Data storage at CERN

Overview:

- Some CERN / HEP specifics
  - Where does the data come from, what happens to it
- General-purpose data storage @ CERN
- Outlook

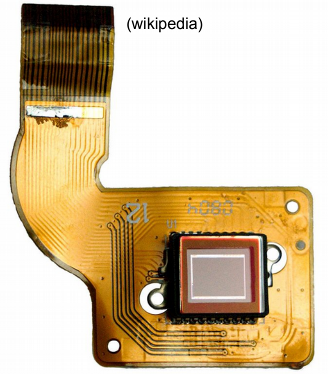
CERN vs Experiments

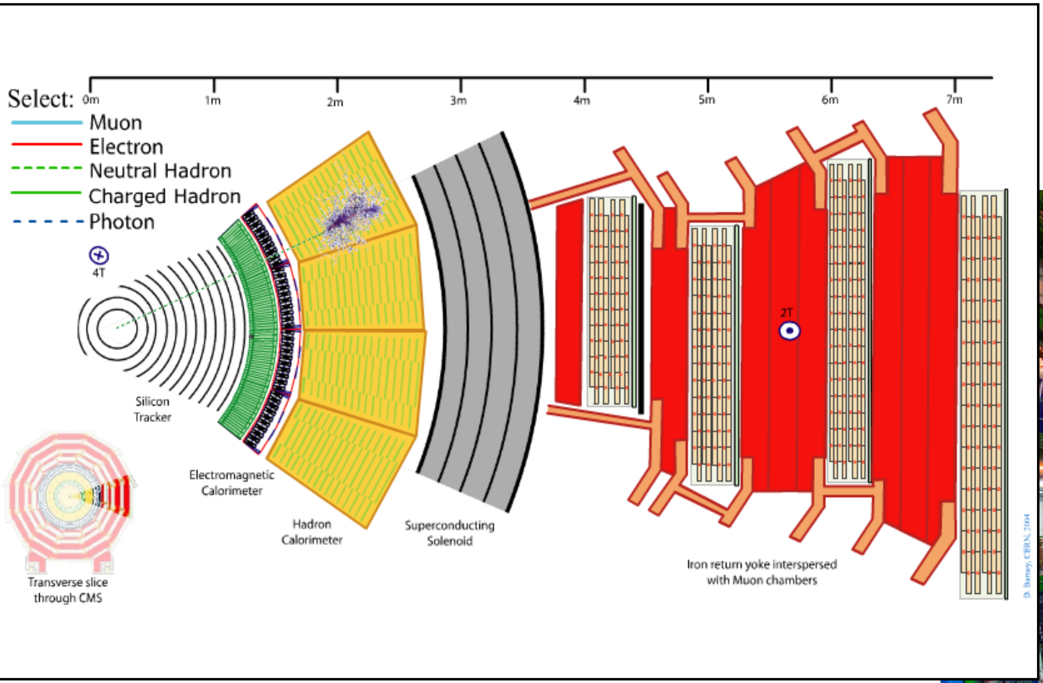
- CERN = (Conseil* Européen* pour la Recherche Nucléaire*) → European Organization for Nuclear Research
  - Est.1954, international treaty, 21 states
  - Provides lab facilities: water, electricity, cooling, offices, network, computing, various flavours of particle beams,..
  - ~2300 staff – clerical, engineers, firemen, ..
- Experiments: international scientific collaborations, own funding
  - "HEP": high-energy physics
  - Build & install detectors
  - Use lab facilities ("MoU")
    - Computing: yearly review
  - Generate & use & manage data
  - 10...3000 physicists each
- WLCG: computing grid for LHC

