reliability has many sub-parameters



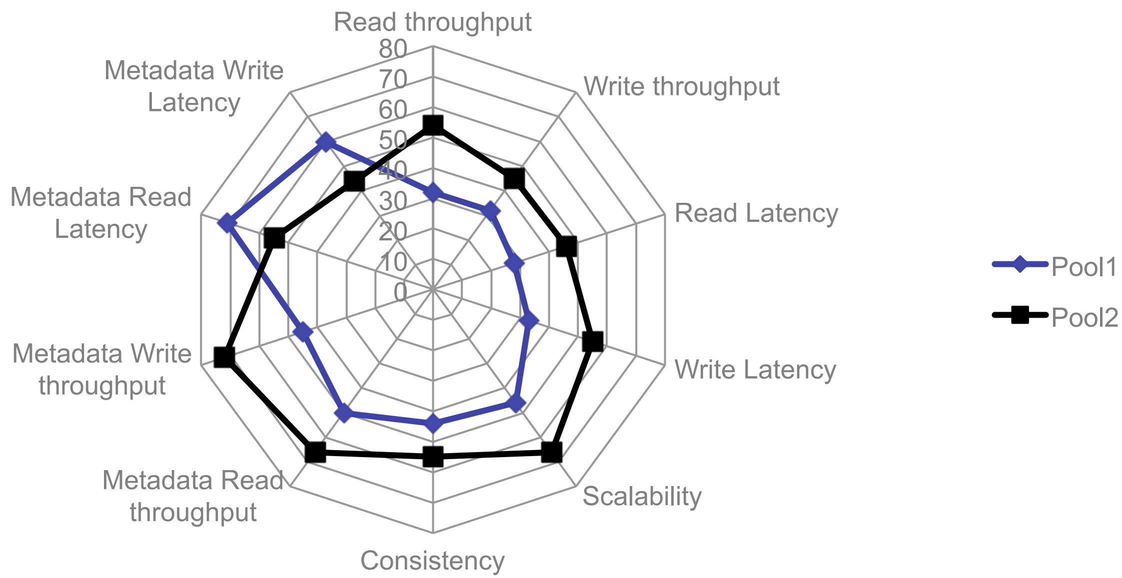





# And reality is complicated



- 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”
- And where is scalability ?




# Tools needed



- Needed:
  - Tools to transfer data effectively across pools of different quality
  - Storage elements that can be reconfigured “on the fly” to meet new requirements without moving the data
- Examples
  - Moving petabytes of data from a multiuser disk pool into a reliable tape back-end
  - Increasing the replica factor to 5 on a pool containing condition data requiring access form thousands of simultaneous users
  - Deploying petabytes of additional storage in few days


# Roles Storage Services



- Three main roles
  - Storage (store the data) Size in PB
  - Distribution (ensure that data is accessible) Availability
  - Preservation (ensure that data is not lost) Reliability

| | |
|---|---|
| Tier-0 (LHC Experiments) | Storage, data distribution, data preservation |
| ↓ Tier-1 | Storage, data distribution, data preservation |
| ↓ Tier-2, Tier-3 | Storage, data distribution |

# Reliability …



- You can achieve high reliability using standard disk pools
  - Multiple replicas
  - Erasure codes
- Here is where tapes can play a role
  - Tapes ??

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

# Do we need tapes ?



- Tapes have a bad reputation in some use cases
  - 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



