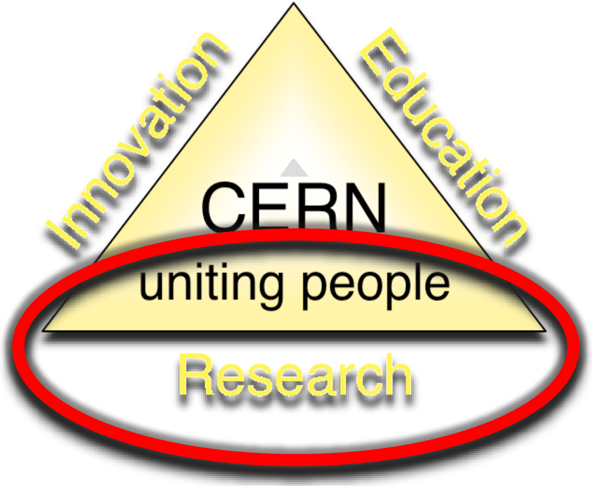
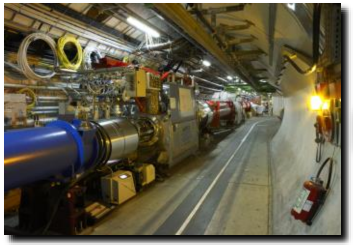
CERN

The mission of CERN



We are here



Accelerating particle beams

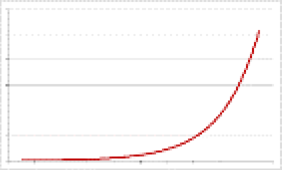
Detecting particles (experiments)

Large-scale computing (Analysis)

Discovery

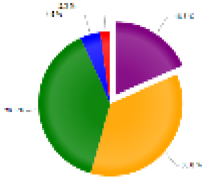
The need for computing in research

- Scientific research in recent years has exploded the computing requirements
  - Computing has been the strategy to reduce the cost of traditional research



At constant cost, exponential growth of performances

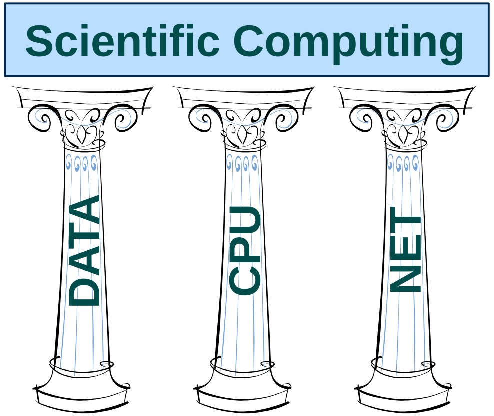
  - Computing has opened new horizons of research not only in High Energy Physics



Return in computing investment higher than other fields: Budget available for computing increased, growth is more than exponential

The need for storage in computing

- Scientific computing for large experiments is typically based on a distributed infrastructure
- Storage is one of the main pillars
- Storage requires Data Management…



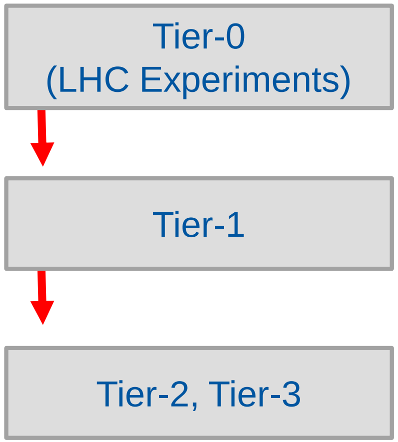
Roles Storage Services

- Three main roles
  - Storage (store the data)
  - Distribution (ensure that data is accessible)
  - Preservation (ensure that data is not lost)