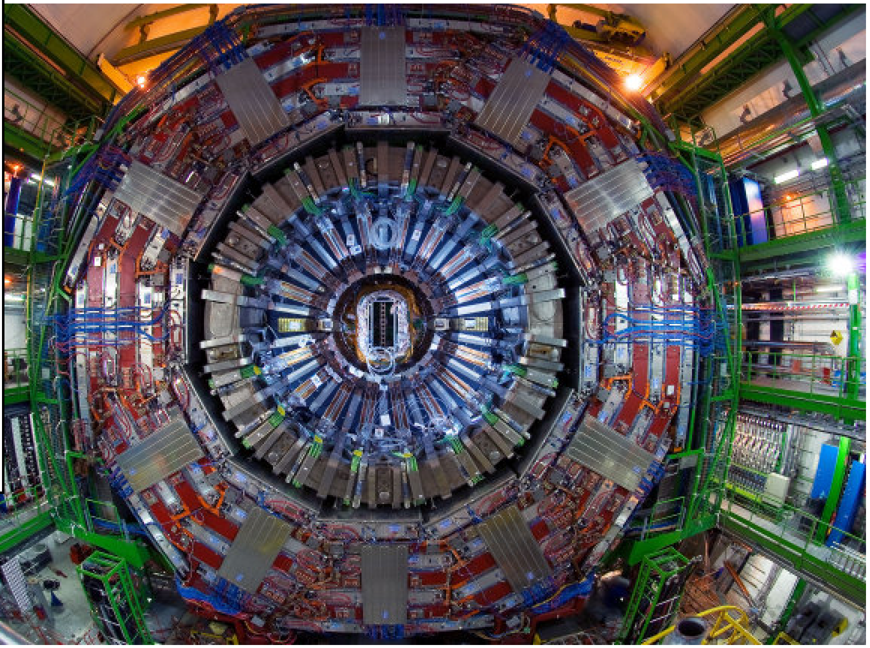
*: wrong, nowadays

“Data taking” in HEP

- Think “digital video camera”..
  - Unwieldy & complicated & expensive

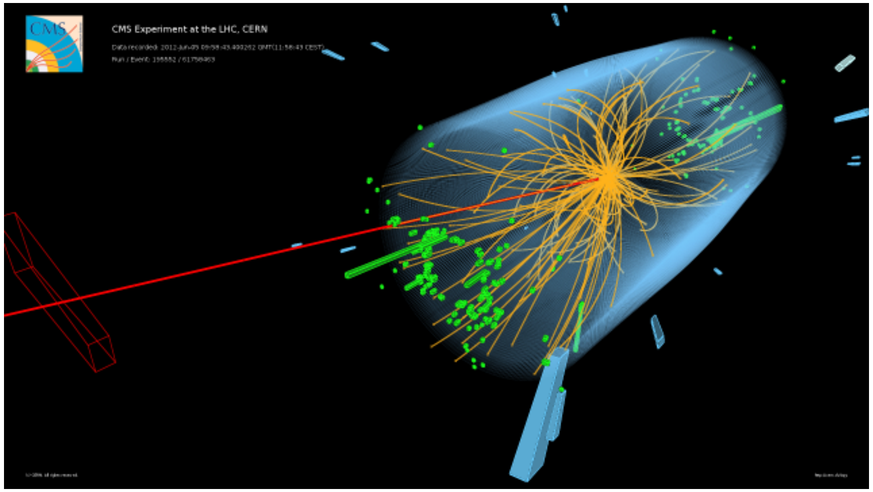




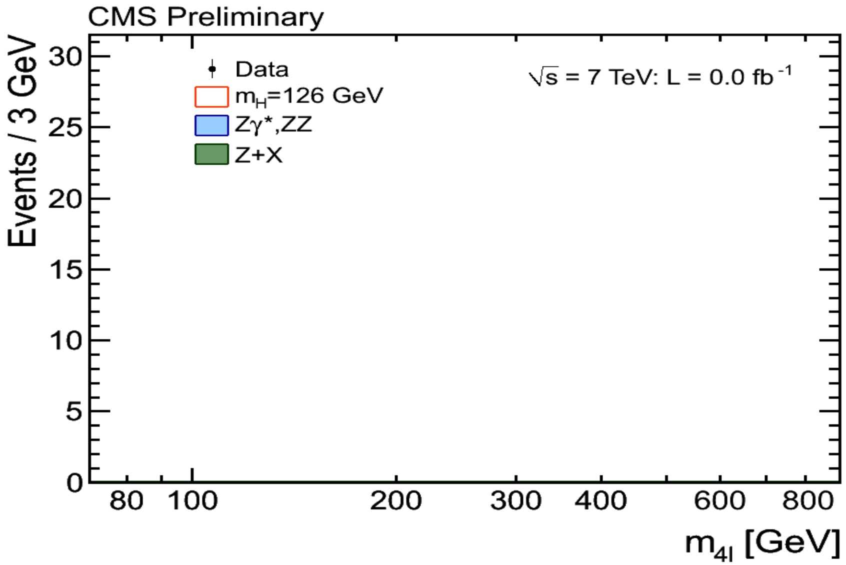


“Data taking”

- .. but gives 4d “pictures”


CMS
CMS Experiment at the LHC, CERN
Data recorded: 2012-Jun-05 09:58:43.400262 GMT(11:58:43 CEST)
Run / Event: 195552 / 61758463

- Significant postprocessing required: calibration, track reconstruction

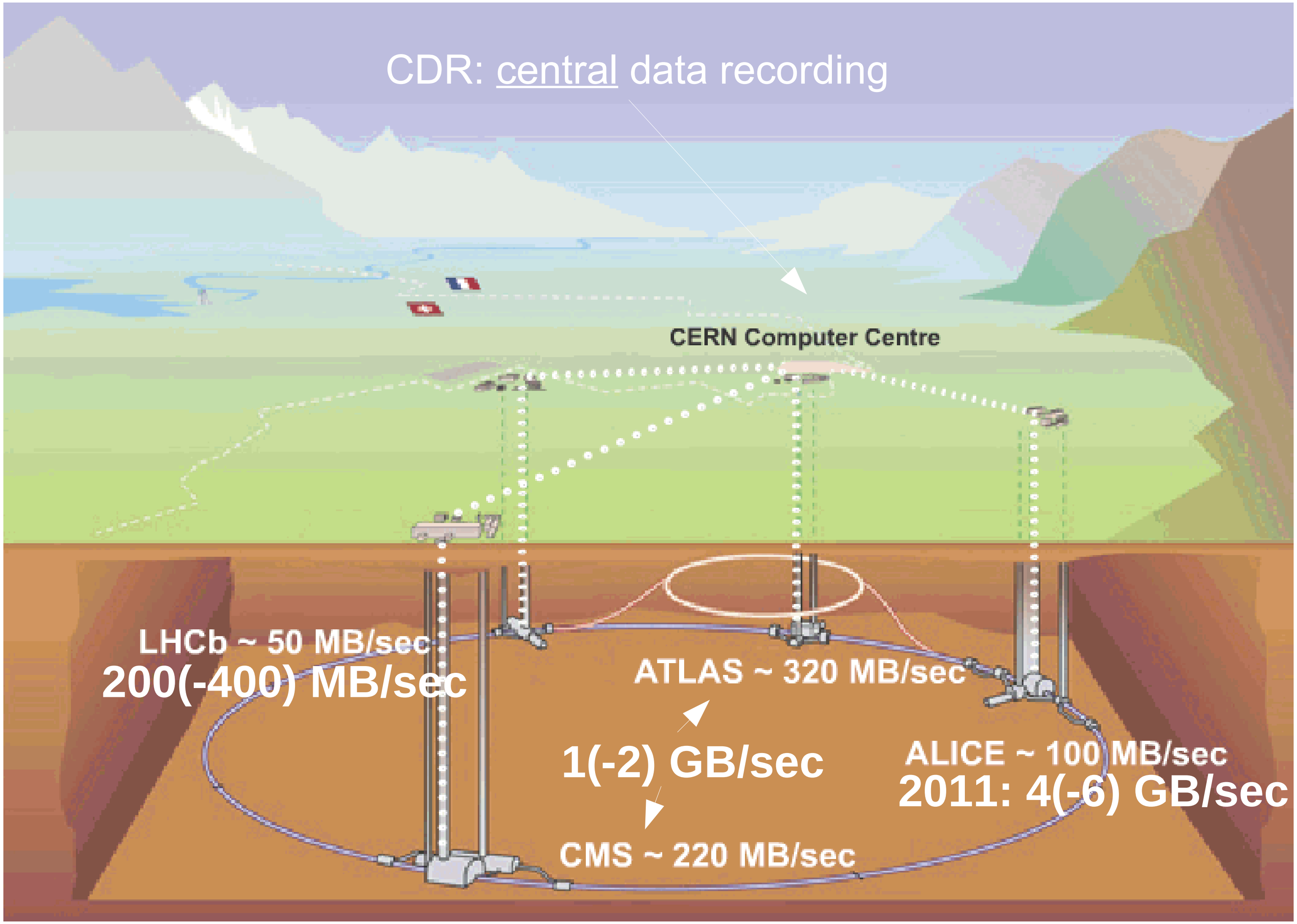


- .. and subsequent analysis

- Result: good statistics on very rare events ≈

scientific papers.

