Data Preservation for the HERA Experiments at DESY

Dirk Krücker

DESY

DPHEP Workshop, CERN

9.6.2015

HELMHOLTZ
| ASSOCIATION

DESY

HERA

- HERA was the largest particle accelerator at DESY

- It was the first internationally funded accelerator project and the joined effort of 11 countries

- Started in 1992,the storage ring served the international particle physics community for over 15 years

- The HERA experiments H1, ZEUS and HERMES finished data taking in 2007

- Up to now – and for the foreseeable future – no other electron-proton accelerator has explored electron-proton interaction at higher energies
  → Unique dataset
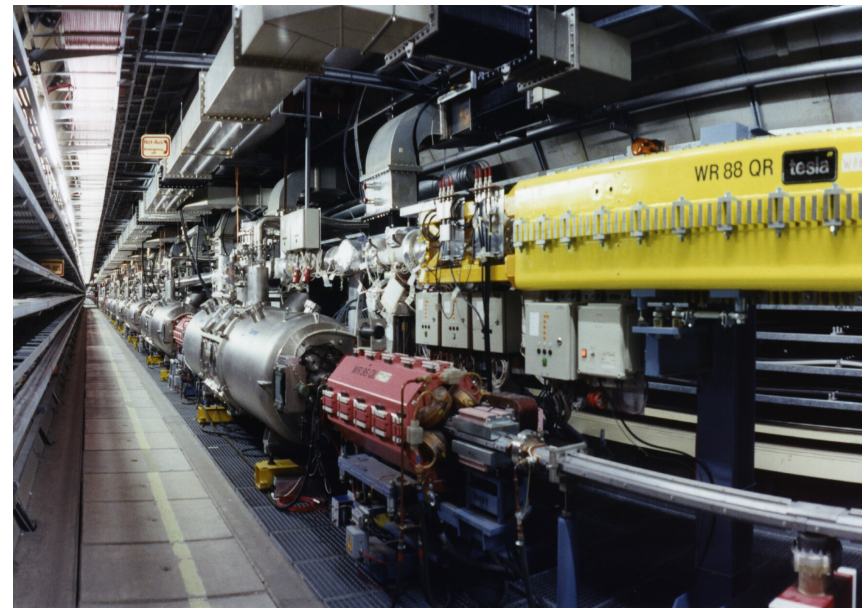
Aspects of Data Preservation
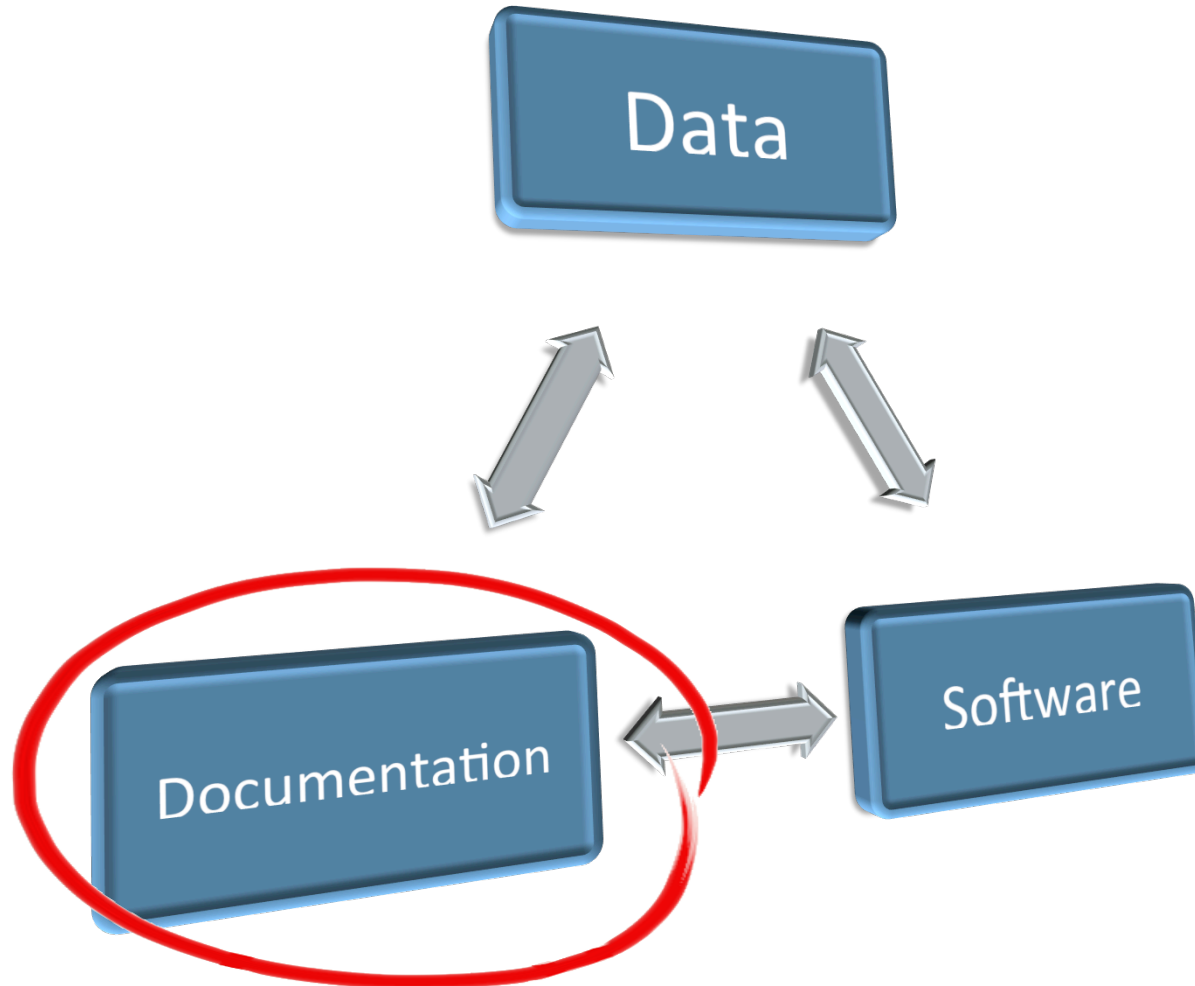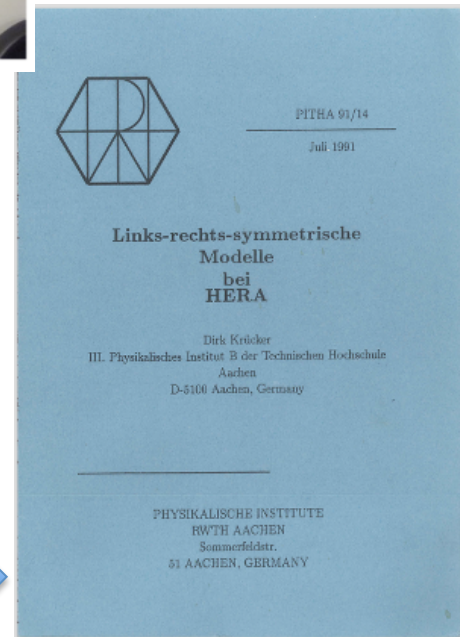
Documentation

before

after

DESY Library

During the past years great care has been taken to collect all available digital and non-digital documentation accumulated by the experiments over years

