iPRES2013

Cost(s) of Curation

Bit Preservation Experience & Costs for the LHC

Jamie.Shiers@cern.ch

4C Workshop at iPRES 2013

International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

Context

- Context is ***bit preservation*** for the LHC experiments ***as a service (~1FTE@CERN)***
  - **Could be extended to others as part of a Collaborative Data Infrastructure**
- Archive is ***tape***: active is disk (also has curation costs)

- Data on **major migrations** over several decades also available:
  - Platform: e.g. mainframe->clusters(minis)->farms(micros->PCs)->grid->cloud->?
  - Data: 200TB-1PB data format migrations
  - Languages: Fortran+X->c/c++; Build systems; Repositories; Documentation;
  - Major s/w packages: CERNLIB, PAW, GEANT, ROOT, … (many authors, many SLOCs, …)
  - ***Up-coming: re-writes for new architectures***

- Experience from major HEP labs worldwide
- From running experiments + “resurrection(s)”

- **Manpower costs: a factor to an order of magnitude higher? (Per migration)**
- **Hard to envisage “as a service” but “support teams” do help a lot**

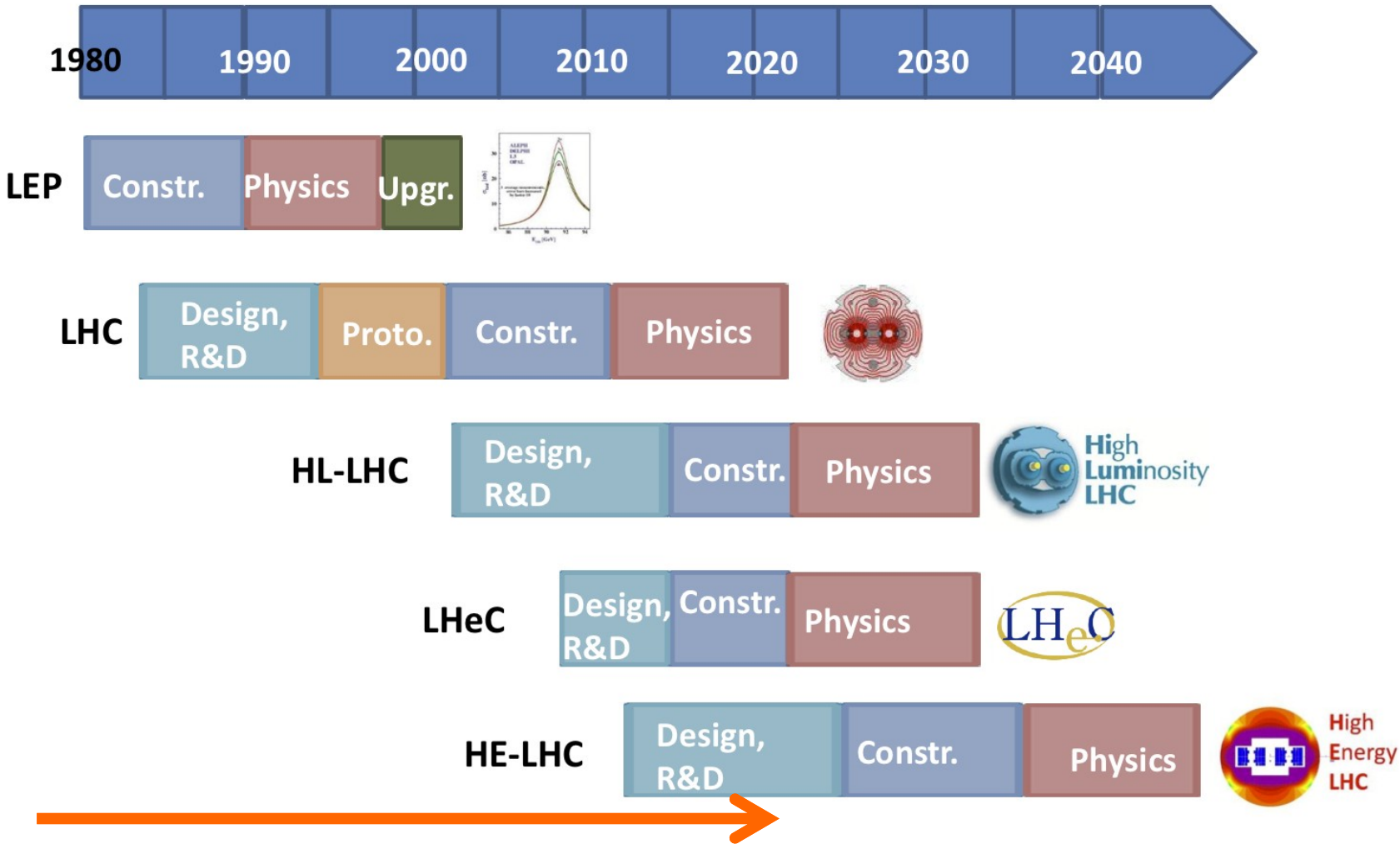
(Exa-)Scale

- Total "physics" data stored at CERN: **100 PB**
  - 70% of this is LHC data ("Run1")
  - 30% is LEP (~1PB) and other pre-LHC experiments
  - ➢ **At least 29PB is lost! "Unlinked"**

- LHC and its successors will run until ~2040
- Total data volume **~1EB**
  - Growth rate: 25PB / year in 2012;
    50PB / year in 2015? >100PB / year in 2020??
  - 2020+ rates: ~1PB / day to archive storage

