DPHEP

Data Preservation in High Energy Physics

Science Demonstrator

Overview

- Introducing DPHEP
- The Data Preservation use-case
- Mapping to EOSC services
- Status of deployment
 - What was easy
  - What was challenging
- Outlook

What is (HEP) data?　　　(And its not just "the bits")

Digital information
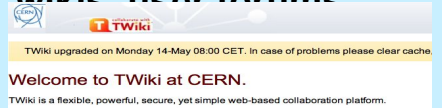The data themselves, volume estimates for preservation data of the order of **a few to 10 EB**

Other digital sources such as databases to ~~considered~~

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

ROOT
An Object-Oriented Data Analysis Framework
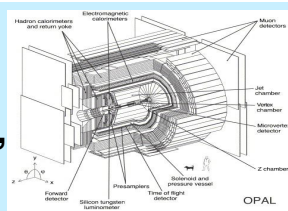
CERNLIB Access
- Access to the CERN Program Library is free of charge to all HEP users worldwide.
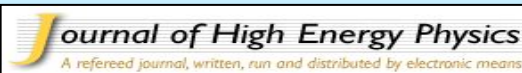- Non-HEP academic and not-for-profit organizations: 1KSF/year

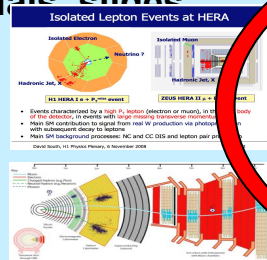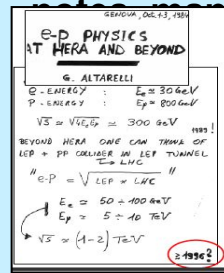## Meta information
Hyper-news, messages, wikis, user forums,

TWiki upgraded on Monday 14-May 08:00 CET. In case of problems please clear cache.

Welcome to TWiki at CERN.
TWiki is a flexible, powerful, secure, yet simple web-based collaboration platform.

SLAC NEWS CENTER

Publications
arXiv.org

inSPIRE

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

PHYSICS LETTERS B

EPJ C
Particles and Fields

Documentation
Internal publications, ~~notes, manuals, slides~~

Isolated Lepton Events at HERA

Expertise and people

LEP (HL-)LHC Timeline

| | 1980 | 1990 | 2000 | 2010 | 2020 | 2030 | 2040 |

LEP — Constr. Physics Upgr.

Data management, software, distributed computing (100TB/experiment)

LHC — Design, R&D / Proto. / Constr. / Physics

DM R&D, DBs, WLCG, EGI, Major Data Migrations!!

HL-LHC — Design, R&D / Constr. / Physics

ESFRI Roadmap "landmark project"
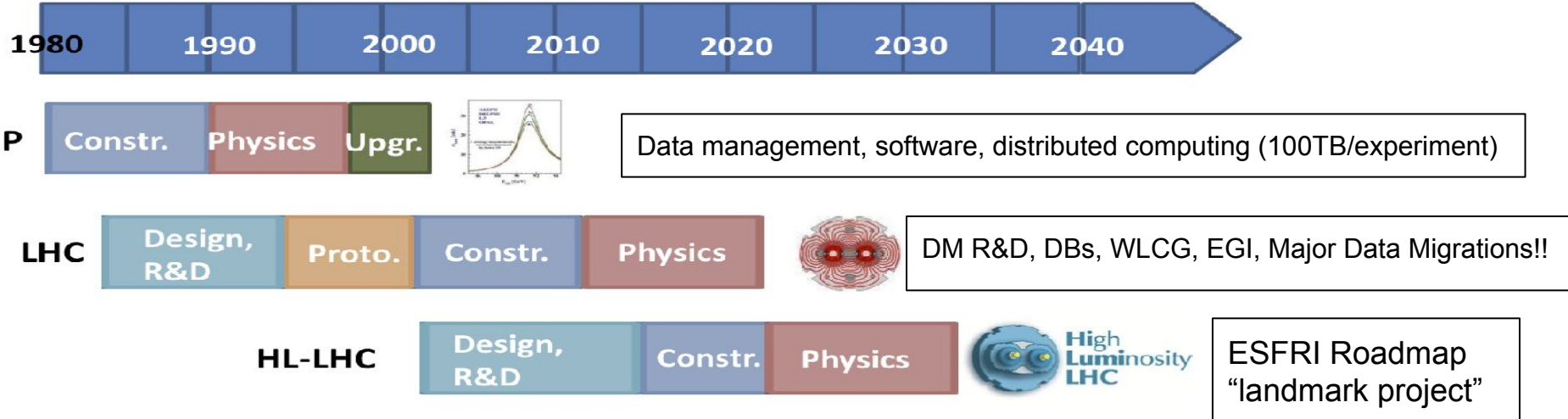
- Robust, stable services over several decades
- Data preservation and re-use over similar timescale
- Need to support transparent data migrations
- Data growing, 100TB, 100PB… Exabytes…
  - But DMPs could be the same (now and tomorrow)
  - And today's data volumes may be trivial for tomorrow's storage