- Non-digital documentation have been catalogued and stored in the DESY library archive
- Some part of non-digital documentation has been digitized
- Web servers

My first encounter with the HERA data preservation effort

Documentation

Web servers are non-trivial

- In the past different, experiment specific solution had been built up over years
  - Dead links
  - Scripts
  - Dependence on storage
    - /afs/desy.de group directories for the experiments
    - As long as AFS exits at DESY the experiment group names will be kept but some day AFS will be gone
  - Still in use for administrative task: Information on remaining analyses, authors lists

For preservation we provide

- Web server with static version of the experiment webpage (plain html) on virtual machines with isolated storage maintained by the DESY 'web office'
  - conflicting views
    - Production system vs. long-term preservation
  - time consuming transition - complexity easily underestimated
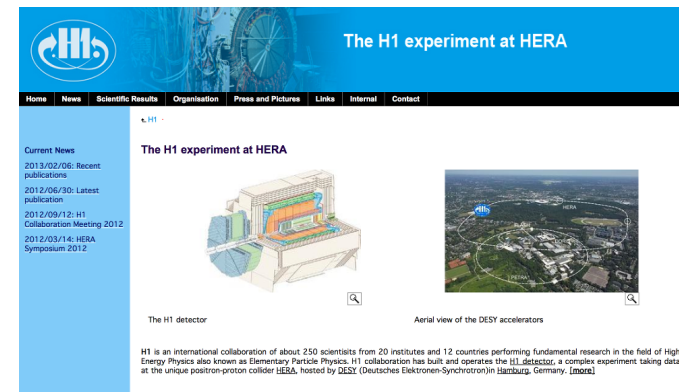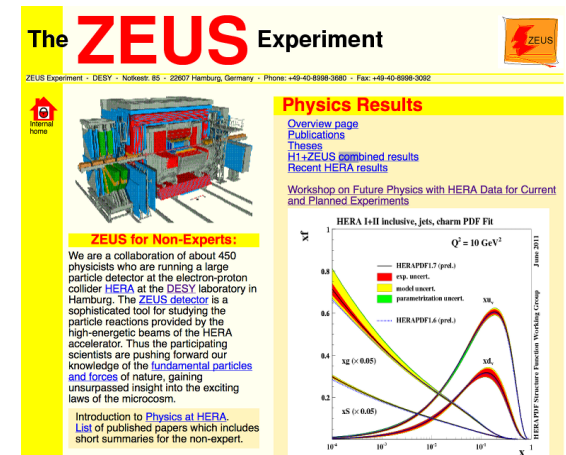
Present Situation DESY DPHEP

- Had been a group of H1, ZEUS and HERMES people together with DESY-IT and DESY Library
  - Steadily decreasing number of persons
- Funding ended as expected
  - **H1 and ZEUS 31ˢᵗ Dec 2014**
  - HERMES 31ˢᵗ Dec 2012
  - By now **no dedicated manpower for DP** left for the experiments
  - 1 **FTE @ DESY IT** (me) until May 2016
- But there are still ongoing analyses and there is interest in the data beyond that
  - Workshop last year: **Future Physics with HERA Data for Current and Planned Experiments**
    https://indico.desy.de/conferenceDisplay.py?confId=10523

  No dedicated person-power for DP from the experiments from now on
&& Ongoing analyses work
=> **How to serve the needs best?**

Strategy

- Transition from experiment specific to **institutional solution**
  - Storage, computing, web services and documentation
- DESY IT department and Library
  - If the data, the knowhow documented in webpages etc. should survive on the long run the solutions must be provided by the institution DESY and not the few remaining people from the experiments or single individuals
- **Short and long-term availability** - 2fold strategy
  - Assure the long-term availability of the data -> **tape archive**
  - Support on-going analyses work/unfinished PhD theses
  - Keep the data easily accessible -> **disk pools**
- Our main effort during the last months was related to bit-preservation

Aspects of Data Preservation
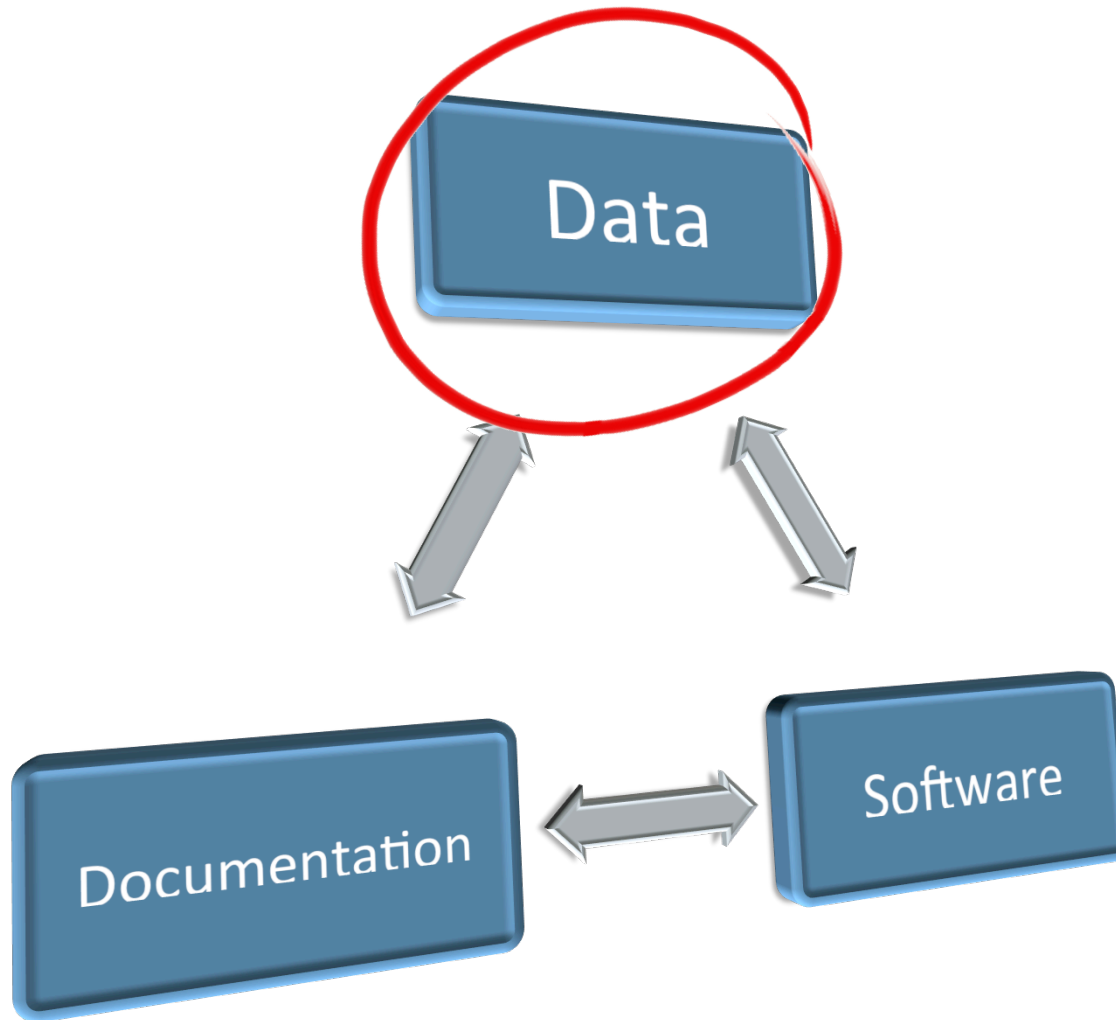
Transition Old -> New Store