Tier 0 at CERN: Acquisition, First pass reconstruction, Storage & Distribution

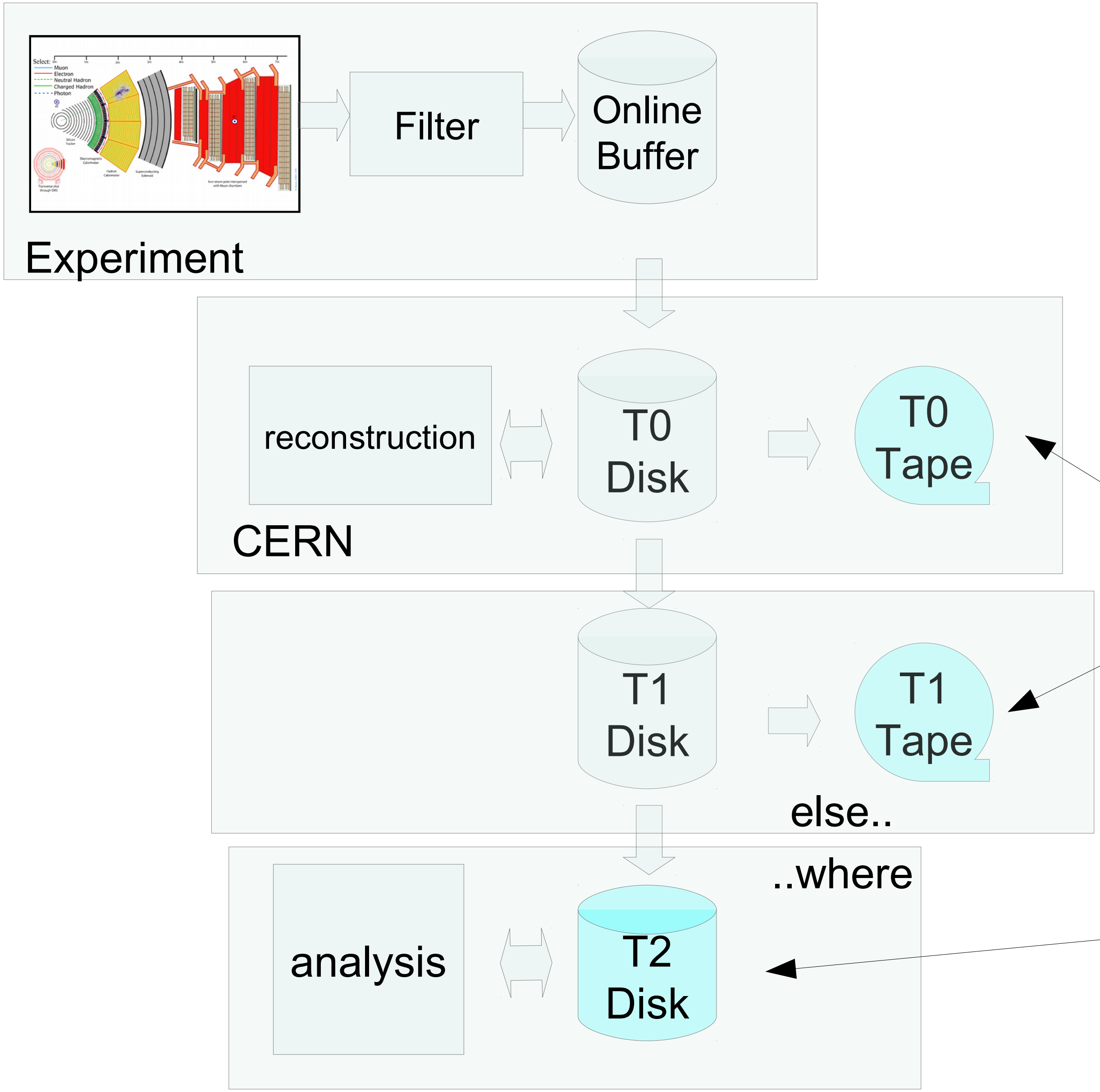


Some history of scale…

Date	Collaboration sizes	Data volume, archive technology
Late 1950's	2-3	Kilobits, notebooks
1960’s	10-15	kB, punchcards
1970’s	~35	MB, tape
1980’s	~100	GB, tape, disk
1990’s	700-800	TB, tape, disk
2010’s	~3000	PB, tape, disk

For comparison:
1990’s: Total LEP data set ~few TB
Would fit on 1 tape today
Today: 1 year of LHC data ~25 PB

HEP data flow - schematic



Filter

Online Buffer

Experiment

reconstruction

T0 Disk

T0 Tape

CERN

T1 Disk

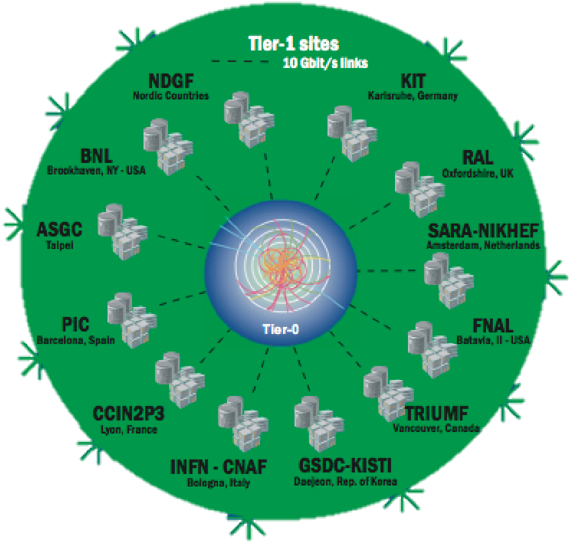
T1 Tape

else..