# Do we need tapes ?

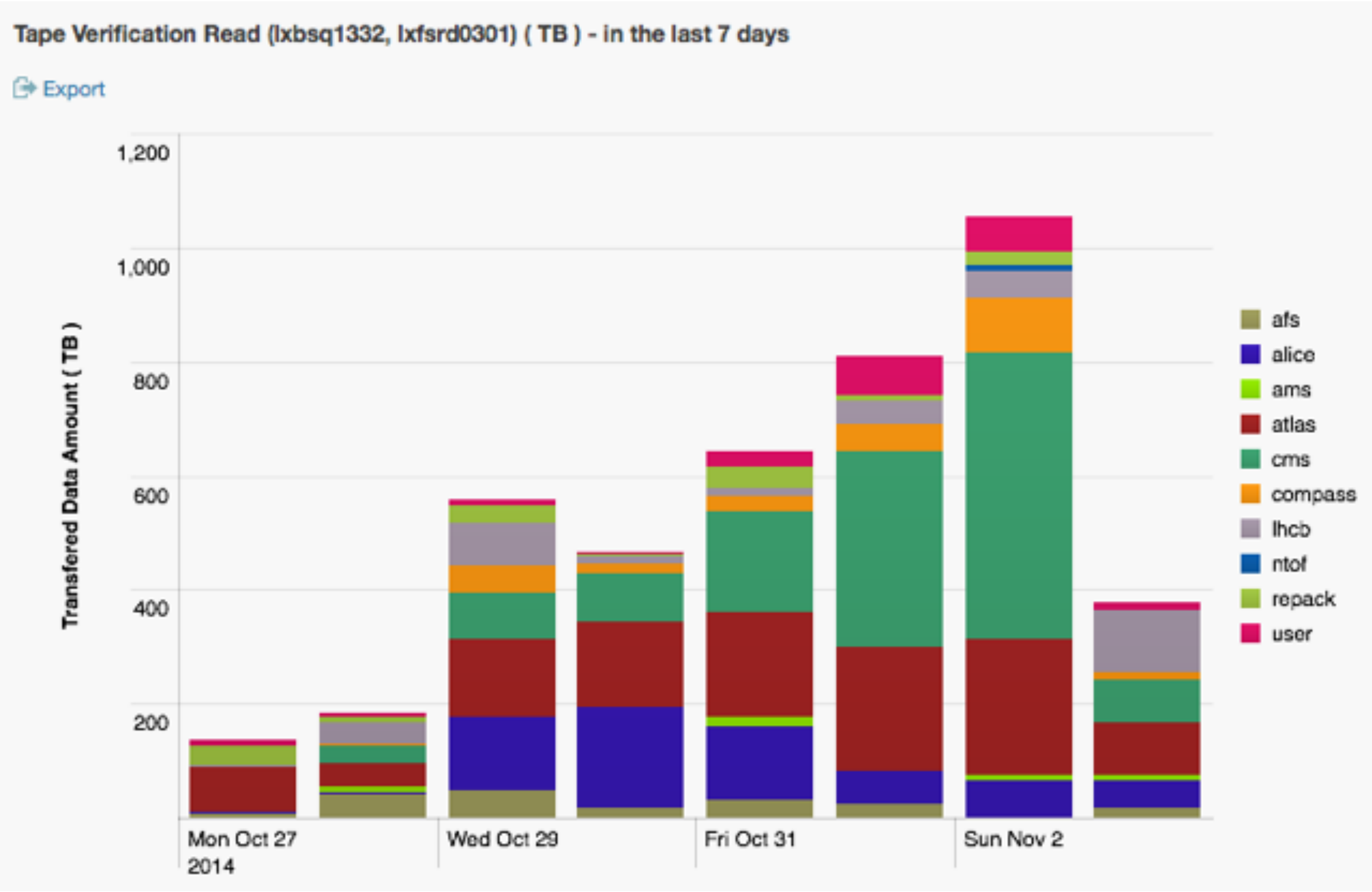


- Tapes have a bad reputation in some use cases
  - Slow in random access mode
    - high latency in mounting process and when seeking data (F-FWD, REW)
  - Inefficient for small files (in some cases)
  - Comparable cost per (peta)byte as hard disks

- Tapes have also some advantages
  - 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