- At DESY different dCache instances for mass storage are maintained
  - dCache is an abstraction layer between hardware and filesystem
    - Hardware: disks, tape robot
    - Access: different protocols; from dcap,xrood,NFSv4.1 to WebDAV
- All HERA data had been stored on the **HERA dCache**
  - HERA dCache is more than 13 years old and cannot be reasonably maintained
  - HERA dCache had been set **read-only** end of 2014
  - Planned **shutdown** in 2015
- Good opportunity to clean up and define the HERA legacy dataset!
  - Mainly done but the last 10% is difficult - as always …
  - Data, MC, log files etc.
- Transfer data to new store:
  dphep area on **DESY dCache**

CMS dCache   Atlas dCache   HERA dCache   →   DESY dCache
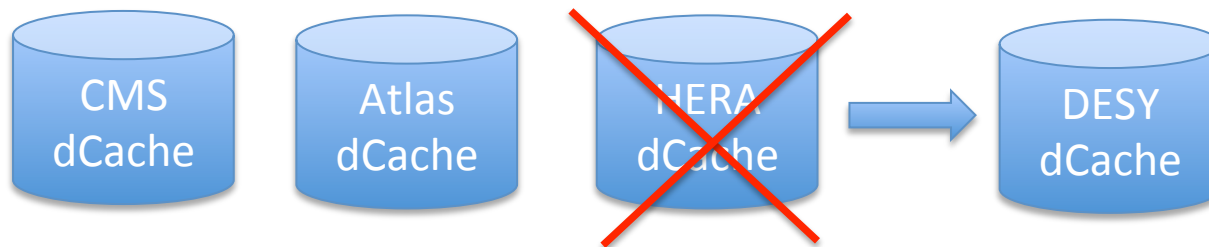
Storage Structure
DPHEP on DESY dCache
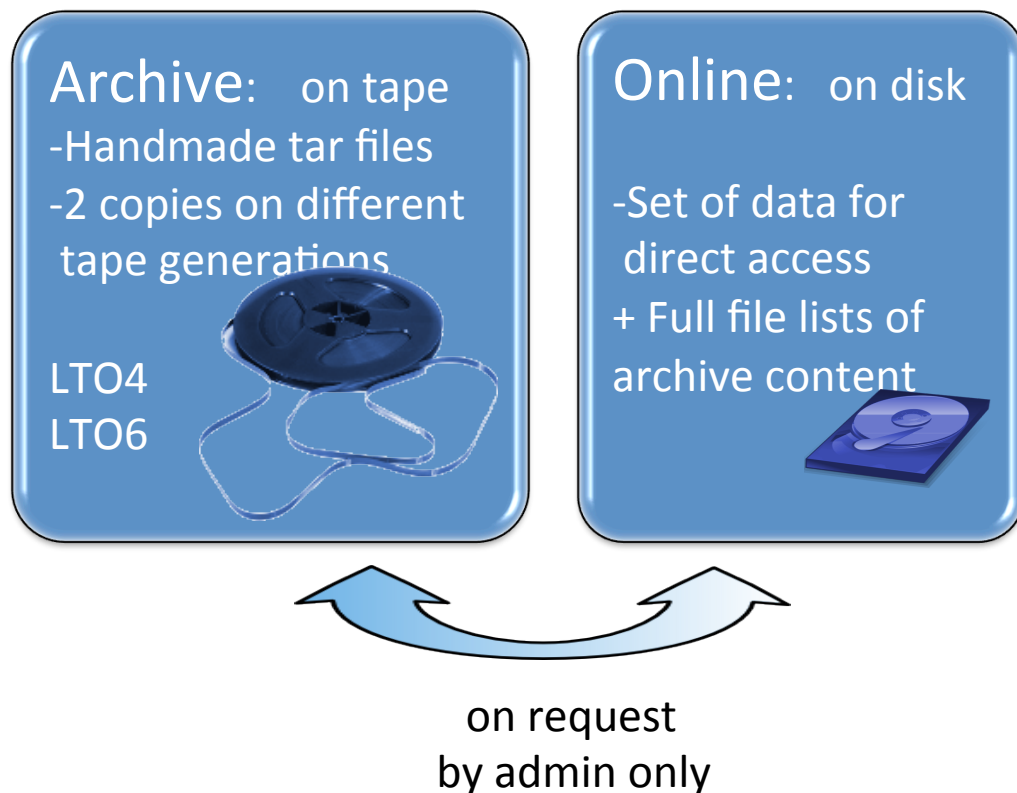
A twofold system

Archive: on tape
- Handmade tar files
- 2 copies on different tape generations

LTO4
LTO6

Online: on disk

- Set of data for direct access
+ Full file lists of archive content

on request
by admin only

- **Archive part 1.2 PiB**
  - All data identified for preservation will be available in 2 redundant tape copies on 2 different generations of tape cartridges (LTO4/LTO6)
  - The archive is not generally accessible by the user
  - Handmade tar files
- **Online part 700 TiB**
  - 47 disk pools
  - Can be mounted r/o where needed i.e. batch cluster or WGS