- Normally, there is <u>at least</u> **one** other copy elsewhere

LHC Timeline



CASTOR archive in Numbers

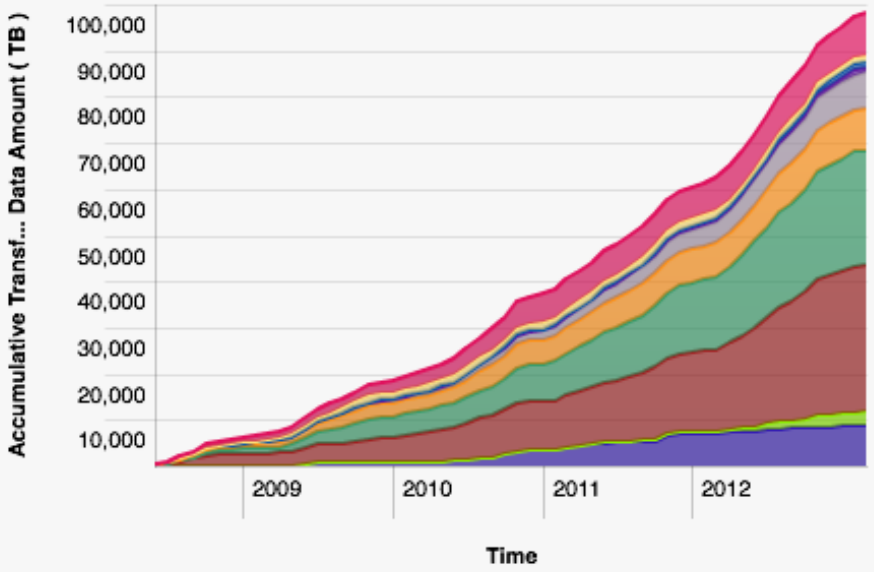
CERN IT Department

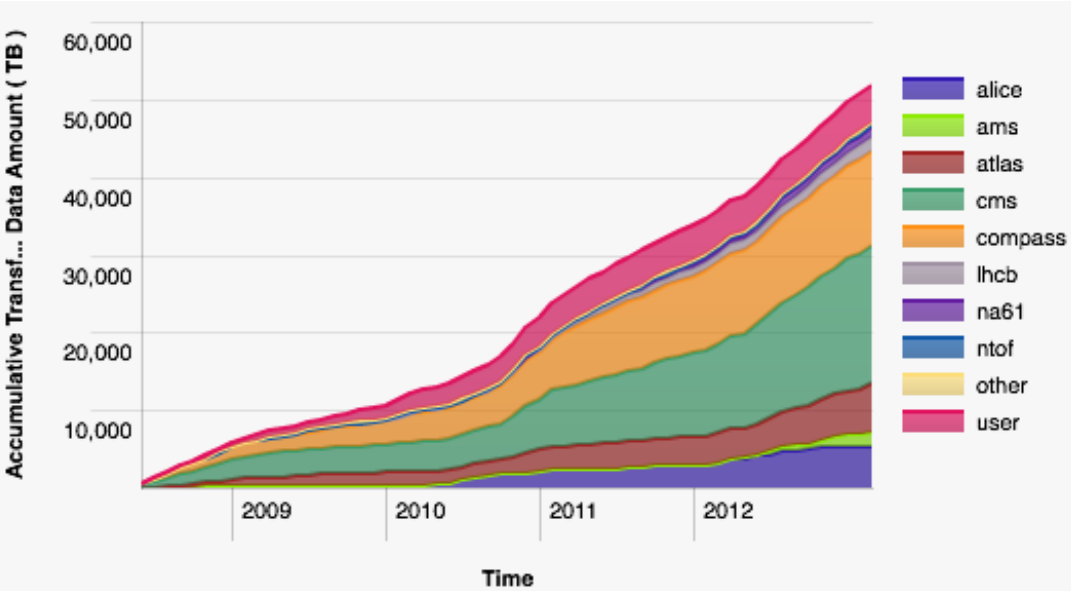
Data:
88PB (74PiB) of data on tape; 245M files over 48K tapes
Average file size ~360MB
1.5 .. 4.6 PB new data per month
Up to 6.9GB/s to tape during HI period

Lifetime of data: infinite

Infrastructure:
~ 52K tapes (1TB, 4TB, 5TB)
7 libraries (IBM and Oracle) – 65K slots
90 production + 20 legacy enterprise drives
15PB disk cache (staging + user access) on ~750 disk servers

CASTOR