# DSS Verifying more than 1PB in 24h



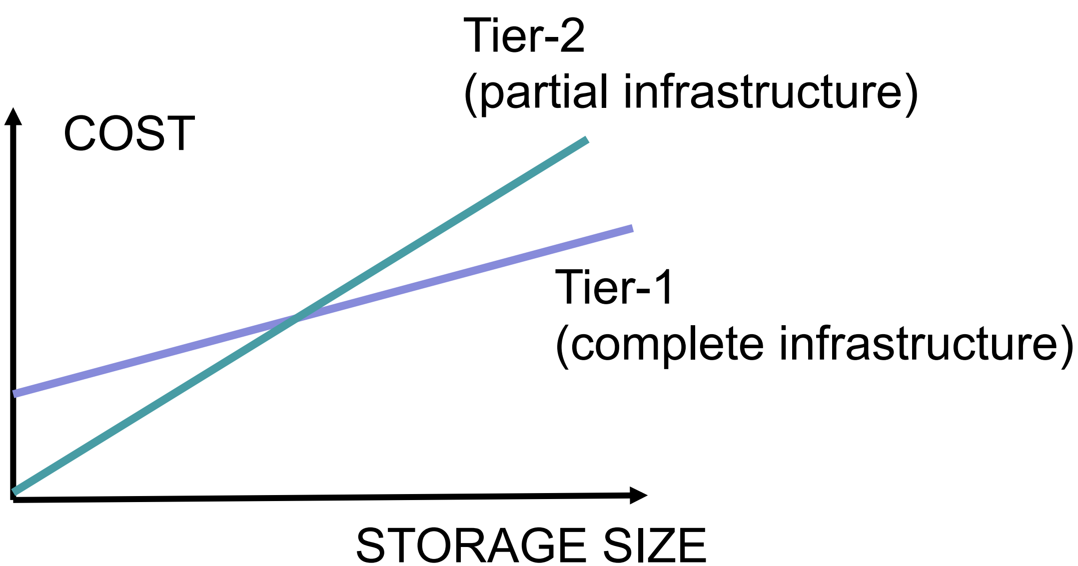



Source: G. Cancio / CERN

# Practically …
# which solution ?

# … not a single answer

- There are fixed costs for the skilled manpower that is required to manage the high efficiency but complex solutions
  - Disk technology, Database technology, Tape technology, Networks, Databases…




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

# DSS AT CERN

- Many options available
  - interoperability is a key requirement for WLCG
  - DPM, dCache
- AT CERN
  - Separation between HOT (disks) and COLD (tapes) storage
  - CASTOR (used for the tape storage backend)
  - EOS (used for the disk storage frontend)
    - guaranteed low-latency file access for analysis
    - collaborative storage platform with space and access regulations
    - expand/exchange/shrink during operation
    - semi-automatic disk ejection & healing – reduced maintenance


# Why this diversity



- Example:
  - DPM is a consolidated product which is fully supported by CERN. It meets all requirements of a small site
    - Integrates well within both xroot and http data federations
  - EOS is the new design for a scalable infrastructure designed to minimize management cost. It is ready to meet the requirements for larger sites
    - Larger functionalities


# Summary



- Storage for High Energy Physics or Scientific computing can be
  - Too simplistic
  - Simple
    - Typical for Small Tier-2
  - Scalable and Reliable
    - Typical for a Tier-1 / Large Tier-2
  - Too complex
- Behind the software you use, it is always the people that makes the difference