Storage Structure
DPHEP on DESY dCache

A twofold system

Oracle SL8500
tape robot

Online: on disk

-Set of data for
 direct access
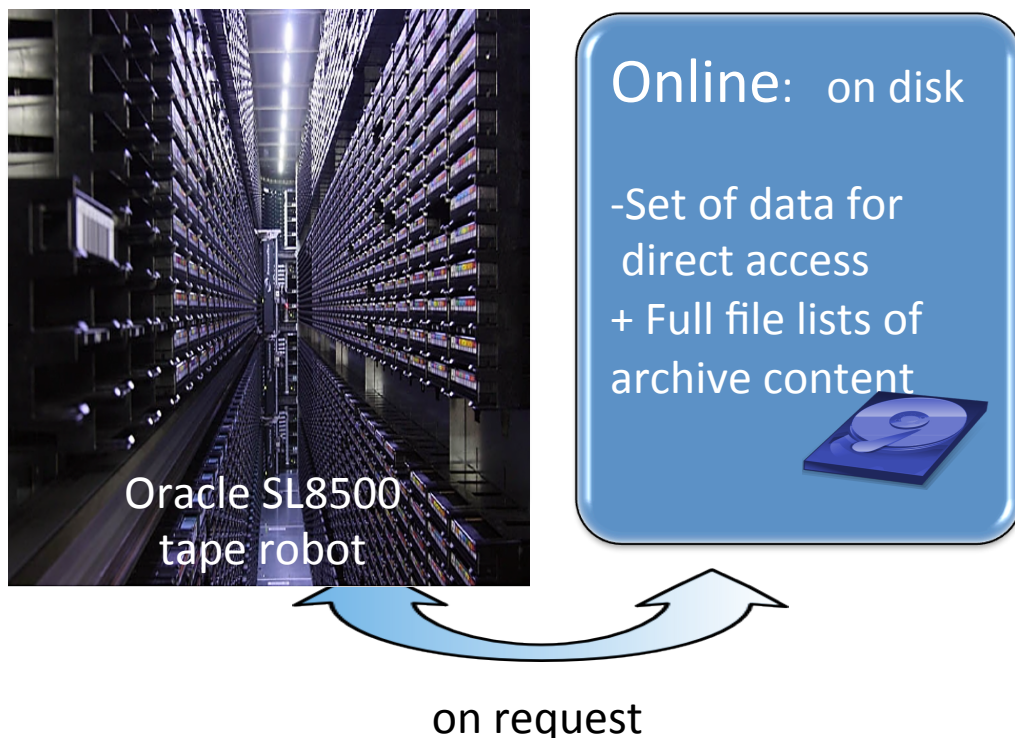+ Full file lists of
archive content

on request

- **Archive part 1.2 PiB**
  - All data identified for preservation will be available in 2 redundant tape copies on 2 different generations of tape cartridges (LTO4/LTO6)
  - The archive is not generally accessible by the user
  - Handmade tar files
- **Online part 700 TiB**
  - Disk pools
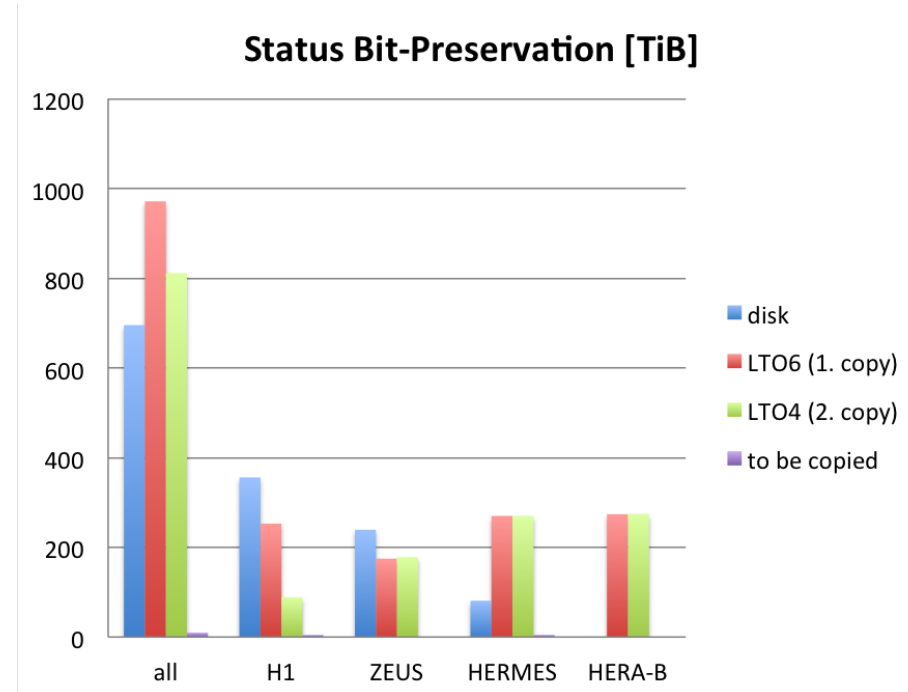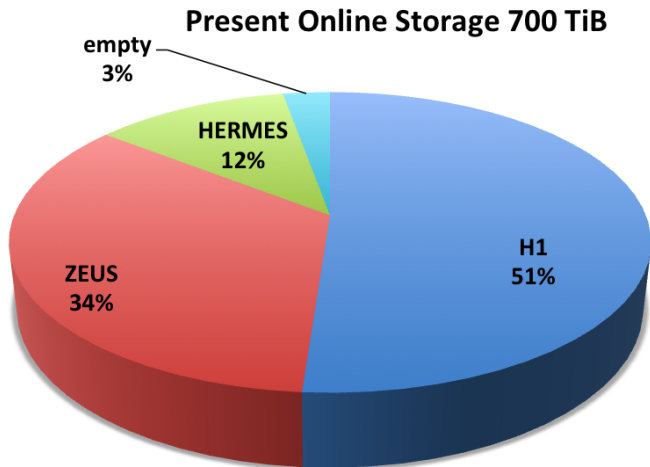 - Can be mounted r/o where needed i.e. batch cluster or WGS

HERA Bit-Preservation

Present Online Storage 700 TiB

empty
3%

HERMES
12%

ZEUS
34%

H1
51%

Status Bit-Preservation [TiB]

- disk
- LTO6 (1. copy)
- LTO4 (2. copy)
- to be copied

Recent activities

- Old HERA dCache had been set **read only**
  – a fixed dataset now
- Online store had been extended to **700 TiB**
  – sufficient for the remaining data
  – scratch area 24TiB for ongoing work provided
- ZEUS dataset **finalized** and copied to new store
- HERMES is reorganizing the online store
- H1 data largely defined but larger datasets for online than expected

For the Statistics Enthusiasts: Present Storage Content

H1	Hermes	Zeus	HeraB	type
798872	6458415	974552	846059	files
4545	7269	5531	4109	tar file container
106	403	268	392	LTO4 (800G) tapes
80	114	74	110	LTO6 (2.4T) tapes
356	81	239	0	TBytes online
88	270	178	276	TBytes on LTO4 tape
253	270	174	276	TBytes on LTO6 tape

About 9 million files, many small files
(2x) 1.2 PiB on tape
~2 years work – sorting out – copying old tapes etc.