CERN Advanced STORage manager





Tape archive verification

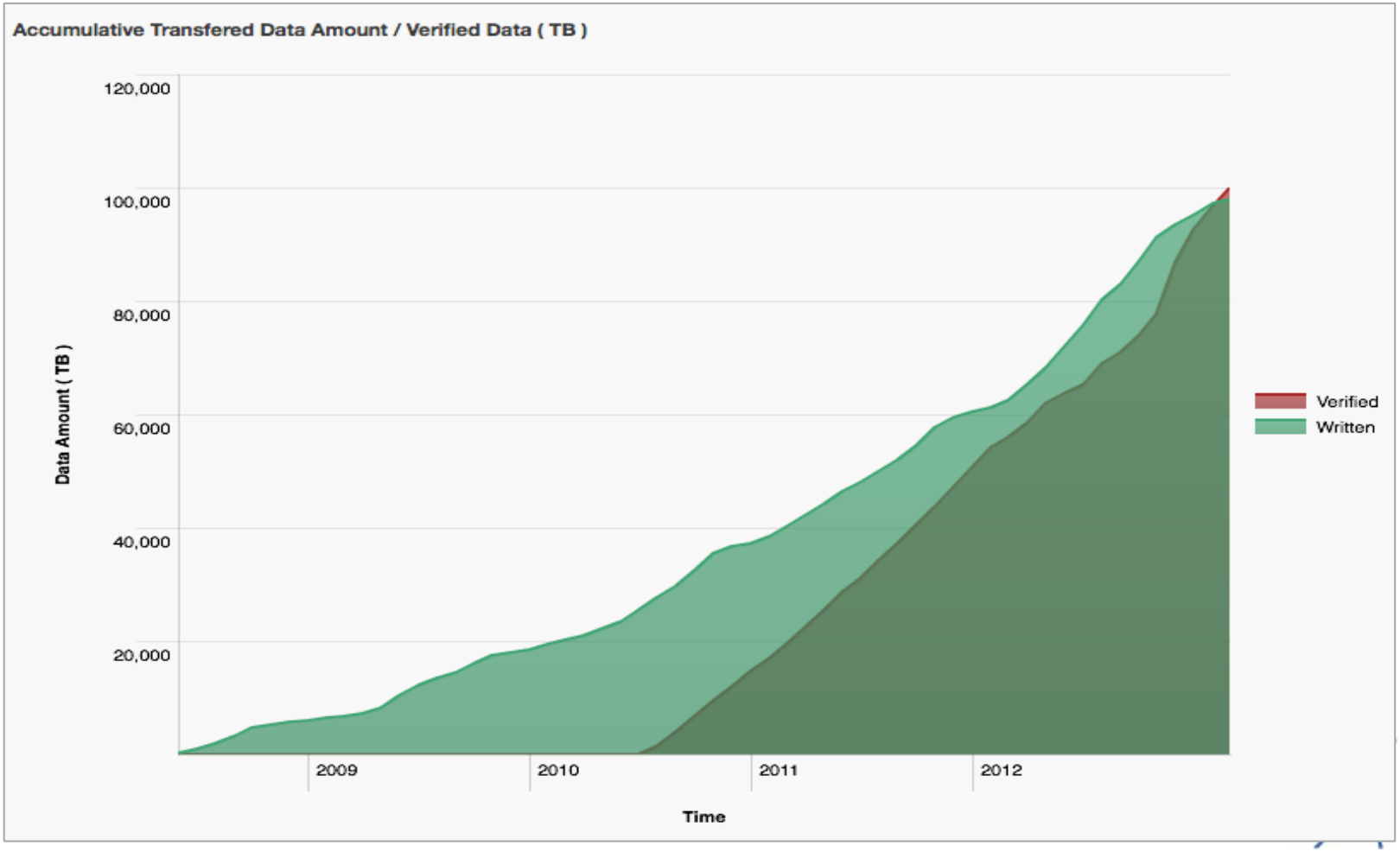
- Data in the archive cannot just be written and forgotten about.
  - Q: can you retrieve my file?
  - A: let me check… err, sorry, we lost it.
- Proactive and regular verification of archive data required
  - Ensure cartridges can be mounted
  - Check data can be read+verified against metadata (checksum/size, …)
  - Do not wait until media migration to detect problems
- Several commercial solutions available on the market
  - Difficult integration with our application
  - Not always check *your* metadata