Size in PB + performance

Availability

Reliability



Data production,
Storage, data distribution, data preservation

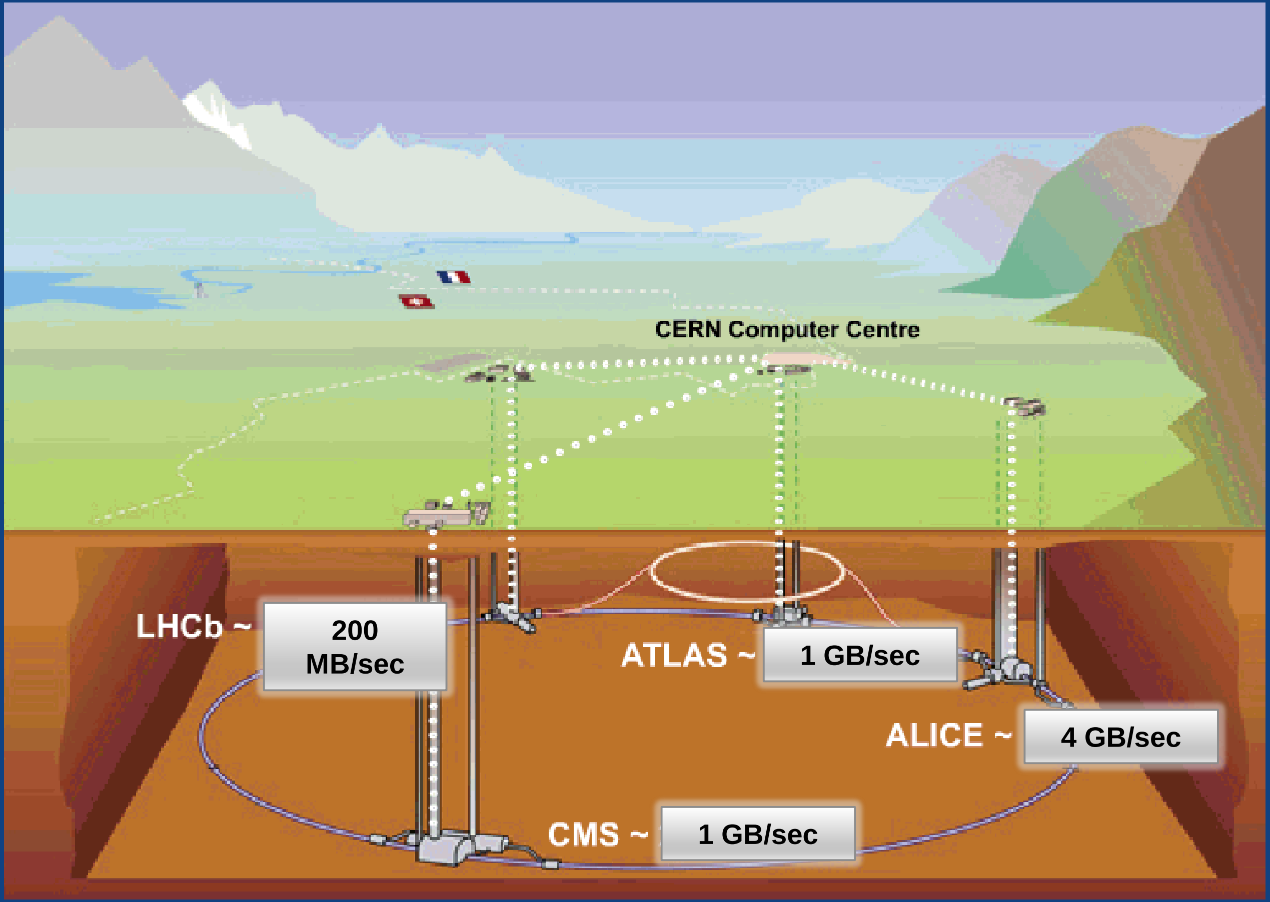
Storage, data distribution, data preservation

Storage, data distribution

“Why” data management ?

- Data Management solves the following problems
  - Data reliability
  - Access control
  - Data distribution
  - Data archives, history, long term preservation
  - In general:
    - Empower the implementation of a workflow for data processing

At CERN: Acquisition, First pass reconstruction, Storage, Distribution, and Data Preservation



CERN Computing Infrastructure

Overview: Data Centre 27-Nov-1015 @ 11:05 a day ago to a few seconds ago

MEYRIN DATA CENTRE

	last_value
Number of Cores in Meyrin	121,255
Number of Drives in Meyrin	70,847
Number of 10G NIC in Meyrin	5,587
Number of 1G NIC in Meyrin	21,707
Number of Processors in Meyrin	21,533
Number of Servers in Meyrin	11,598
Total Disk Space in Meyrin (TB)	122,909
Total Memory Capacity in Meyrin (TB)	480

WIGNER DATA CENTRE

	last_value
Number of Cores in Wigner	43,360
Number of Drives in Wigner	23,184
Number of 10G NIC in Wigner	1,399
Numer of 1G NIC in Wigner	5,071
Number of Processors in Wigner	5,422
Number of Servers in Wigner	2,714
Total Disk Space in Wigner (TB)	71,745
Total Memory Capacity in Wigner (TB)	172