Creating new tapes and checking the data integrity

Creating new tapes and checking the data integrity

old tapes

PostgresSQL DB

new DESY-SE dCache

Pool

single file 1..n

dcap

scratch area

copy job

tar

NFS 4.1

tar file - file i

Pool

old HERA dCache

PostgresSQL DB

checksum calculated from within the

new tapes

Looks complicated

checksum as calculated when file entered the store in the past

control job
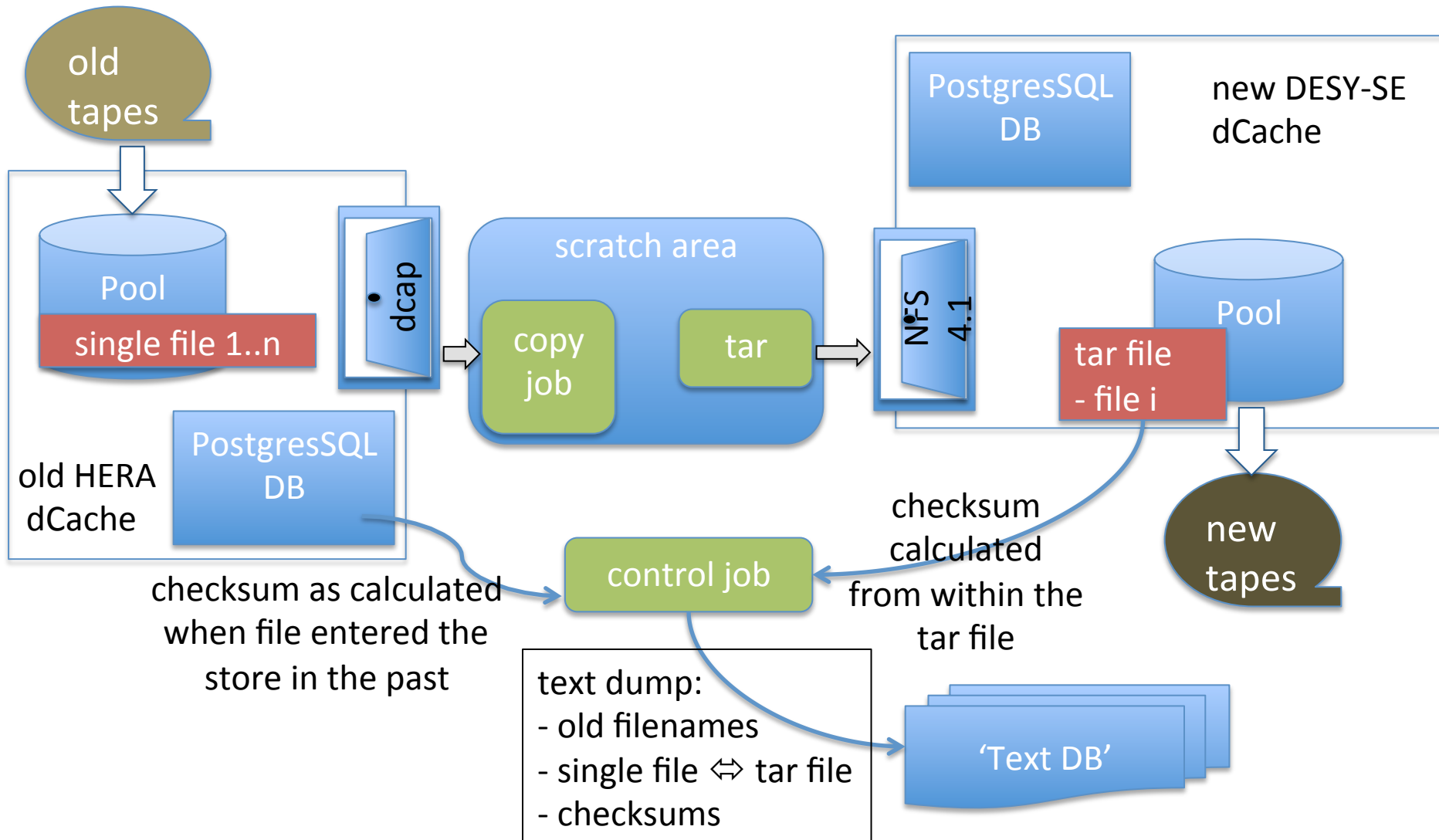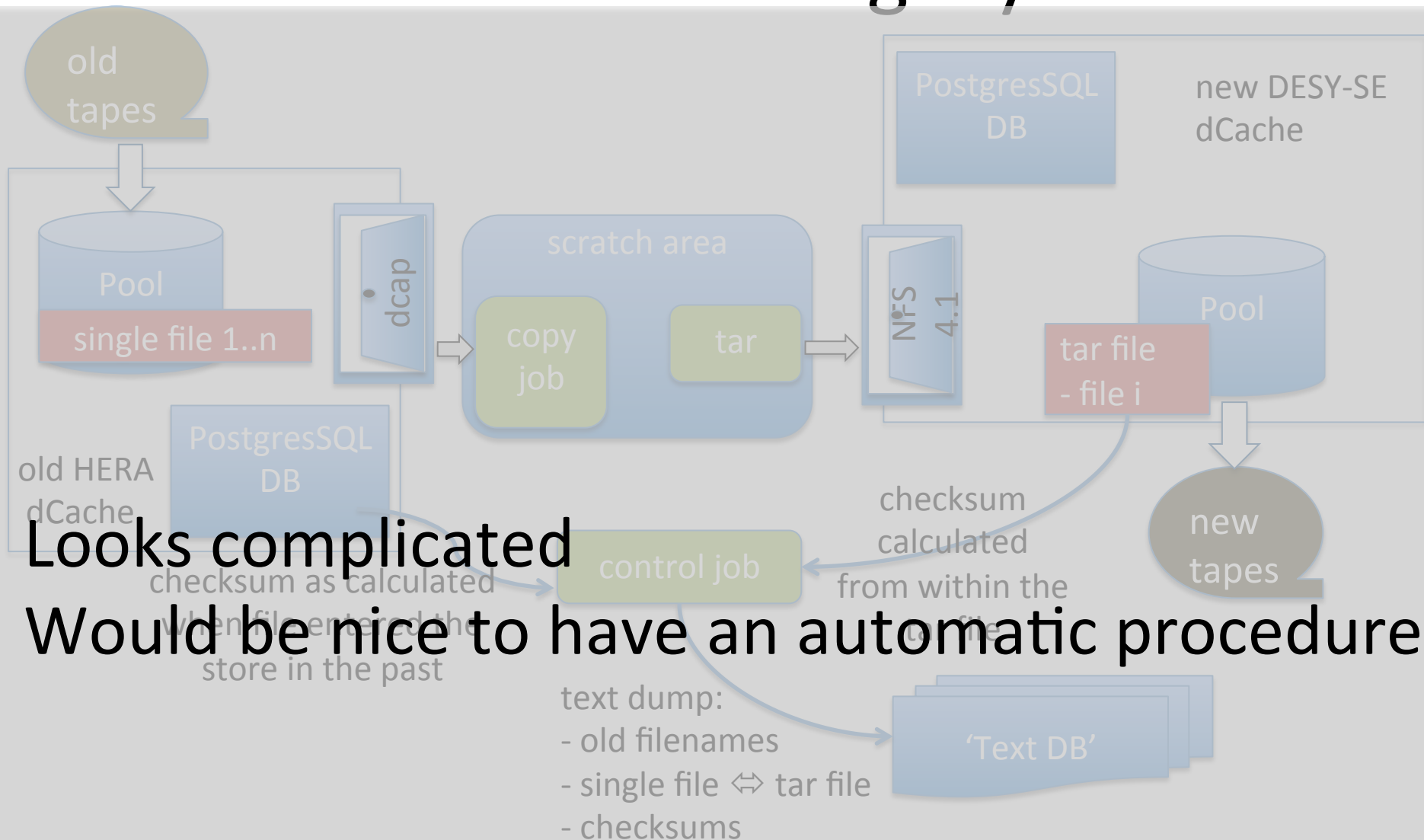
Would be nice to have an automatic procedure