- In 2010, implemented and deployed a background scanning engine:
  - Read back all newly filled tapes
  - Scan the whole archive over time, starting with least recent accessed tapes

Verification: first round completed!

CERN IT Department

- Up to 10-12 drives (~10%) for verification @ 90% efficiency
- Turnaround time: ~2.6 years @ ~1.26GB/s
- Data loss: ~ 65GB lost over 69 tapes

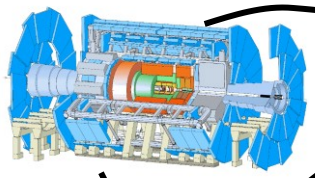


Summary

- Exa-scale bit preservation has costs in hardware + media + manpower
- Manpower: loosely coupled to volume (~1FTE)
- Hardware: ~10% extra investment in drives (LHC)
- Media: costs affected by technology choices & evolution (vendor + consumer)
- Manpower costs do not (need not) dominate
- Plan to share / coordinate also via RDA & HEPiX: input to building Collaborative Data Infrastructures

➢ Caveat: several examples of major data loss during repack exercises – some unrecoverable!

Data Handling and Computation for Physics Analysis



detector

event filter (selection & reconstruction)

reconstruction

raw data

event summary data

processed data

event reprocessing

batch physics analysis

analysis

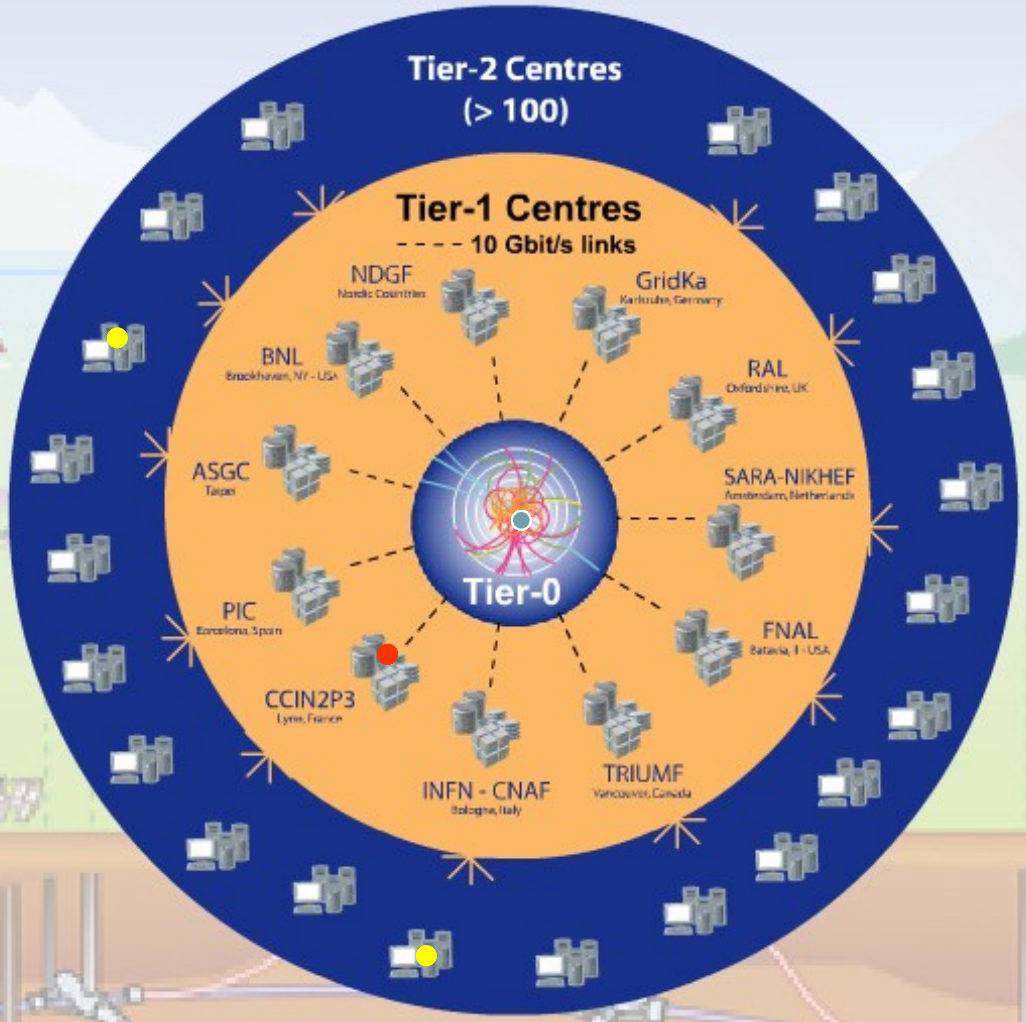
analysis objects (extracted by physics topic)

interactive physics analysis

event simulation

simulation

Tier 0 – Tier 1 – Tier 2



Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (>200 centres):
- Simulation
- End-user analysis