# Areas of developments in storage at CERN

## For EOS (Reserve slides)

Source: A.J.Peters/ CERN

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


# EOD design

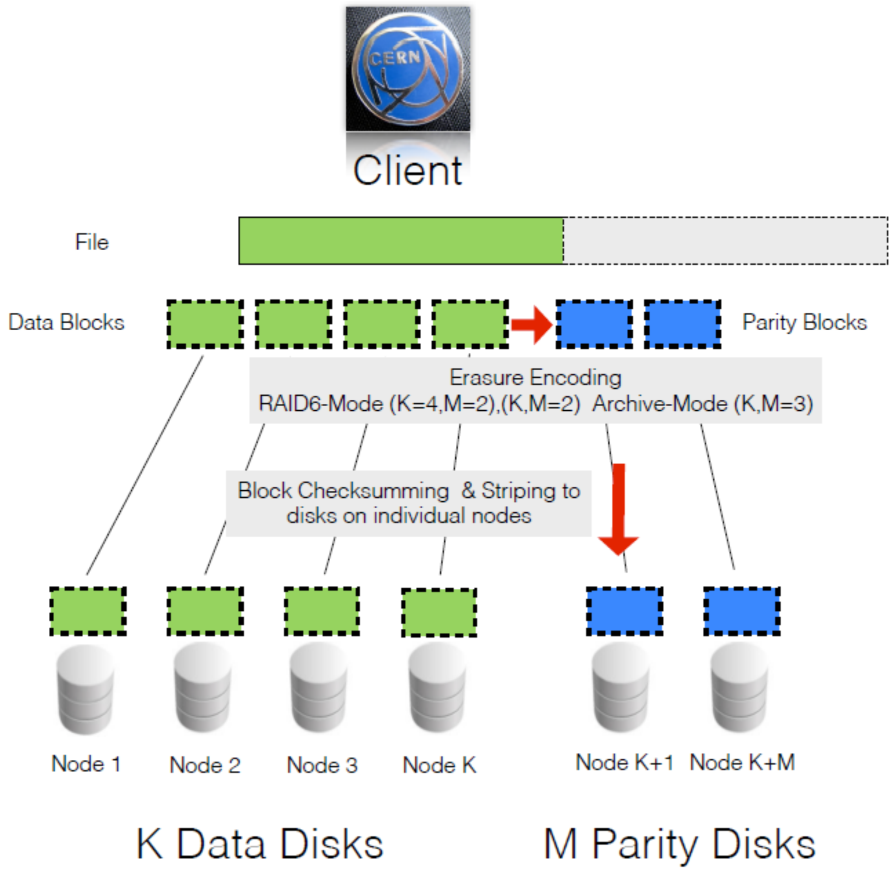


- No need for a central relational database
- Designed for infinite scalability and arbitrary reliability
- expand/exchange/shrink during operation
  semi-automatic disk ejection & healing

# DSS EOS

CERN IT Department

- # Arbitrary Reliability – RAIN
  - Redundant Array of Independent Nodes




**RAIN**

- Block Striping allows parallel IO
  - boost single file performance

    e.g. with K=4 280 MB/s streaming write & 400 MB/s streaming write

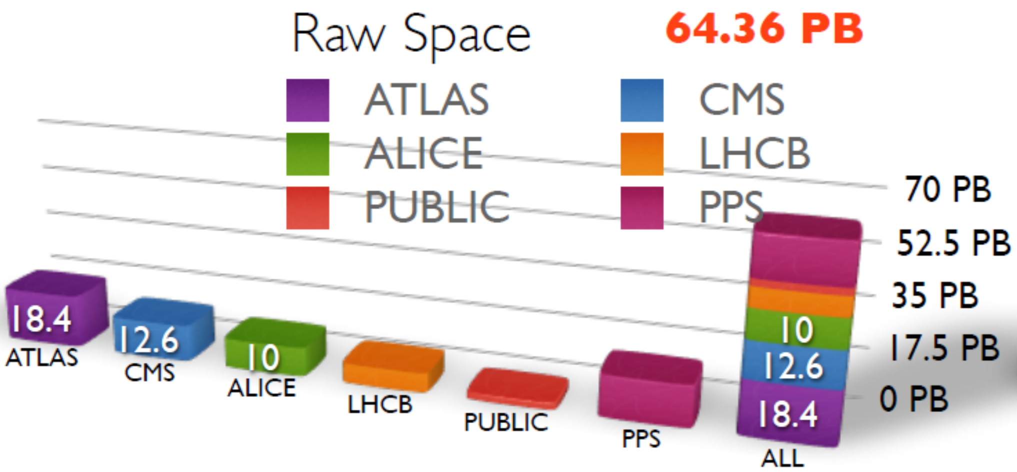
- Erasure Encoding + Block Checksumming allows
  - error correction on the fly
  - up to M concurrent disk failures without data loss

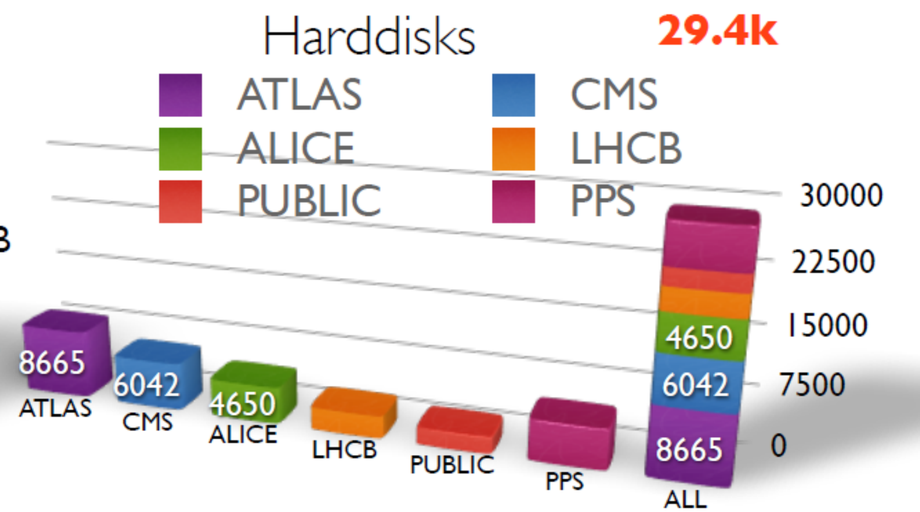
- disk space savings compared to file replication