text dump:
- old filenames
- single file ⇔ tar file
- checksums

'Text DB'

DESY dCache Small Files Service

Small files are not tape friendly

Archive: on tape
-Handmade tar files
-2 copies on different
 tape generations

LTO4
LTO6

Online: on disk

-Set of data for
 direct access
+ Full file lists of
archive content

on request
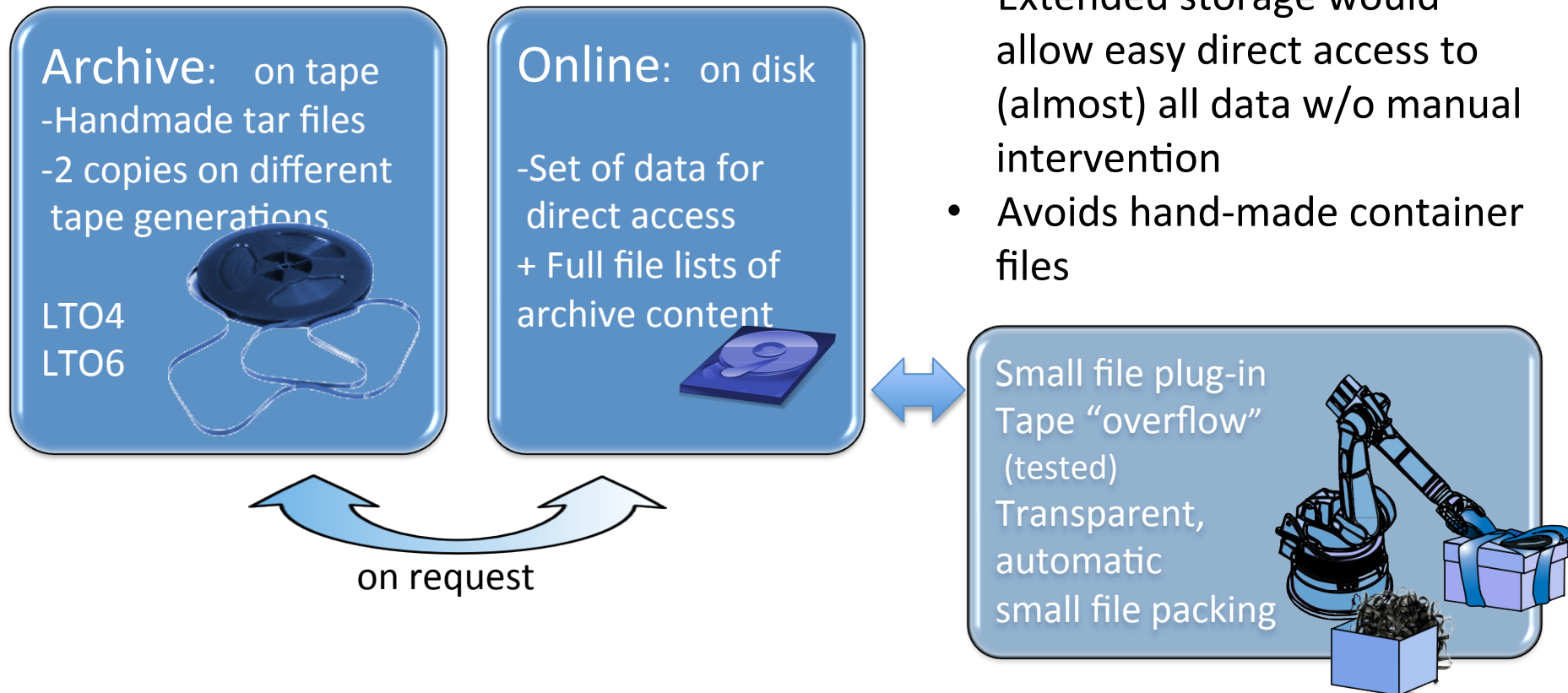
- **Automated tar file**
- Extended storage would allow easy direct access to (almost) all data w/o manual intervention
- Avoids hand-made container files

Small file plug-in
Tape "overflow"
 (tested)
Transparent,
automatic
small file packing

DESY dCache Small Files Service

Small file plug-in
Tape "overflow"
with
transparent,
automatic
small file packing
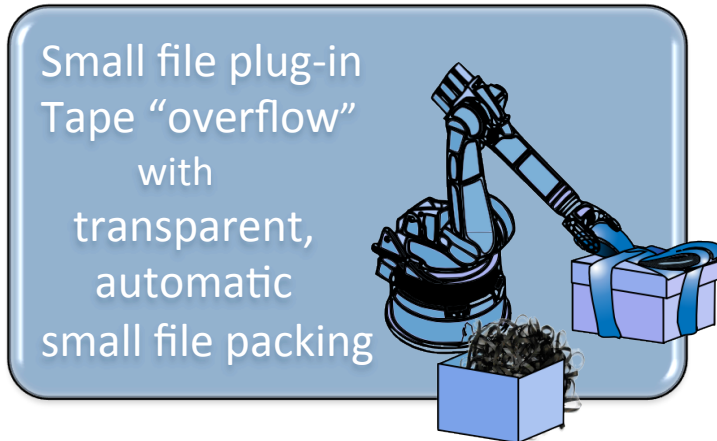
- Files are transparently packed and unpacked from container files using dCache's tape interface
  - User only sees the directory
- A first step to automatize such procedures
- Can be attached to any recent dCache version
- All file metadata is stored in dCache
  - No additional text files etc. necessary information in dCache DB
- We will not change the approach in the middle of the project.
  CHEP15 talk/paper:
  'Data preservation for the HERA experiments at DESY using dCache technology '