The main 2013-14 LHC consolidations

1695 Openings and final reclosures of the interconnections

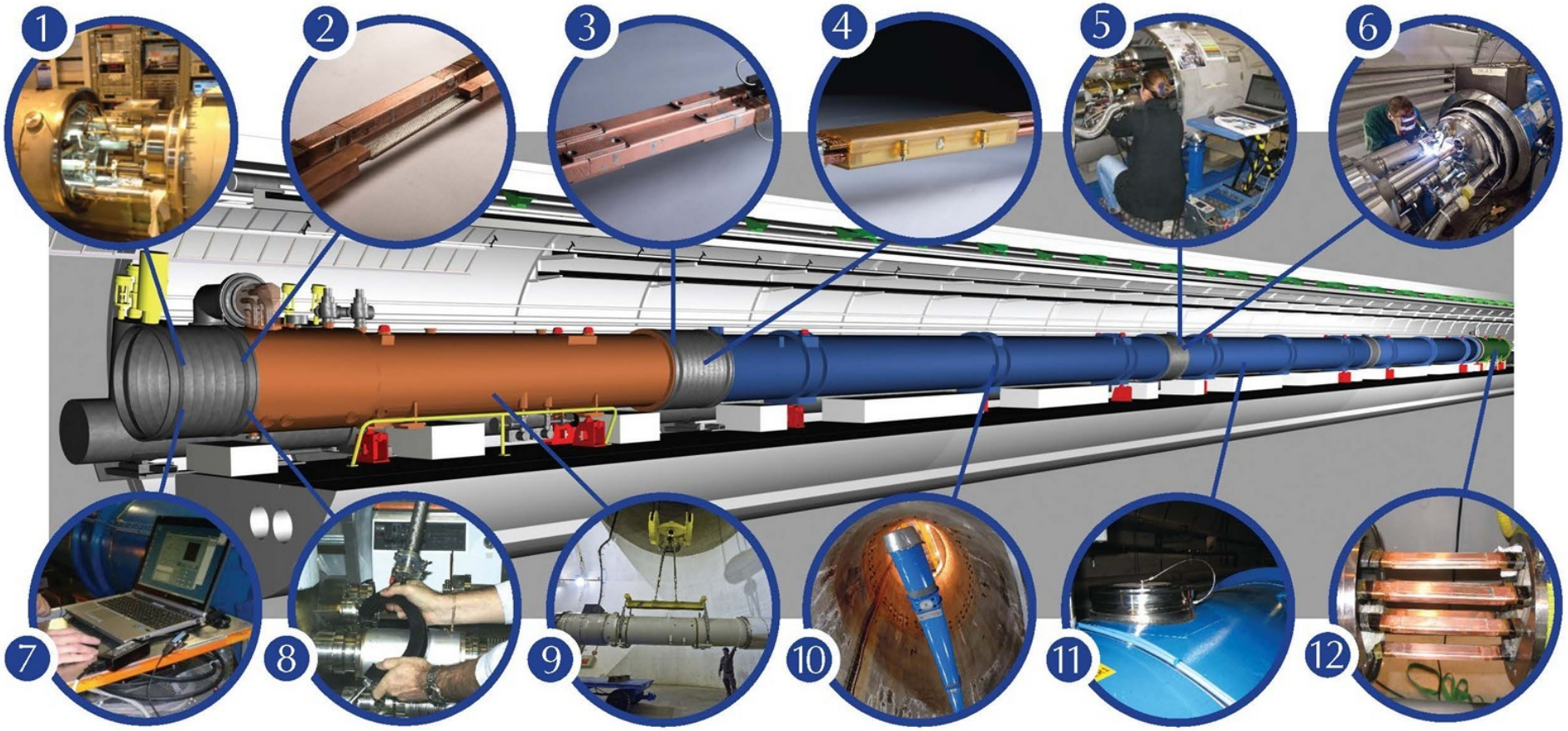
Complete reconstruction of 1500 of these splices

Consolidation of the 10170 13kA splices, installing 27 000 shunts

Installation of 5000 consolidated electrical insulation systems

300 000 electrical resistance measurements

10170 orbital welding of stainless steel lines



18 000 electrical Quality Assurance tests

10170 leak tightness tests

4 quadrupole magnets to be replaced

15 dipole magnets to be replaced

Installation of 612 pressure relief devices to bring the total to 1344

Consolidation of the 13 kA circuits in the 16 main electrical feedboxes

2020 Vision for LT DP in HEP

- ***Long-term – e.g. LC timescales*: *disruptive change***