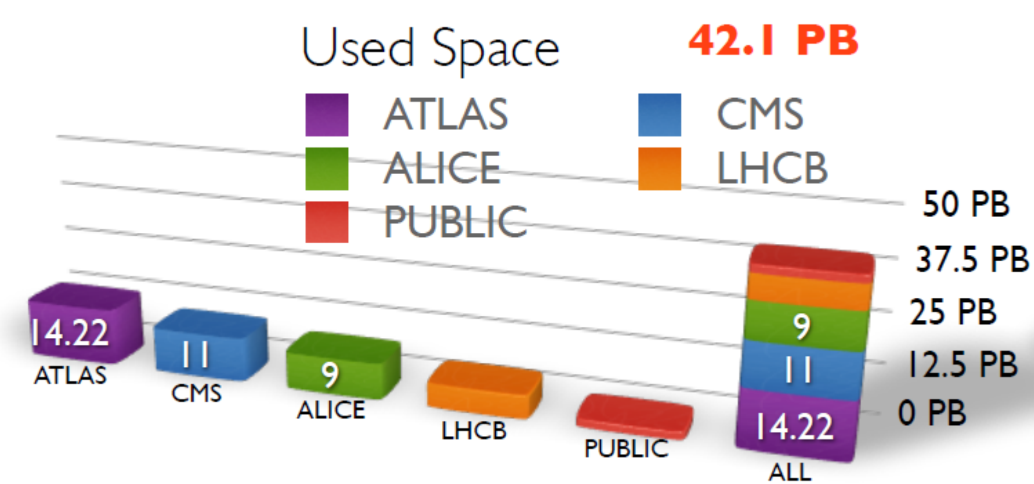
  e.g. 25% less space needed for (4+2) vs. 2 replica

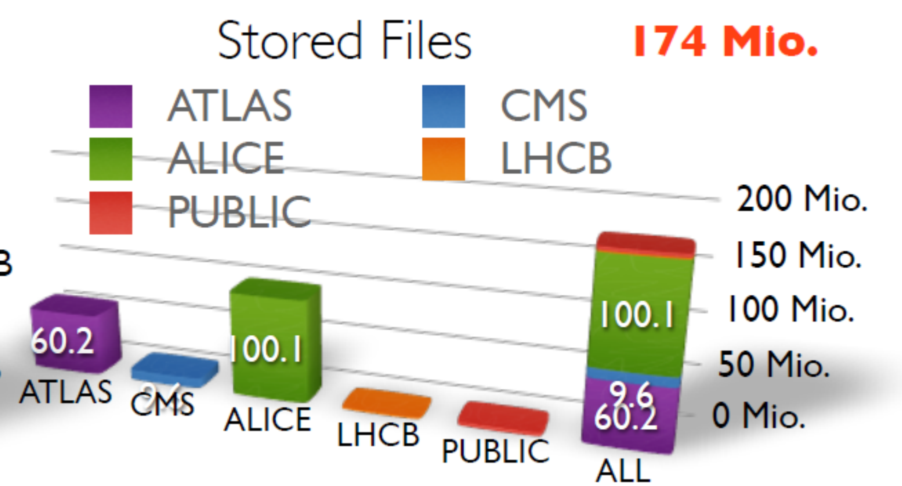
# DSS EOS in production at CERN

CERN IT Department




*EOSPUBLIC started 4.6.2013


# EOS in production



- Five multi-PB installations
  - few thousand disks per instance
- Simplified life-cycle management workflows for on-going replacement/repair of hardware
- JBOD disks (no RAID controller) using software RAIN
- low-latency file access with in-memory namespace - 174M files
- fine-grained access control and quota management
- GRID storage element

# DSS



- Accessed from thousands of CERN-local and remote batch nodes with Kerberos and GSI authentication
- Full support for XRootD, GridFTP, HTTP(S) & WebDav protocol - and – mountable with FUSE as a remote file system