..where

analysis

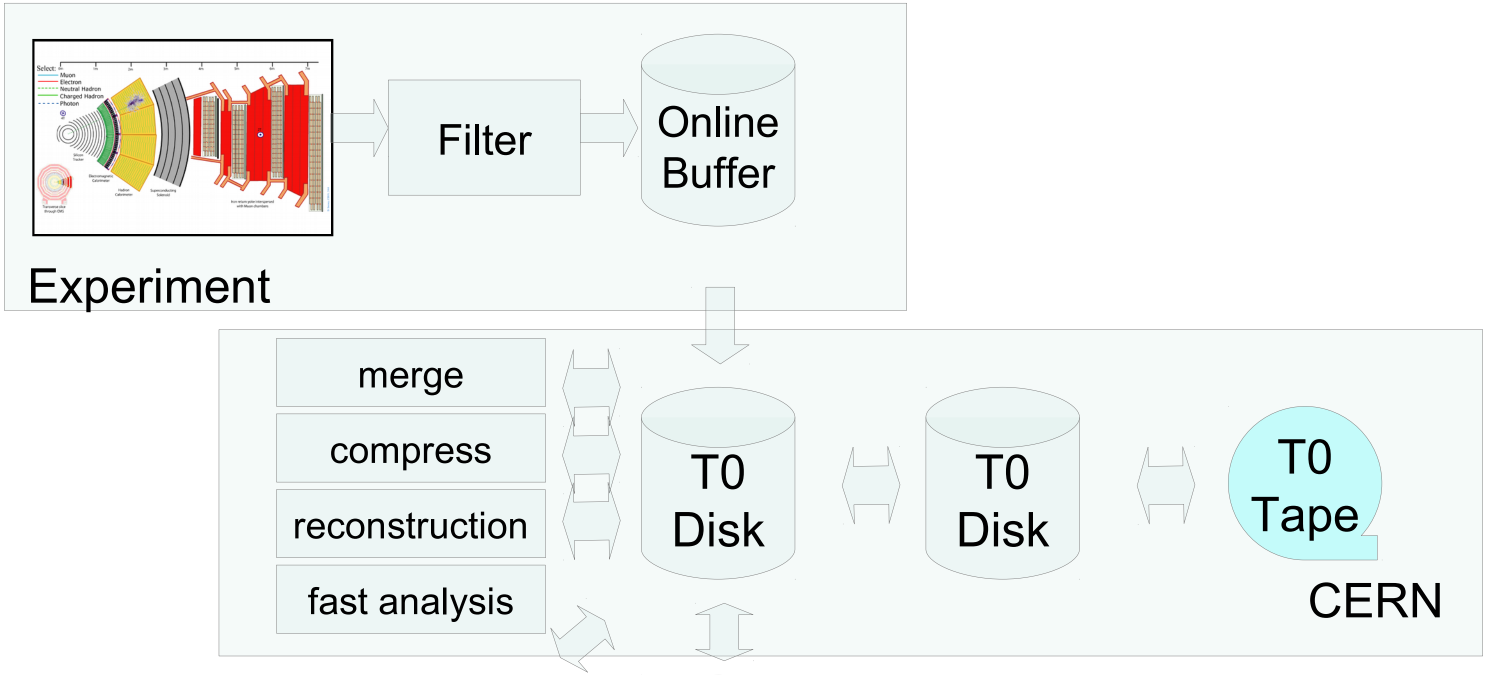
T2 Disk

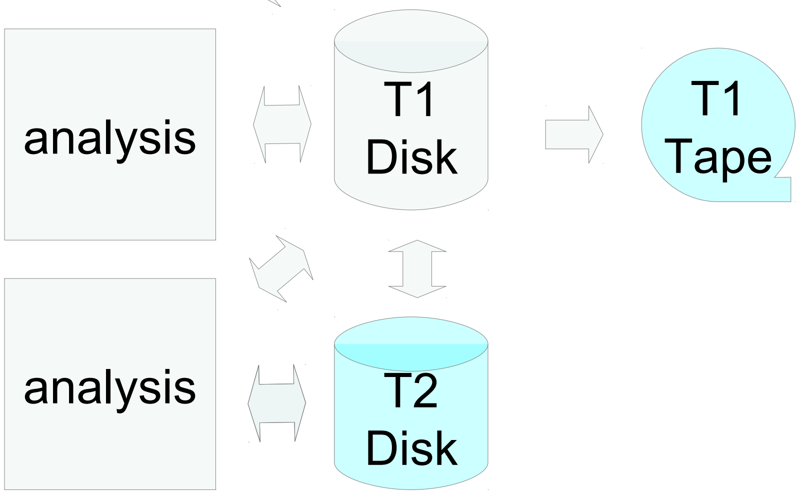


Custodial copy

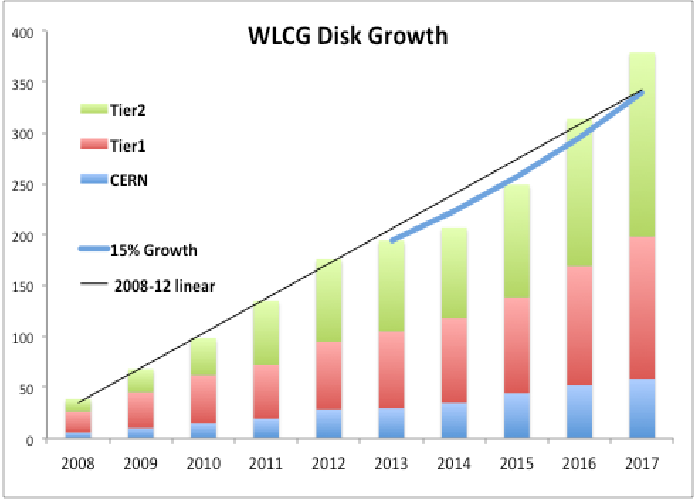
Working copies

HEP data flow - realistic

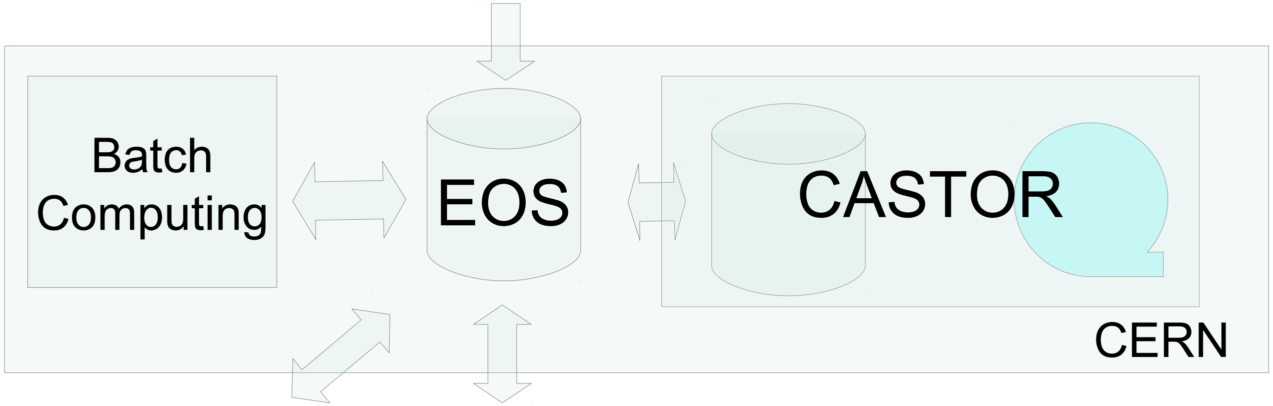




Side note: CERN is small..



Physics data - CERN IT part



- Physics Storage systems in CERN-IT:
  - **CASTOR**: HSM
  - **EOS**: diskonly low-latency access, recent
- Both:
  - Homegrown
  - [Non|HEP]-standard protocols (**XrootD**, RFIO, SRM, gridftp)