NETWORK AND STORAGE

	last_value
Tape Drives	104
Tape Cartridges	26,340
Data Volume on Tape (TB)	125,493
Free Space on Tape (TB)	40,224
Routers (GPN)	134
Routers (TN)	29
Routers (Others)	97
Switches	3,574

CPUs

Network

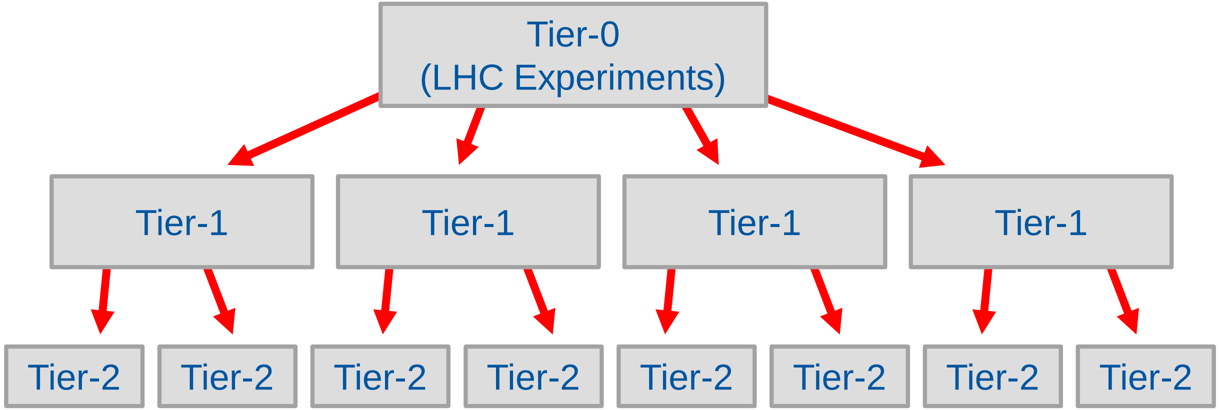
Databases

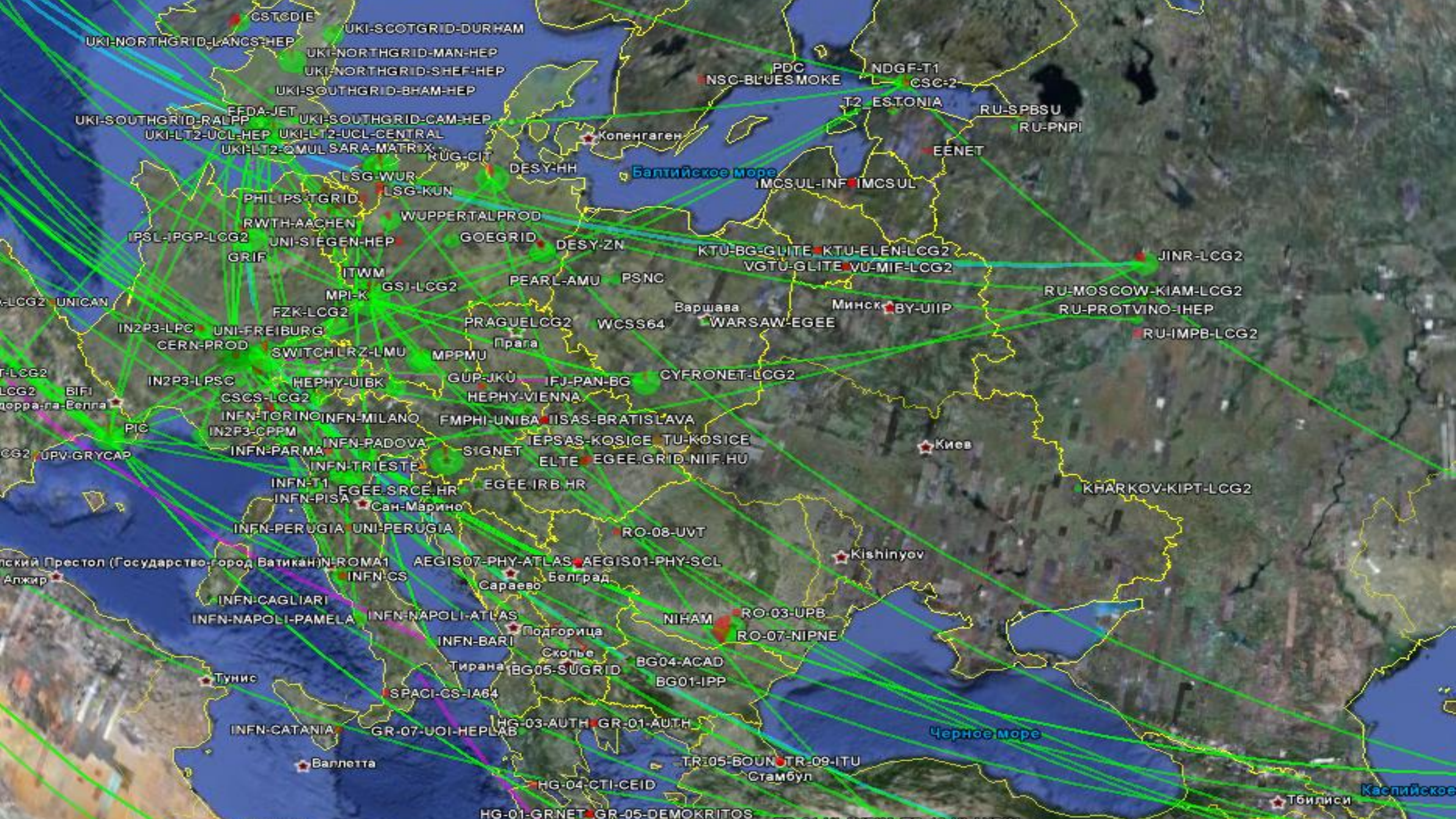
Storage

Infrastructur
e

Dataflows for Science

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



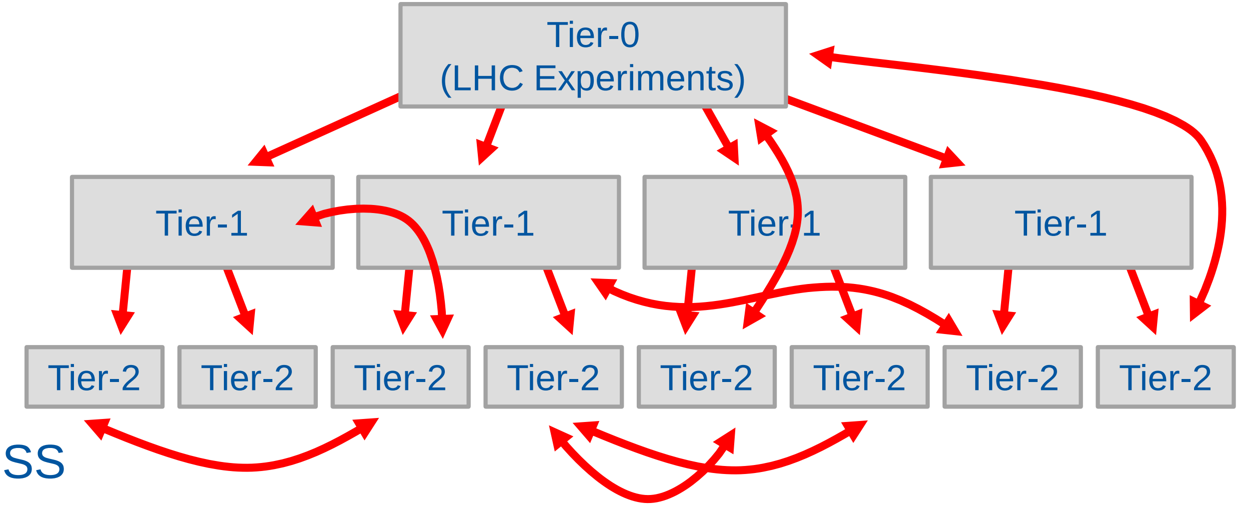

CSTCDIE
UKI-SCOTGRID-DURHAM
UKI-NORTHGRID-LANCS-HEP
UKI-NORTHGRID-MAN-HEP
UKI-NORTHGRID-SHEF-HEP
UKI-SOUTHGRID-BHAM-HEP
EFDA-JET
UKI-SOUTHGRID-RALPP
UKI-SOUTHGRID-CAM-HEP
UKI-LT2-UCL-HEP
UKI-LT2-UCL-CENTRAL
UKI-LT2-QMUL
SARA-MATRIX
RUG-CIT
LSG-WUR
LSG-KUN
DESY-HH