Data Preservation - Demonstrator Use-case

Goal: Demonstrate "best practices" regarding data management in the arena of LTDP, "open" data (sharing and re-use) - how we can realize this on the EOSC.

- PIDs for data and metadata stored in TDRs
- DOIs for documentation
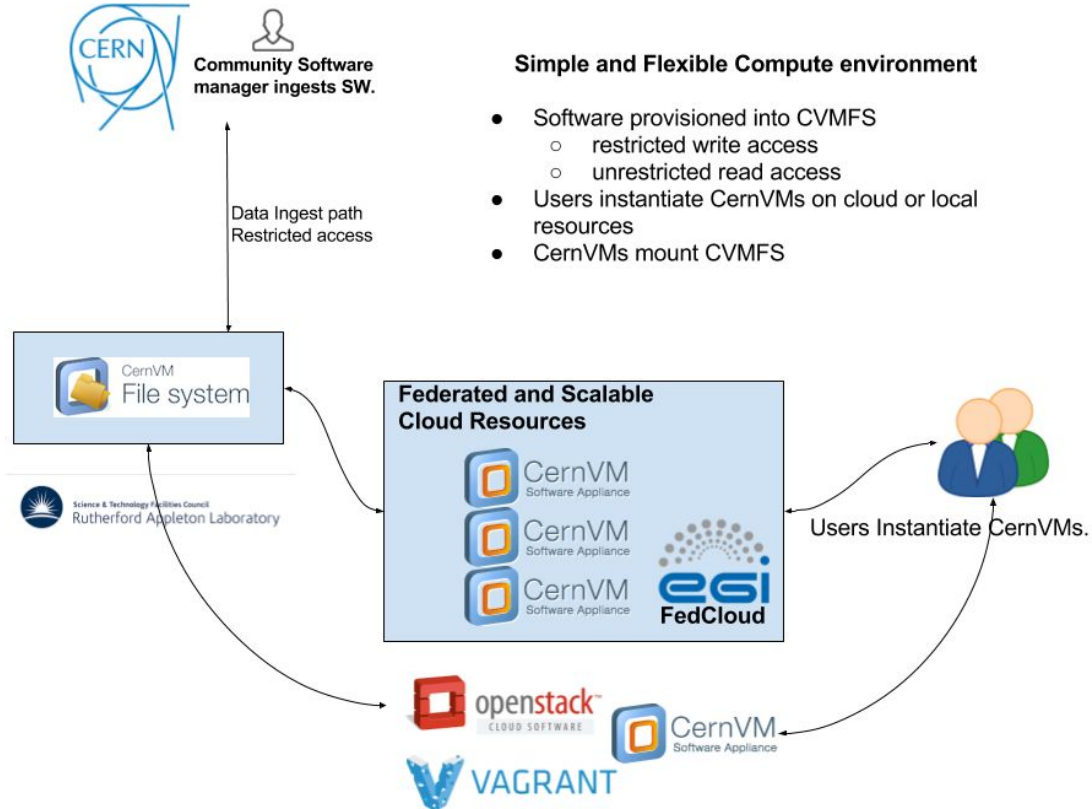- Expose and Archive the SW + environment