CASTOR HSM

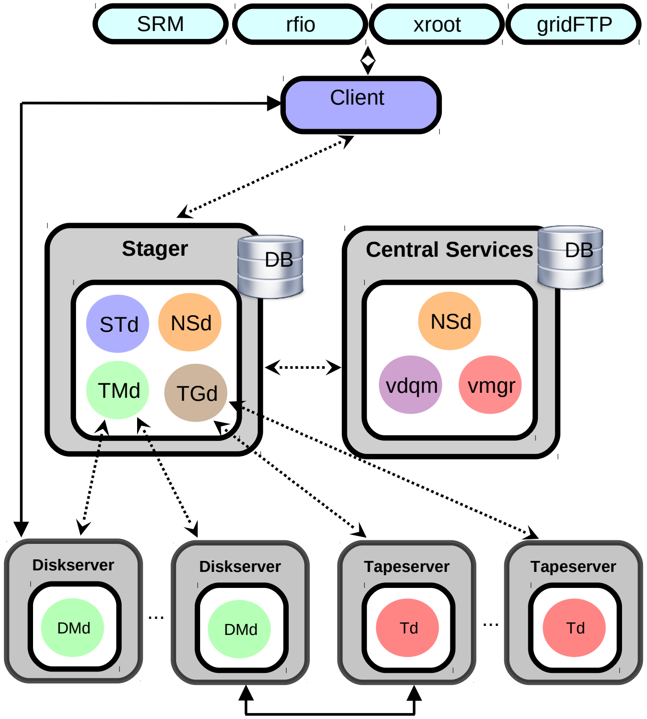
CASTOR
CERN Advanced STORage manager

92PB 316M 350k

- Born in 1999
- Common Namespace
- Main Role: **data recording**,

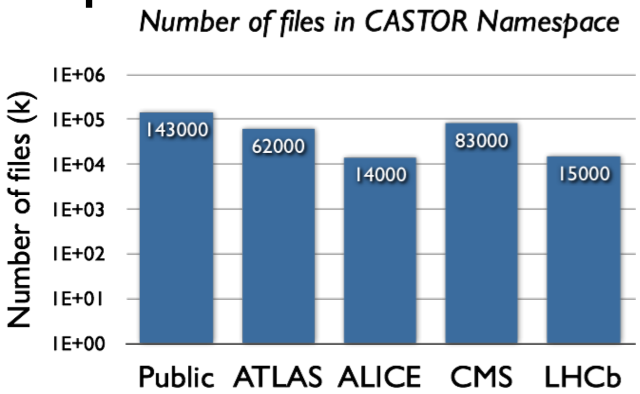
  Tier-1 data export, production activities
- Mainly **tape-backed data**
- Focus on tape performance

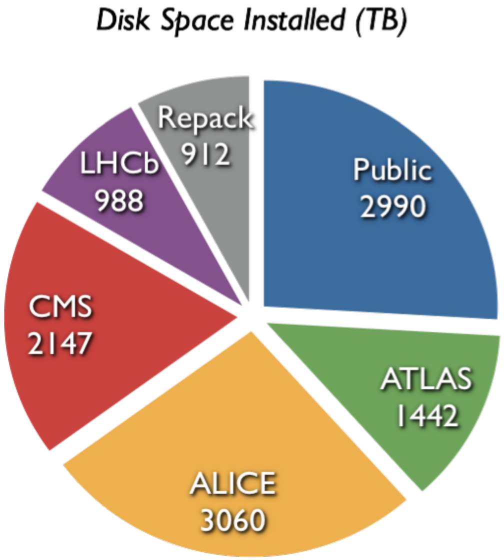
  (latency can be high)
- **Database** centric
- Not optimized for concurrent access:
  - Currently Raid-1 configuration
- **Aimed** at DAQ activities: limited transfer slots = QoS
- No (real) quotas..



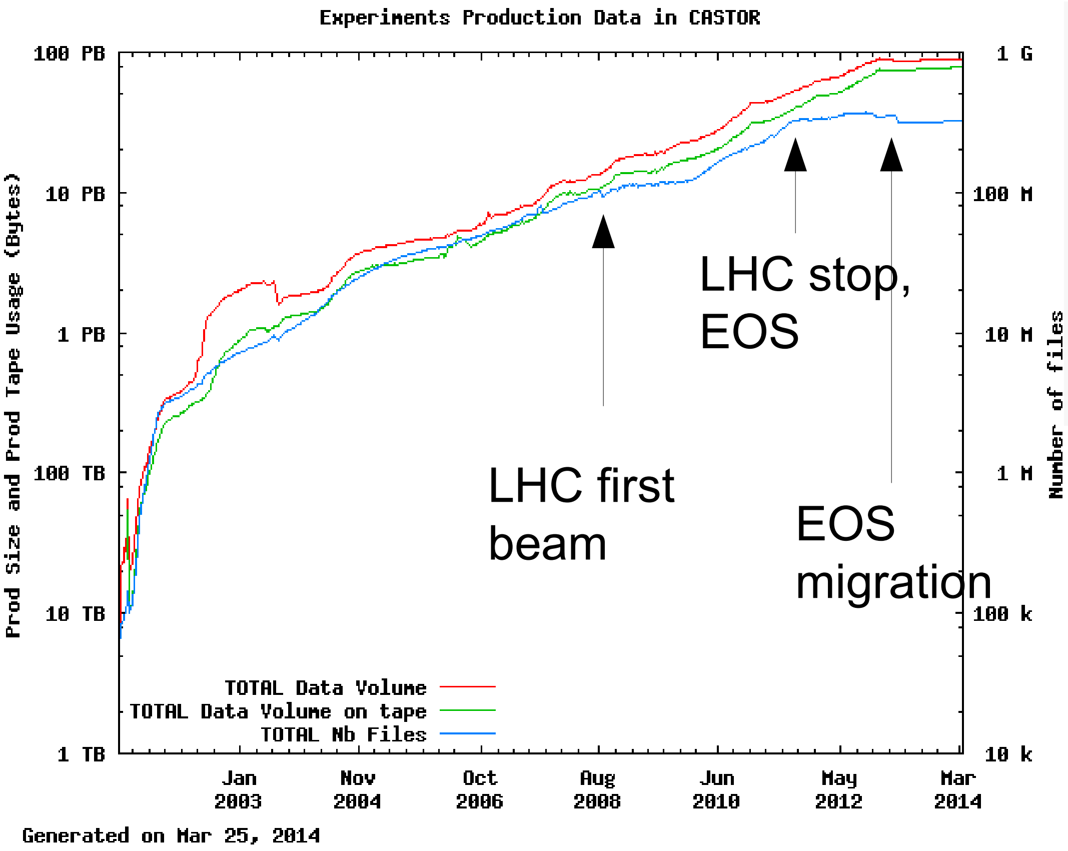
CASTOR: current setup

- 7 instances:
  - LHCs: ATLAS, ALICE, CMS and LHCb
  - PUBLIC: users, non-LHC experiments and specific pool for DAQ activities for AMS, COMPASS, NA61, NA62
  - Repack (tape media migration and compacting) and PPS (pre-production)
- Totals:
  - 92PB, 316M files, ~650 diskservers
- Mature release cycle:
  - ~1 major release per year
- In production also at RAL and ASGC