PHILIPS-TGRID
RWTH-AACHEN
WUPPERTALPROD
IPSL-IPGP-LCG2
UNI-SIEGEN-HEP
GOEGRID
DESY-ZN
GRIF
ITWM
GSI-LCG2
PEARL-AMU
PSNC
MPI-K
FZK-LCG2
UNICAN
IN2P3-LPC
UNI-FREIBURG
PRAGUELCG2
WCSS64
Прага
CERN-PROD
SWITCH
LRZ-LMU
MPPMU
IN2P3-LPSC
HEPHY-UIBK
GUP-JKU
IFJ-PAN-BG
CYFRONET-LCG2
BIFI
CSCS-LCG2
HEPHY-VIENNA
INFN-TORINO
INFN-MILANO
FMPHI-UNIBA
IISAS-BRATISLAVA
PIC
IN2P3-CPPM
IEPSAS-KOSICE
TU-KOSICE
INFN-PARMA
INFN-PADOVA
SIGNET
UPV-GRYCAP
INFN-TRIESTE
ELTE
EGEE.GRID.NIIF.HU
INFN-T1
EGEE.SRCE.HR
EGEE.IRB.HR
INFN-PISA
Сан-Марино
INFN-PERUGIA
UNI-PERUGIA
RO-08-UVT
INFN-ROMA1
AEGIS07-PHY-ATLAS
AEGIS01-PHY-SCL
INFN-CS
Белград
Сараево
Алжир
INFN-CAGLIARI
INFN-NAPOLI-PAMELA
INFN-NAPOLI-ATLAS
Подгорица
INFN-BARI
Скопье
Тирана
BG05-SUGRID