  - By 2020, all archived data – e.g. that described in Blueprint, including LHC data – easily findable, fully usable by designated communities with clear (Open) access policies and possibilities to annotate further
  - Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards

➢ **Vision achievable, but we are far from this today**

What is HEP data?

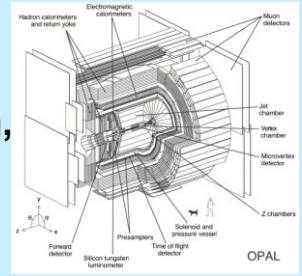
Digital information

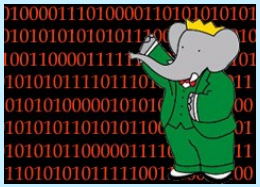
The data themselves, volume estimates for preservation data of the order of a few to 10 PB (+100PB LHC)

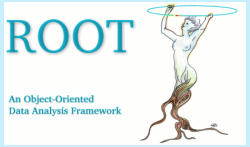
Other digital sources such as databases to also be considered

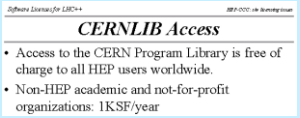
Software

Simulation, reconstruction, analysis, user, in addition to any external dependencies



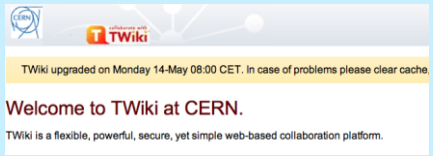






Meta information

Hyper-news, messages, wikis, user forums..