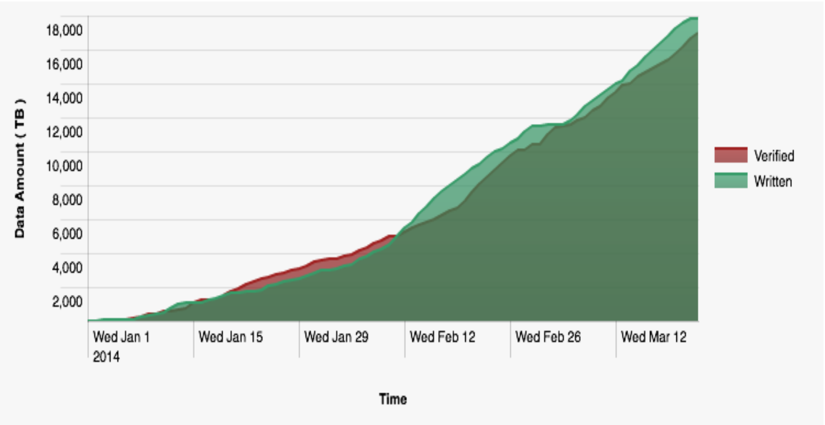
CASTOR historical data



LHC first beam

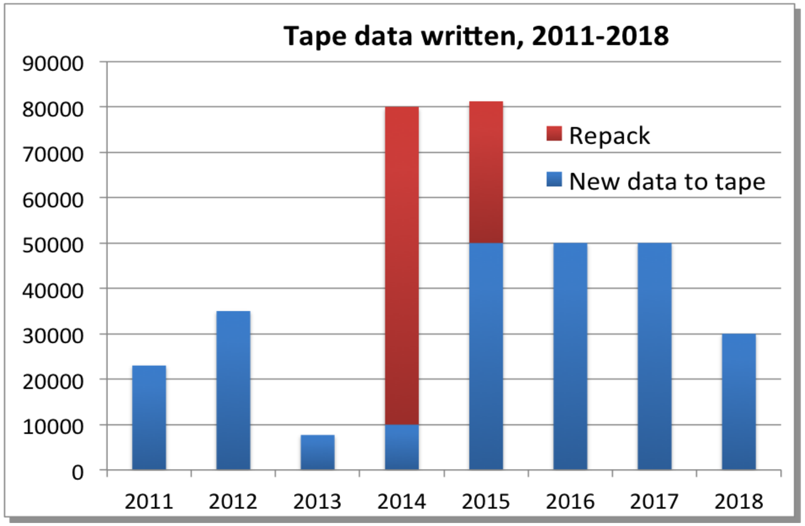
LHC stop, EOS

EOS migration



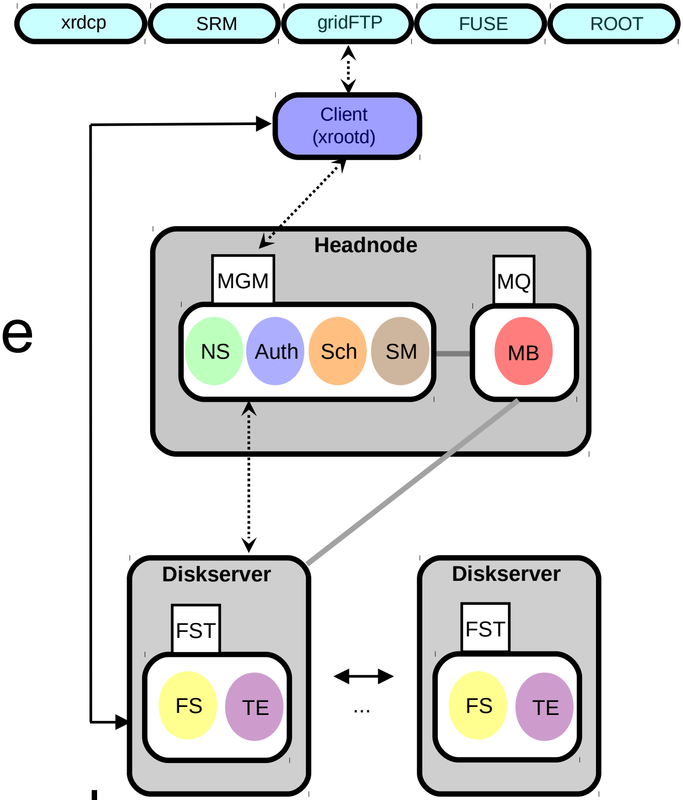
Tape “cold” data verification

Repacking tapes is major activity



EOS

25PB 158M 73k

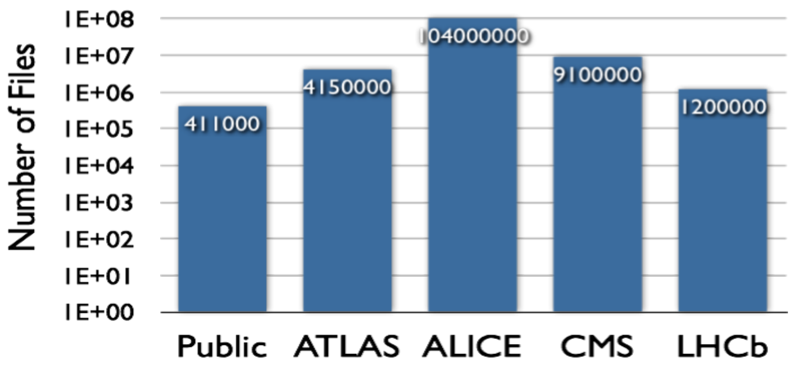


- Born in 2010
- **in-memory namespace** – split per instance
- Main role: end-user analysis
- **Disk-only** storage
- Focus on **low latency**
- Optimized for **concurrency**
- **Multi-replica** on different diskservers
- No limit on transfer slots – throttle via overload
- **Quota** system: users&groups (for volume and files)
- **Strong** authentication: krb5, X509
- Diskserver: **JBOD** configuration

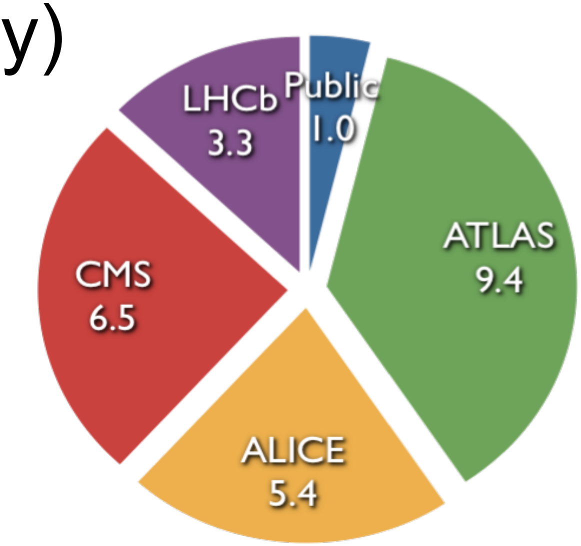
EOS: current setup

- 6 instances:
  - LHCs: ATLAS, ALICE, CMS and LHCb
  - PUBLIC: recently deployed - AMS and COMPASS experiments
- Totals:
  - 20PB, 158M files, ~1100 diskservers
- Release lifecycle driven by functionality
  - ~2 major release per year, constant updates
- Used at Fermilab (Tier-3 functionality)