BG04-ACAD
BG01-IPP
Тунис
SPACI-CS-IA64
INFN-CATANIA
GR-07-UOI-HEPLAB
HG-03-AUTH
GR-01-AUTH
Валлетта
HG-04-CTI-CEID
HG-01-GRNET
GR-05-DEMOKRITOS
PDC
NSC-BLUESMOKE
NDGF-T1
CSC-2
T2_ESTONIA
RU-SPBSU
RU-PNPI
Копенгаген
EENET
Балтийское море
IMCSUL-INF
IMCSUL
KTU-BG-GLITE
KTU-ELEN-LCG2
VGTU-GLITE
VU-MIF-LCG2
JINR-LCG2
RU-MOSCOW-KIAM-LCG2
Минск
BY-UIIP
RU-PROTVINO-IHEP
Варшава
WARSAW-EGEE
RU-IMPB-LCG2
Киев
KHARKOV-KIPT-LCG2
Kishinyov
NIHAM
RO-03-UPB
RO-07-NIPNE
Черное море
TR-05-BOUN
TR-09-ITU
Стамбул
Тбилиси

Computing Services for Science

- 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
- 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

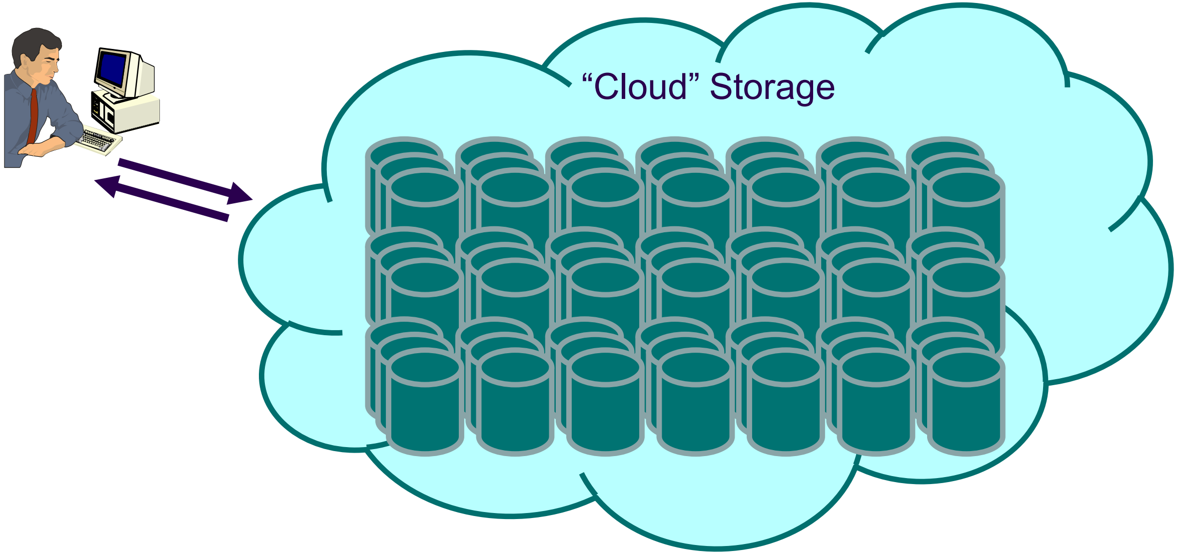


Why storage is complex ?