Publications

arXiv.org

INSPIRE

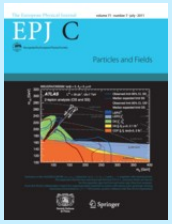
Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,) and is updated at regular intervals.

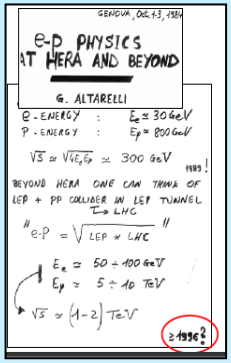
Journal of High Energy Physics

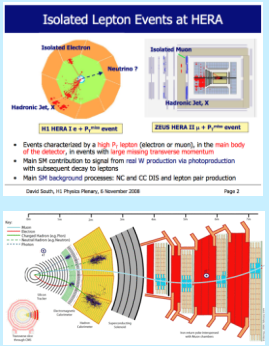
A refereed journal, written, run and distributed by electronic means



Documentation

Internal publications, notes, manuals, slides





Expertise and people

DESY

Documentation projects with INSPIRE

- Internal notes from all HERA experiments now available on INSPIRE
  - Experiments no longer need to provide dedicated hardware for such things
  - Password protected now, simple to make publicly available in the future

- The ingestion of other documents is under discussion, including theses, preliminary results, conference talks and proceedings, paper drafts, ...
  - More experiments working with INSPIRE, including CDF, D0 as well as BaBar

Where are we now?

1. ***Initial*** (chaotic, ad hoc, individual heroics) – the starting point for use of a new or undocumented repeat process.
2. ***Repeatable*** – the process is at least documented sufficiently such that repeating the same steps may be attempted.
3. ***Defined*** – the process is defined/confirmed as a standard business process, and decomposed to levels 0, 1 and 2 (the last being Work Instructions).
4. ***Managed*** – the process is quantitatively managed in accordance with agreed-upon metrics.
5. ***Optimizing*** – process management includes deliberate process optimization/improvement.

Software Strategies

- A 3 pronged approach is being considered:
  - Validation frameworks to (semi-)automate continuous migrations
  - Virtualisation tools to preserve complete environments during LHC lifetimes (decades)
  - Software techniques to help design and implement sustainable software
- Given the (very) long lifetime of the LHC, we will have time + opportunity to evaluate pros & cons
  - e.g. during LS2, LS3 etc.

Data Preservation Maturity Model

Level	Metric	Implications
4	Reproducible results by "citizen scientists"	Desired(?) by funding agencies: people able to reproduce an analysis should be awarded "a degree" – beyond what can realistically be afforded?
3	Reproducible results where consumer ≠ producer and outside immediate community	Stronger demonstration of long-term preservation. Knowledge stored is sufficient for physicist outside immediate community to reproduce results
2	**Reproducible results where consumer ≠ producer but within same "larger community", e.g. LHC (ATLAS / CMS; CDF / D0, …)**	**Highly desirable for "minimal" long-term preservation. "Knowledge" stored is sufficient for a physicist from a different collaboration (but within same overall programme) to reproduce results**
1	Reproducible results where consumer = producer	Required during lifetime of collaboration
0	N/A	Data is lost: logically or physically. This is probably the reality for the bulk of pre-DPHEP experiments (and even some of those??)

- Scale (complexity) is probably "exponential"

Software Preservation Maturity Model

Level	Metric	Implications
4	Reproducible results by "citizen scientists"	Desired(?) by funding agencies: people able to reproduce an analysis should be awarded [illegible] – beyond what can realistically be af[illegible]
3	Reproducible results where consumer ≠ producer and outside immediate community	Stronger demonstration of lo[illegible] preservation. Knowledge stored is suf[illegible] physicist outside immediate commu[illegible] reproduce results
2	**Reproducible results where consumer ≠ producer but within same "larger community", e.g. LHC (ATLAS / CMS; CDF / D0, …)**	**Highly desi[illegible] minimal" long-term prese[illegible] knowledge" stored is sufficient for a p[illegible] om a different collaboration (but within [illegible] overall programme) to reproduce results**
1	Reproducible results wh[illegible] consumer = produc[illegible]	Required during lifetime of collaboration
0	N/A	Data is lost: logically or physically. This is probably the reality for the bulk of pre-DPHEP experiments (and even some of those??)