Aspects of Data Preservation
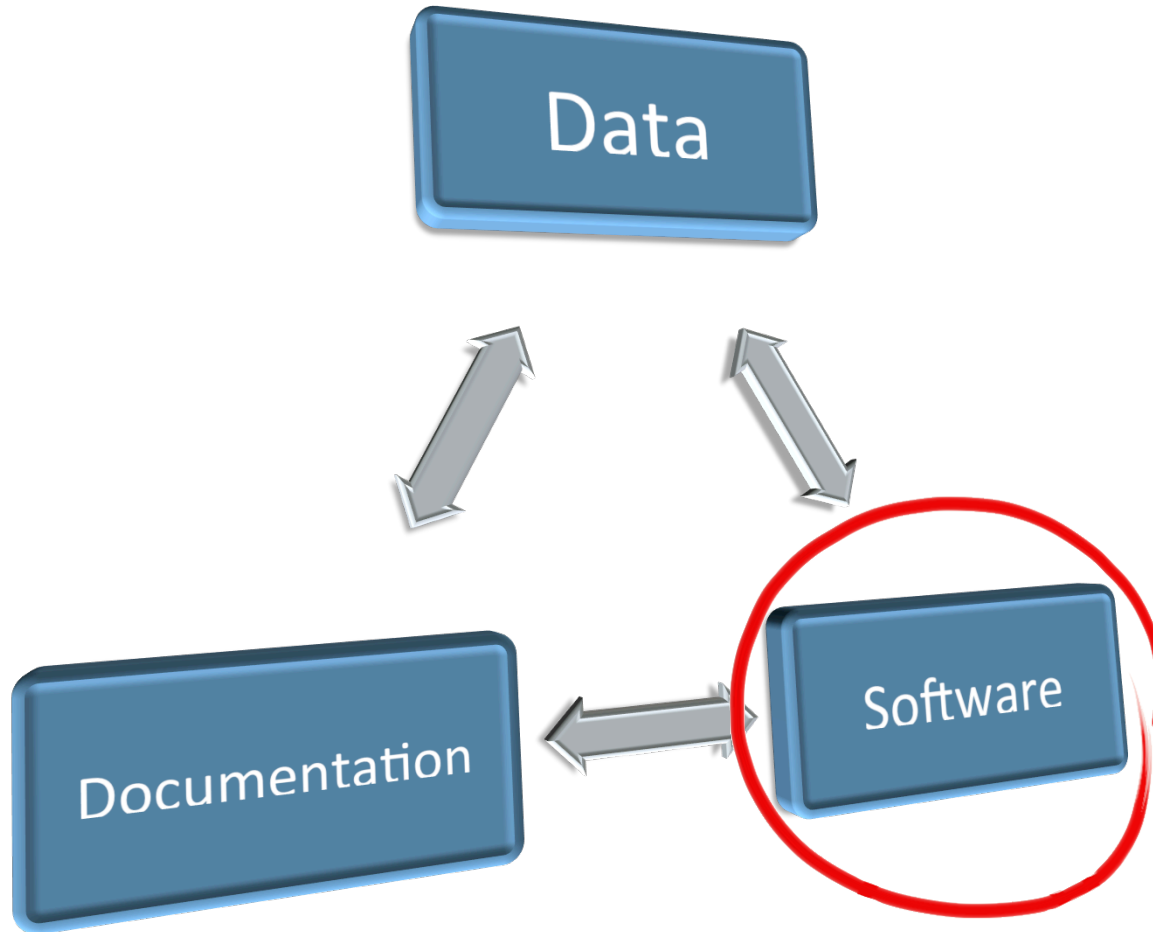
Software preservation

Experiment data is meaningless without a software to read it

- For this historic period the experiments typically started with
    - Fortran codes with custom made data formats,
    - did analyses in PAW and,
    - moved at some point to C++ and ROOT
    - > The preservation data consists of presently used (ROOT5) and historic formats (PAW, ZEBRA, BOS)
- Dependencies on several external libraries/ environments
- There are different philosophies to tackle the problem

Philosophies

Raw data approach

- Keep
  - Raw data
  - Software for processing
  - Calibration, alignment etc. databases
  - Reconstruction software
- Keep the software alive
  - The abillity to analyze the data from scratch
- As a by-product keep the ability to do MC production i.e. Detector simulation

Processed Data approach

- Keep
  - the latest state-of-the-art (calibration, alignment etc.) processed data in a recent data format
- Analyse high level data only
- Loose the ability of MC production

Philosophies

Raw data approach

- Keep
  - Raw data
  - Software for processing
  - Calibration, alignment etc. databases
  - Reconstruction software
- Keep the software alive
  - The abillity to analyze the data from scratch
- As a by-product keep the ability to do MC production i.e. Detector simulation

Level 4:
Preserve the reconstruction and simulation software as well as the basic level data

Processed Data approach

- Keep
  - the latest state-of-the-art (calibration, alignment etc.) processed data in a recent data format
- Analyse high level data only
- Loose the ability of MC production

Level 3:
Preserve the analysis level software and data format

Philosophies

Raw Data approach

- Keep
  - Raw data
  - Software for processing
  - Calibration, alignment etc. databases
  - Reconstruction software

- Keep the software alive
  - The ability to analyze the data from scratch

  As a by-product keep the ability to do MC production i.e. Detector simulation

Processed Data approach

- Keep
  - the latest state-of-the-art (calibration, alignment etc.) processed data in a recent data format

- Analyse high level data only

- Loose the ability of MC production

H1

ZEUS

Reality is mixed

- Both experiments want to save a large set of recent MC production in there latest file formats

- ZEUS e.g. in addition to ROOT ntuple saves PAW ntuple

- Strong psychological barrier to decide that certain data shall be thrown away for ever