- Both approaches coexists
- Data is pre-placed
  - It is the role of the experiments that plans the analysis
- Data globally accessible and federated in a global namespace –the middleware is used for access
  - always attempt to take the local data or redirects to the nearest remote copy
  - jobs designed to minimize the impact of the additional latency that the redirection requires

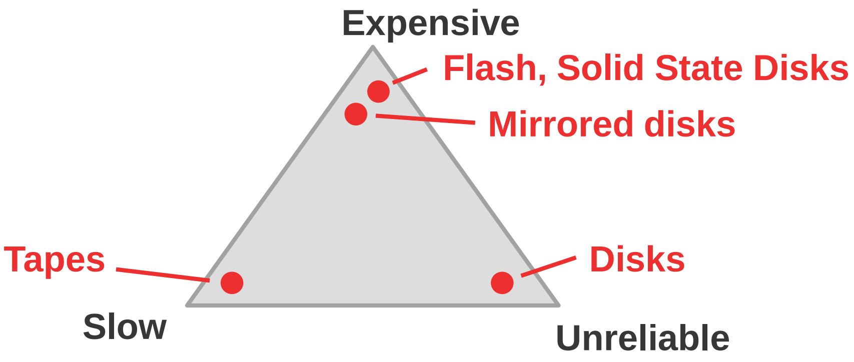
Which storage model ?

- 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



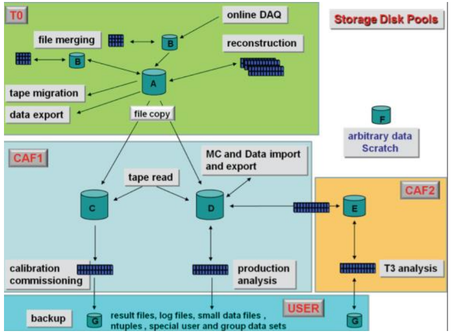
… but some limitations

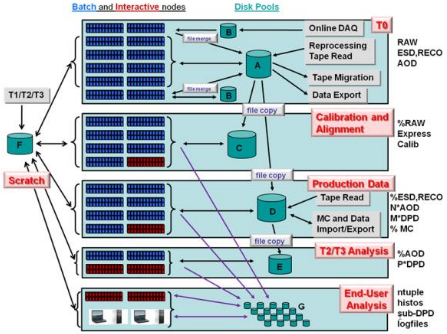
- Different storage solutions can offer different quality of services
  - Three parameters: Performance, Reliability, Cost
  - You can have two but not three
- To deliver both performance and reliability you must deploy expensive solutions
  - Ok for small sites (you save because you have a simple infrastructure)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

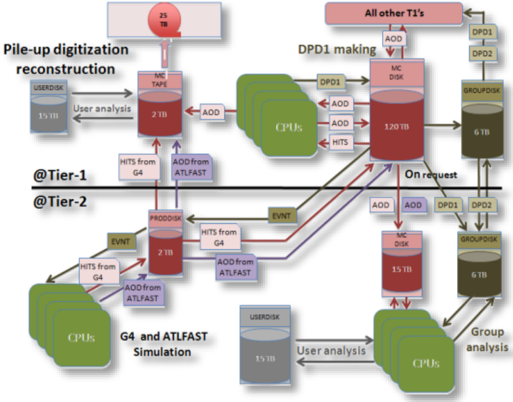


So, … what is data management ?