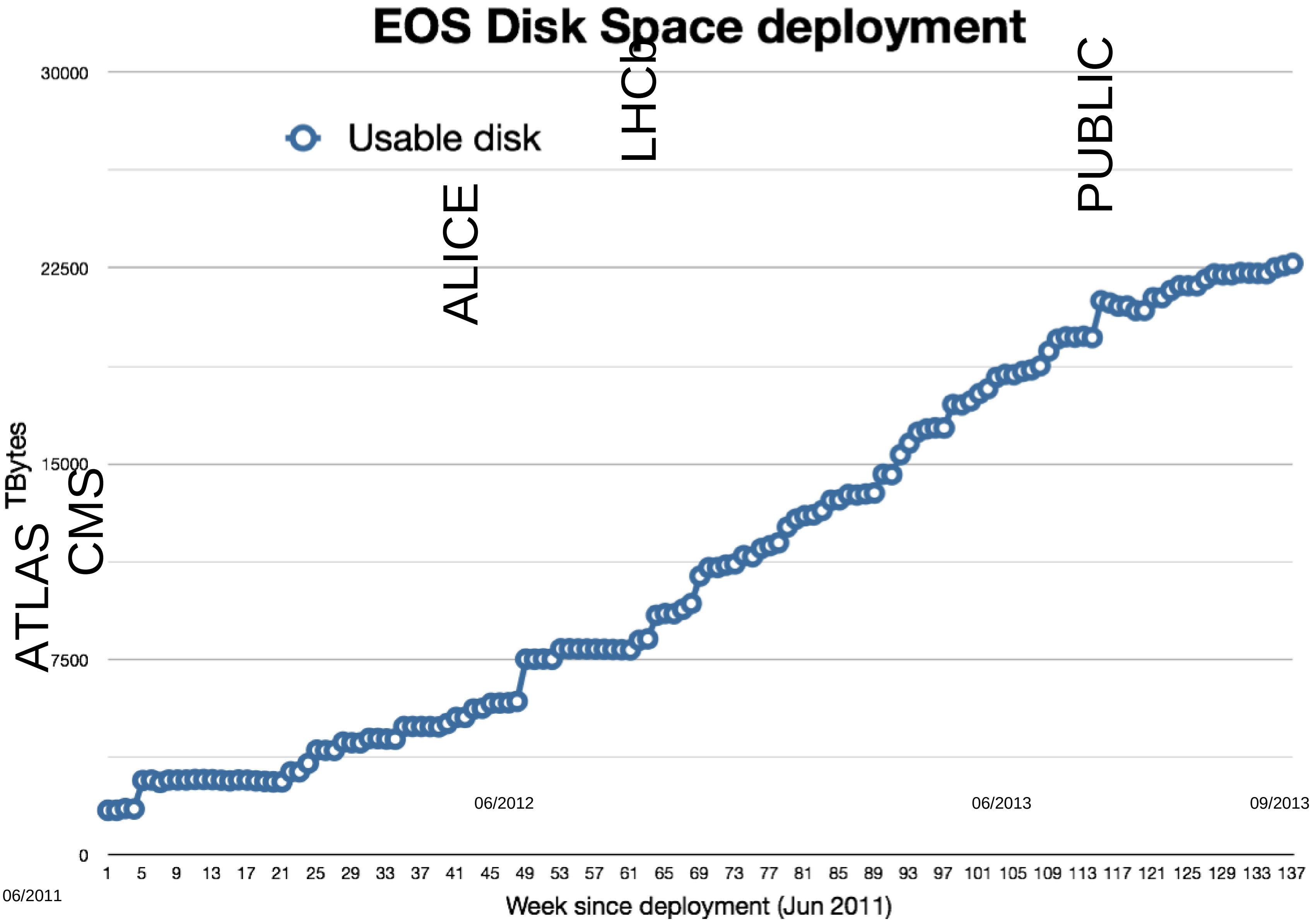
Number of files on EOS Namespace



Disk Space Installed (PB)



EOS deployment



other data storage services

(Data storage – everywhere at CERN...)

- Experiments:
  - ex. DAQ 'disk buffers' – up to several days of data taking
  - Often: prototype solution, but trouble with running long-term
- Department & group-scale independent solutions
  - CERN IT services ought to cover these..
  - .. but not always do.
  - (“NIH”-syndrome?)
- (Structured data / databases – not considered here)

Here: looking at CERN IT(-DSS) services.

General-Purpose storages

- General-purpose shared filesystems

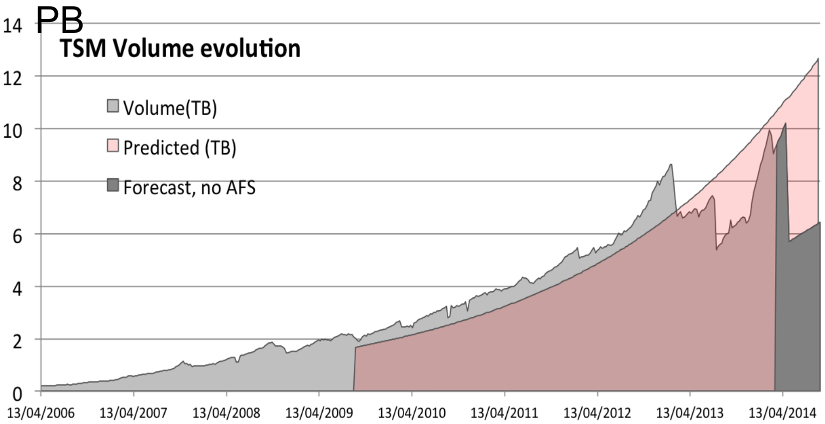
  Home directories = Untrusted clients = strong authentication

  - **AFS – Linux/Mac -** (see CERN site report), 950TB, 3G files
  - DFS - Windows-only
  - Future: (NFSv4), (FUSE-mounted EOS), (local FS+OwnCloud)
- “cluster filesystems”

  Typically: weak authentication

  - Not used for “computing” at CERN
    - general problem: “open” network + (too) many machines
    - CERN computing is “embarrassingly parallel”
    - CERN computing is worldwide
  - (experiments have own networks)
  - **NetApp** Filers: NFSv3 (190TB, 210M files)
    - Re-export: **CVMFS**, higher-level services (TWiki, VCs, ..) , DB
    - “standard”, “will not void warranties”

- Archive/tape
  - **TSM** (9PB, 2.1G files)
    - (See CERN AFS Backup talk)
    - Most machines are *not* backed up
  - (CASTOR)
  - (EOS reed-solomon)



- Block storage: **CEPH**

- R&D: Hadoop/HDFS, Huawei S3, SWIFT, ..