- In the best of all worlds we would keep the software alive i.e. compilable on the latest linux with the latest library versions
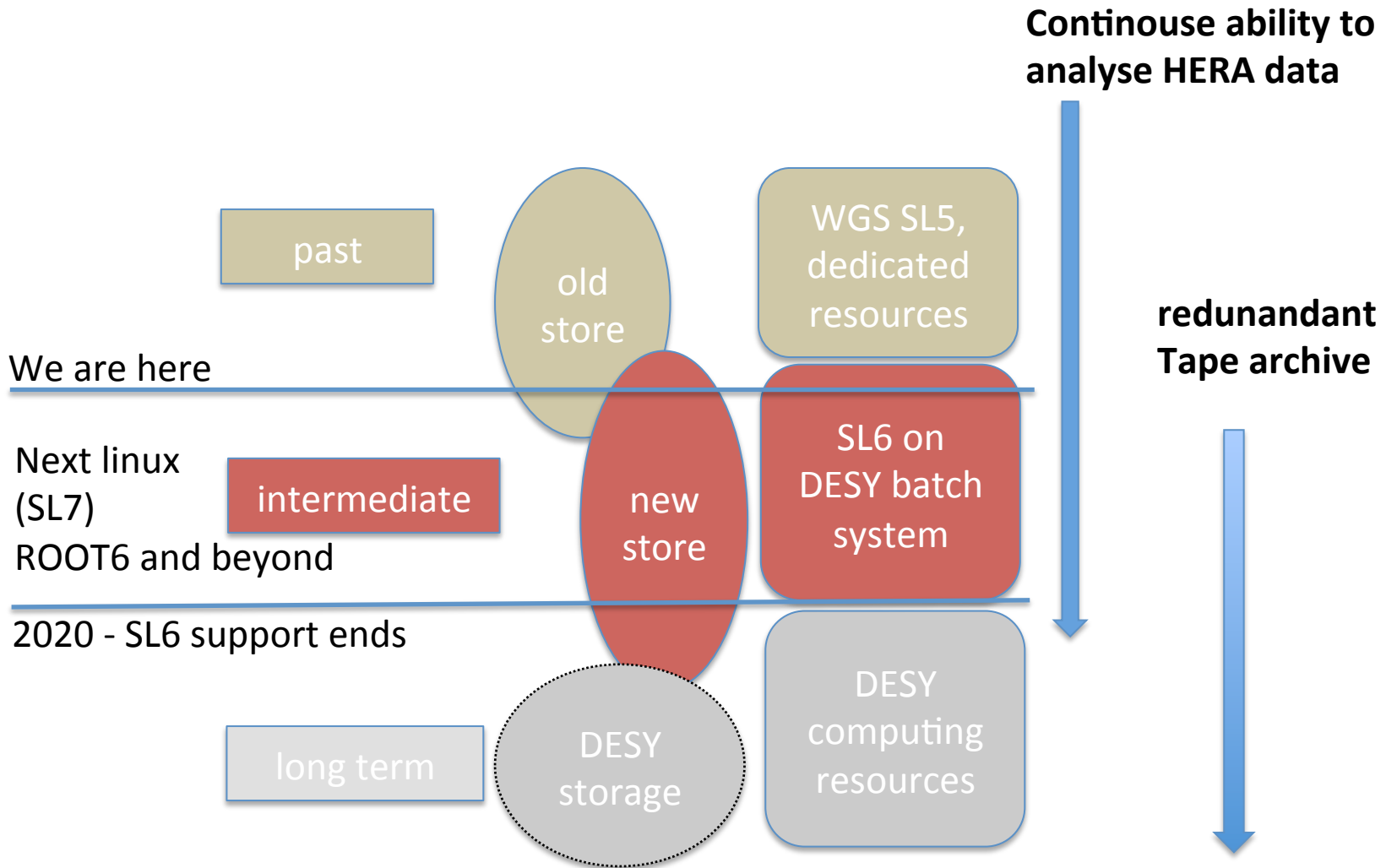
SP System
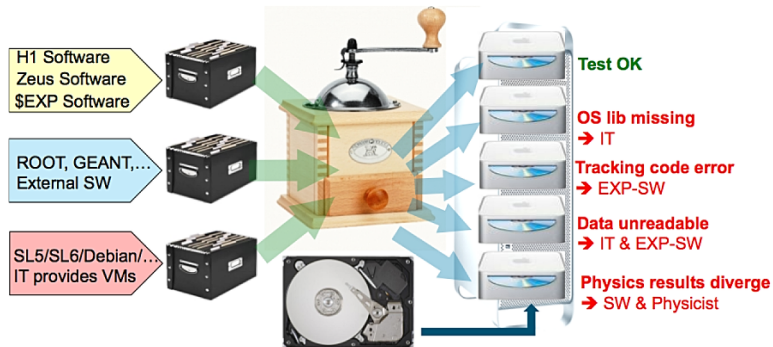
- HERA experiments tried to migrate software for as long as possible instead of just freezing it
- The idea of the sp-system is to help perform migrations to newer software versions and environments, where transitions are performed often and validated by a comprehensive set of tests provided by the experts
- There is always a running version of the software with the latest versions of system libraries and other external dependencies
- Recipe for production environment

- Pilot project in **2010**
  - Concept similar to *Continuous Integration*
  - But test driven development/maintenance had been an relatively new concept to the generation of experiment software
  - $\Rightarrow$ large amount of work to define exhaustive tests
  - **Full design of SP system not implemented**
    - Many dependencies in exp. soft. that may distort the physical output
    - A test that the software compiles is easy but are the (slightly different) results physically sound?
    - By design not a fully automated approach
    - Lack of expert person-power (H1/ZEUS software experts are not around anymore)

Replicability is easy to test but checking consistency of physical results is subtle

Software Preservation Status

- The SP system was involved in transition to SL6 and we gained experience with validation environments
  - All experiments now running on **SL6**/NFS4.1 and central DESY IT resources
  - Analyses of HERA data secured until **2020**
  - Experiments use DESY resources e.g. DESY batch sysytem on a **fair share** basis
- For the software preservation we now follow a freezing approach
  - A virtual machine with isolated storage and a well defined set of external libraries
- Transistion beyond 2020 is an open question.
  - May become an urgent question for H1
  - ZEUS uses flat ROOT nTuples
- There are possible new approach meanwhile
  (SP project started 2010)
  - Computing clouds could provide an environment to run a frozen software version in production mode

As a Summary

Continouse ability to analyse HERA data

redunandant Tape archive

WGS SL5, dedicated resources

past

old store

We are here

Next linux (SL7)
ROOT6 and beyond

intermediate

new store

SL6 on DESY batch system

2020 - SL6 support ends

long term

DESY storage

DESY computing resources

Some Observations

Do not underestimate the effort

- Large amount of data (MC) is produced in the final phase of the experiments
  - Size of the data is measure in 'now-a-day' units i.e. what can be handled at the end of the experiments
- Weeding out the data and scrutinize its usefulness is a difficult task
  - Large understandable, psychological barriers to give up data
  - Starts typically late - to include latest processing, latest MC production…
- Data preservation must be prepared while the collaborations exits and person-power is available
  - The HERA experiments did a good job here and started early
- Experiment expertise is fading away quickly after end of funding
  - There is a danger that the required person-power is underestimated
- Some ambitious projects will not come to an end
  - There will always be an concurrency between different task in the final phase. Fallback solutions are always a good idea

END