- Examples from LHC experiment data models



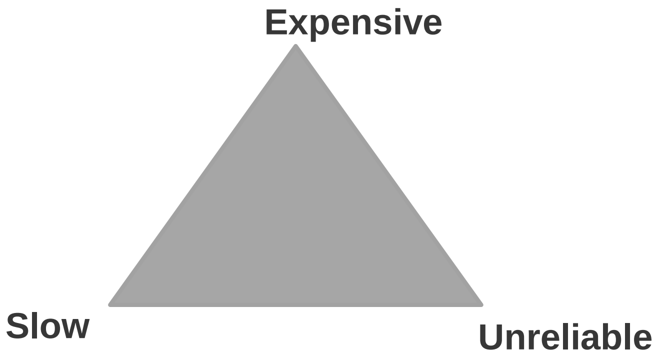




- Two building blocks to empower data processing
  - Data pools with different quality of services
  - Tools for data transfer between pools

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

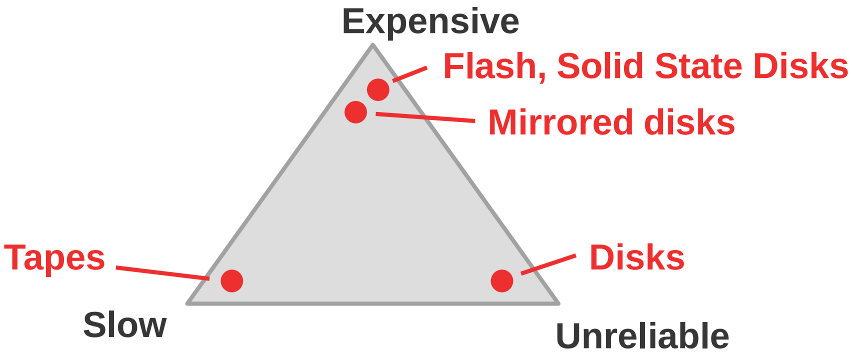


Why multiple pools and quality ?

- 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

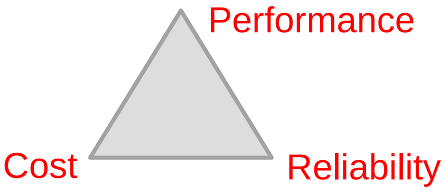
Data pools

- Different quality of services
  - Three parameters: (Performance, Reliability, Cost)
  - You can have two but not three

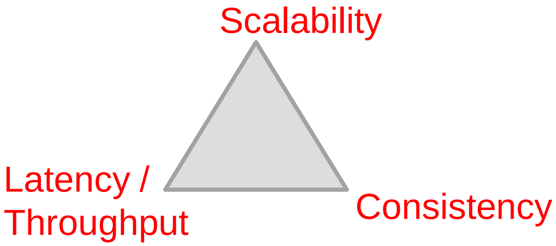


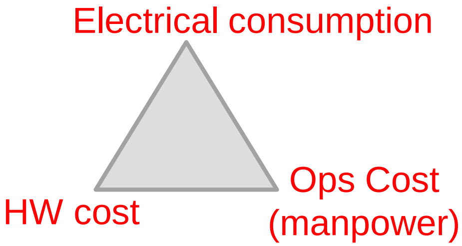
… and the balance is not simple

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



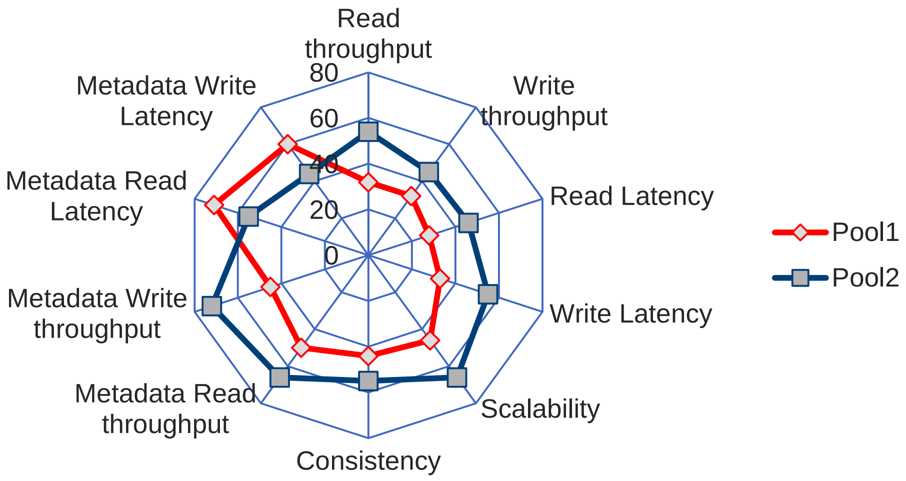
- Performance has many sub-parameters
- Cost has many sub-parameters
- Reliability has many sub-parameters





… and reality is complicated

- Key requirements: Simple, Scalable, Consistent, Reliable, Available, Manageable, Flexible, Performing, Cheap, Secure.
- Aiming for on-demand “quality of service”
- And what about is scalability ?