Equivalent to CERN Open Data Portal but using EOSC resources, thus allowing this solution to be opened to other communities.
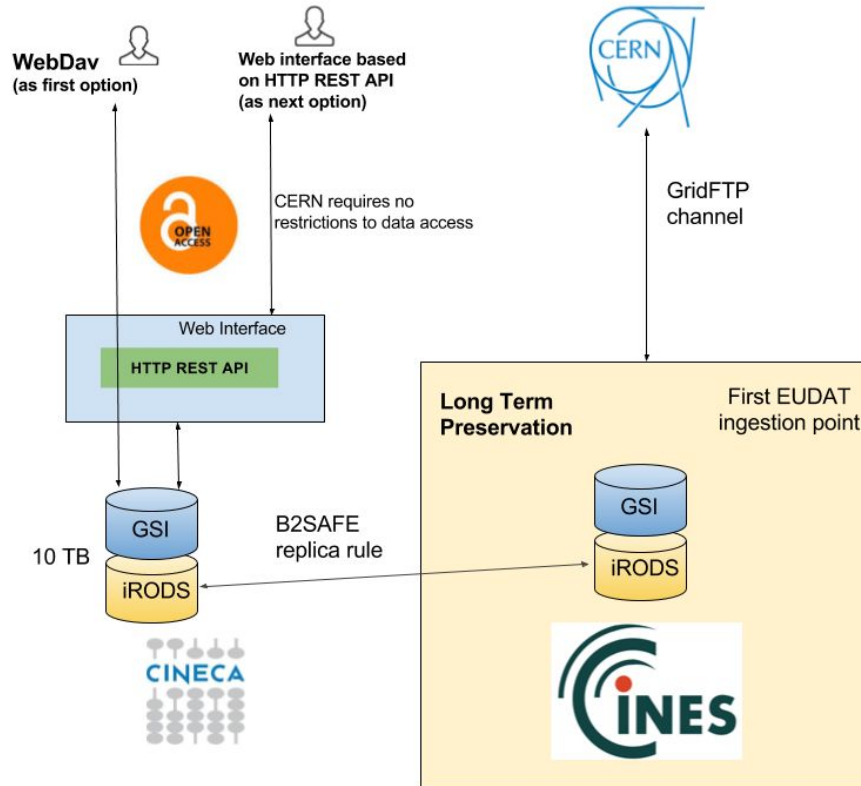
Mapping the use-case to services

Service	HEP	EOSC
Trustworthy Digital Repository (TDR)	CERN Castor+EOS	EUDAT TDR (part of CDI)
PID/DOI systems		EUDAT B2Handle
Digital Library	CERN Document Server	EUDAT B2Share (Zenodo)
Software + Environment	CVMFS + CernVM	CVMFS + CernVM Tested on EGI FedCloud

Mix of EGI and EUDAT services/resources required - good to show interoperation between e-infrastructures.

Software and Environment - Solution

Community Software manager ingests SW.

Simple and Flexible Compute environment

- Software provisioned into CVMFS
  - restricted write access
  - unrestricted read access
- Users instantiate CernVMs on cloud or local resources
- CernVMs mount CVMFS

Data Ingest path
Restricted access

CernVM File system

Science & Technology Facilities Council
Rutherford Appleton Laboratory

Federated and Scalable Cloud Resources

CernVM Software Appliance

CernVM Software Appliance

CernVM Software Appliance

egi FedCloud

Users Instantiate CernVMs.

openstack CLOUD SOFTWARE

CernVM Software Appliance

VAGRANT

Data Archive - solution

Status of Demonstrator

- Software and Environment:
  - CVMFS instance working
  - CernVMs tested on FedCloud and OpenStack/Vagrant
- Document Server:
  - B2SHARE - Documents uploaded to test instance
- Trusted Digital Archive:
  - In progress (big step forward)
  - Discussions regarding roles/requirements of communities and providers mainly done
  - Service to open data still to be deployed
- Conclusion: most of the boxes ticked, BUT the most difficult aspect is still being tackled!

Deployment - levels of difficulty

Relatively Easy: Fedcloud (on demand service)

Relatively Easy: B2Share Document store (on demand service)

Medium Difficulty: CVMFS (people in the loop)

Challenging: Archive solution (lot of people in the loop).

   Lot of discussion required, clarification of what is required and expected from both side. Need to have open data access made this more challenging.