REPRODUCIBLE RESULTS AFTER "PORTING" TO NEW ENVIRONMENT!

Outlook: Tape market evolution

- Tape technology getting a push forward
  - Drive generations last released

Vendor	Name	Capacity	Speed	Type	Date
LTO consortium(*)	LTO-6	2.5TB	160MB/s	Commodity	12/2012
Oracle	T10000C	5.5TB	240MB/s	Enterprise	03/2011
IBM	TS1140	4TB	240MB/s	Enterprise	06/2011

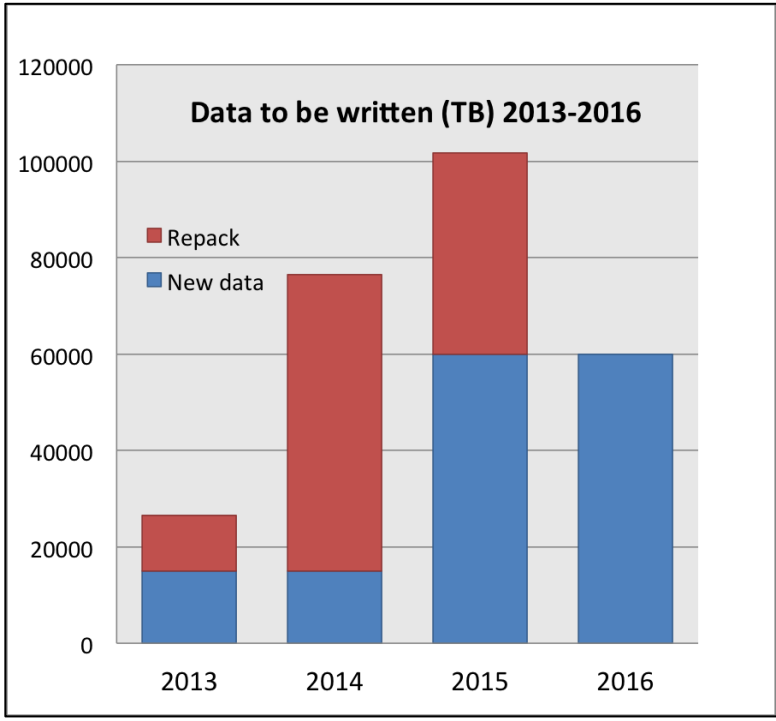
  - Vendor roadmaps exist for additional 2-3 generations, up to 20TB / tape (~2016-17) (+70% capacity / year) – new generations expected 2013/14
  - 35/50TB tape demonstrations in 2010 (IBM/Fuji/Maxell); 125-200TB tapes being investigated by IBM

- Tape market evolving from NEARLINE to ARCHIVING
  - Increased per-tape capacity and transfer speed
  - Little or no increases for mounting/positioning – unsuitable for random access
  - Small-to-medium backup market shrinking (de-duplication, disk-only)
  - Large-scale archive/backup market building up (legal, media, cloud providers - Google: ~6-10EB?)

(*) LTO consortium: HP/IBM/Quantum/Tandberg (drives); Fuji/Imation/Maxell/Sony (media)

Outlook: Media repacking

- Mass media migration or “repacking” required for
  - Higher-density media generations, and / or
  - Higher-density tape drives (enterprise media rewriting)
  - Liberating tape library slots
- Media itself can last for 30 years, but not the infrastructure!
- Repack exercise is **proportional** to the **total size of archive** - and **not** to the fresh or active data
- Next Repack run (expected): 2013/4 - 2016
  - New drive generations appearing “soon”
  - ~100PB to migrate from over 50’000 cartridges
- Data rates for next repack will exceed LHC data rates…
  - Over 3 GB/s sustained
  - Cf . LHC proton-proton tape data rates : ~1-1.5GB/s
- …. but we need to share the drives – **which become the bottleneck**
- Will compete with up to 60PB/year data taking after LS1



- Infrastructure, software and operations must sustain writing up to 0.1EB in 2015 (+ reading!)