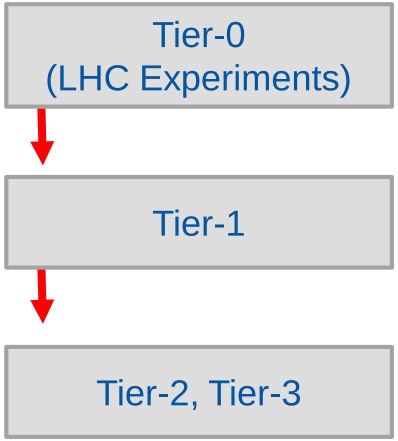


Tools needed

- 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 + performance
  - Distribution (ensure that data is accessible) — Availability
  - Preservation (ensure that data is not lost) — Reliability



Tier-0: Data production,
Storage, data distribution, data preservation

Tier-1: Storage, data distribution, data preservation

Tier-2, Tier-3: Storage, data distribution

Multi site transfers (200 Gbps)

Reliability …

- You can achieve high reliability using standard disk pools
  - Multiple replicas
  - Erasure codes
  - (beware of independence of failures)
- Here is where tapes can play a role
  - Tapes ??

Do we need tapes ?

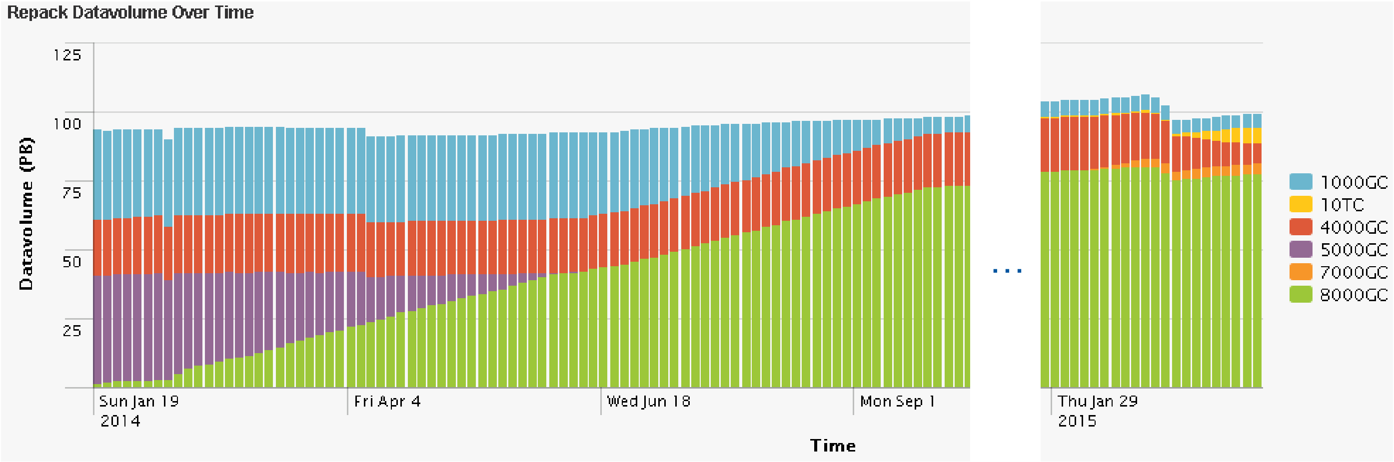
- Tapes have a bad reputation in some use cases
  - Slow in random access mode
    - high latency in mounting process and when seeking data (F-FWD, REW)
  - Inefficient for small files (in some cases)
  - Comparable cost per (peta)byte as hard disks

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 (4x faster)
  - Several orders of magnitude more reliable than disks
    - Few hundreds GB loss per year on 80 PB raw tape repository
    - Few hundreds TB loss per year on 50 PB raw 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
  - Cannot be deleted in few minutes
- Bottom line: if not used for random access, tapes have a clear role in the architecture

Large scale media migration



Summary

- Data Management solves the following problems
  - Data reliability
  - Access control
  - Data distribution
  - Data archives, history, long term preservation
  - In general:
    - Empower the implementation of a workflow for data processing

www.cern.ch