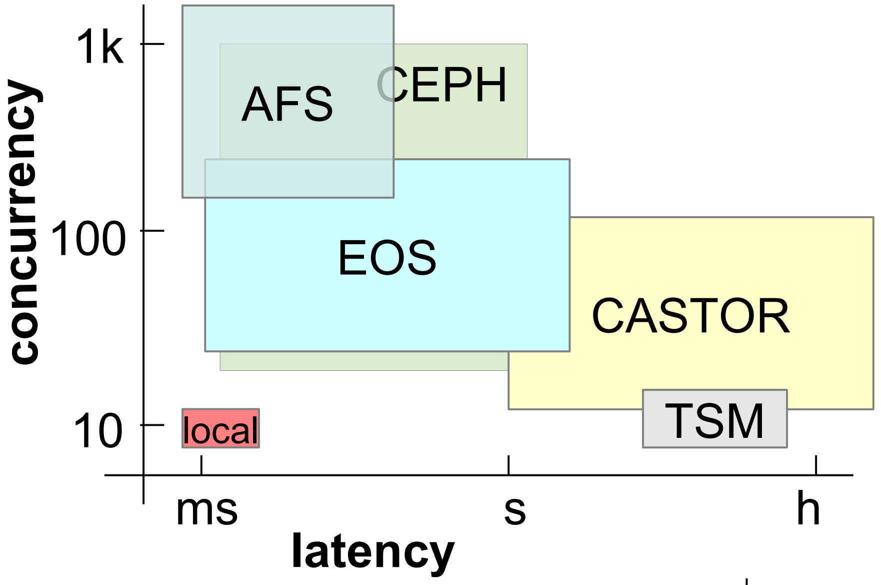
- lots of auxiliary storage services
  - protocol gateways (SRM, gridftp, http/webdav,..)
  - File transfer engines (FTS), **OwnCloud**
  - Bookkeeping, accounting, monitoring,...

Looks like a zoo?

- Yes. But: different Dimensions:
  - Usage
  - Size/#Files
  - I/O pattern
  - Layering
  - Lifecycle (eval/prototype/production/legacy)

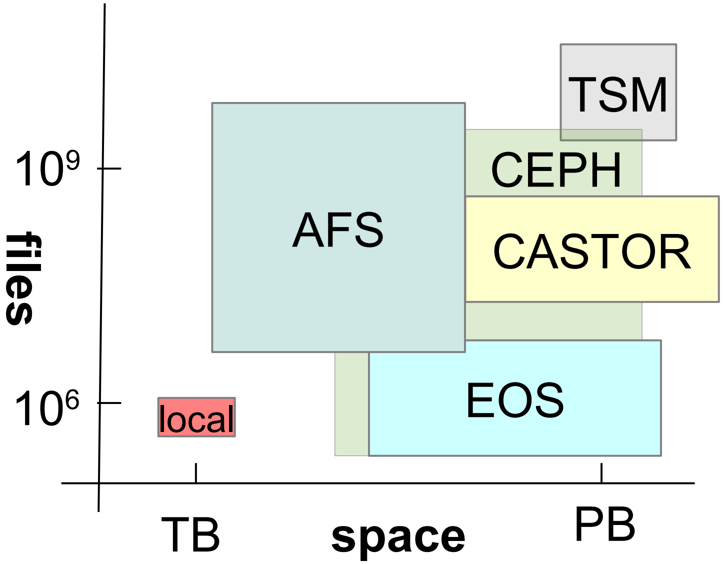
- “tool for the job”-approach
  - (and a bit of history. And Co-Evolution at work..)
  - Historically, lots of “use cases” started on AFS...
    - .. and once “big enough”, moved somewhere else. Mostly.

Size matters



(Orders-of-magnitude only. No sweat.)

- Map new use cases to “our” toolbox
  - Size, performance
  - Acceptable limitations



Our building blocks

- *batchusserverus communalus*
  - 32..64GB RAM; 2..3 HD; Intel or AMD; ScientificLinuxCERN
  - Used for (replicated) headnodes, metadata servers

+ SATA tray(s) + 10GbE →

- *discusserverus vulgaris*
  - 24x 3TB + 3x2TB, 1x 10GbE,
  - used by EOS, CASTOR, AFS, CEPH, TSM, HDFS
  - $$$/TBmonth varies by replication factor (+manpower)

- Build up higher-value services from more basic services:
  - OwnCloud = EOS+FUSE+HTTP or NetApp+HTTP
  - CVMFS = NetApp+HTTP
  - AFS backups = (TSM or CASTOR) + scripts

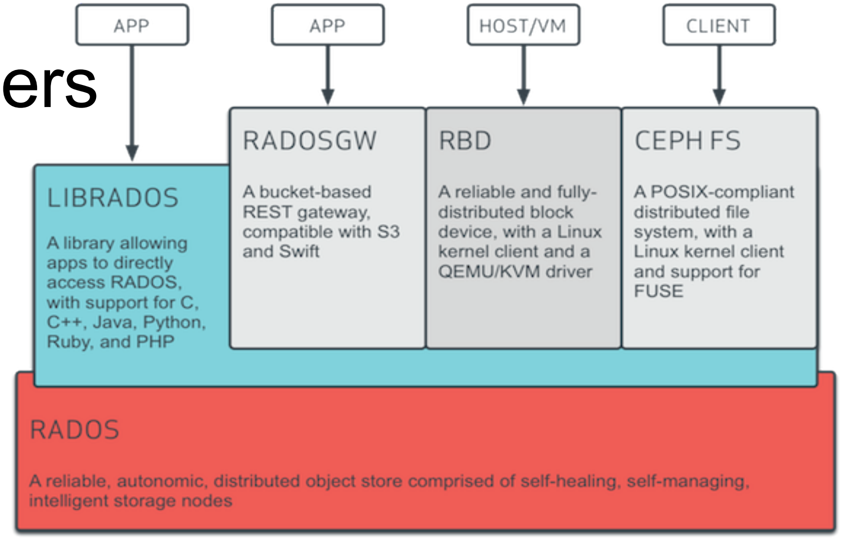
- Redo when needed

All is well?

- Problems / inefficiencies:
  - Manpower issues: training, split
    - But: No formal standby for any data service)
  - Overprovisioning, per-service safety margins
    - But: tend to split too-big services anyway into “instances”
  - “One size fits all” rarely does
    - c.f AFS servers are half-empty – not enough cold data
    - idle CPU capacity on diskservers
    - Spare disks on other machines (CC has 107PB raw disk)
  - Allocation & procurement cycle is sloow
    - Formal tendering, anti-corruption safeguards, national interests ..

Current: Huge interest in CEPH

- Nice 'blocks' -fit at several layers
- Have 3PB 'prototype'
  - Used for OpenStack (VM images+volumes)


APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
RADOS
A reliable, autonomic, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

- Ogled by CASTOR, EOS, (AFS), (NFS)
  - To replace current talk-to-disk-and-handle-errors layer

- (see D. Van der Steer “Ceph – storage for the cloud”, http://indico.cern.ch/event/300076/)

Outlook

- 2nd computer center in Wigner Institute/Budapest
  - “LAN” access now means 23ms..
- Currently LHC is not running (LS1)
  - Restart for “run 2” early 2015 – expect higher data rates (~2x) and different data flows
  - Other experiments will start earlier
  - Run 3: 75GB/s from ALICE?
- RAID works less and less (disk size vs reconstruction speed) : forced to RAIN

Summary

- Some CERN data storage use cases are “special”, some are not
- Toolbox / building block approach
- Common HW reduces cost
  - But manpower / know-how is an issue
- Nothing cast in stone
  - but some upheavals take a long time..