DPHEP

Data Preservation in High Energy Physics

Science Demonstrator

Community PI: Jamie Shiers
Shepherds: John Kennedy, Matthew Viljoen

EOSC pilot
The European Open Science Cloud for Research Pilot Project

Pisa 13.09.2017

Overview

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

Note: This Demonstrator was undertaken using manpower from Shepherds and voluntary dphep community contributions.

DPHEP - Collaboration

- Partners: BNL, CERN, CSC, DESY, Fermilab, IHEP, IN2P3, INFN, IPP, KEK, SLAC, STFC…
- The collaboration aims to create a natural forum for the high energy physics community to foster discussion, archive consensus, and transfer knowledge on technological solutions and the diverse governance applying to the preservation of data, software, and know-how in the high energy physics community.
- “Active” since 2009 - workshops, study groups etc
  - High level but also many pragmatic and practical people ensuring that things get done
- Blueprint paper available - http://arxiv.org/pdf/1205.4667 (2012 update)

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

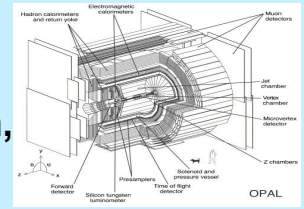
Digital information

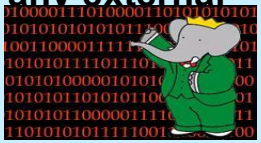
The data themselves, volume estimates for preservation data of the order of a few to 10 EB

Other digital sources such as databases to [illegible]sidered

Software

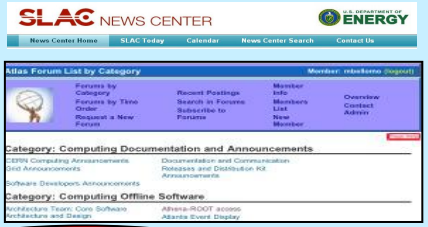
Simulation, reconstruction, analysis, user, in addition to any external





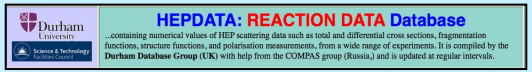
Meta information

Hyper-news, messages, wikis, user forums



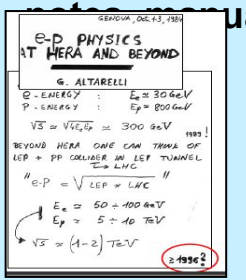
Publications

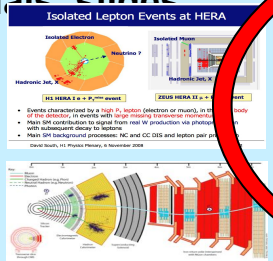
arXiv.org



Documentation

Internal publications, notes, manuals, slides

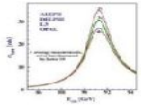




Expertise and people

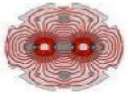
LEP (HL-)LHC Timeline

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

High Luminosity LHC

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

CERN

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

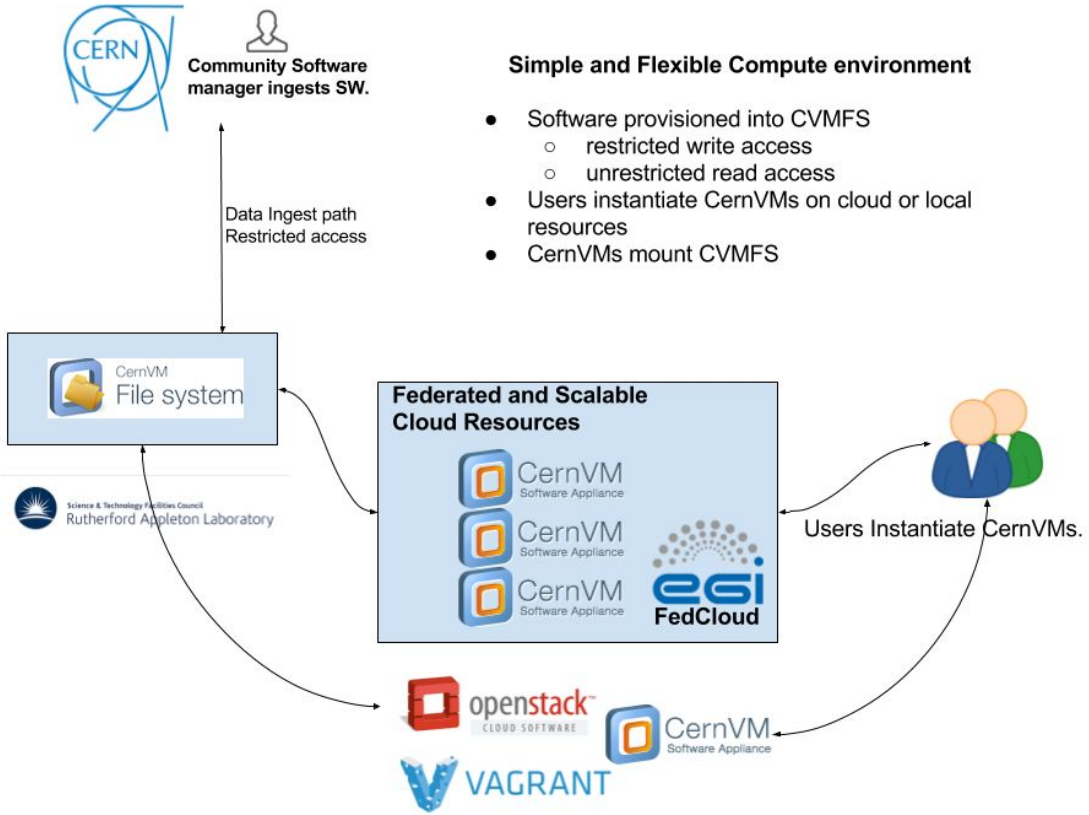
Equivalent to CERN Open Data Portal but using EOSC resources, thus allowing this solution to be opened to other communities.

Mapping the use-case to services

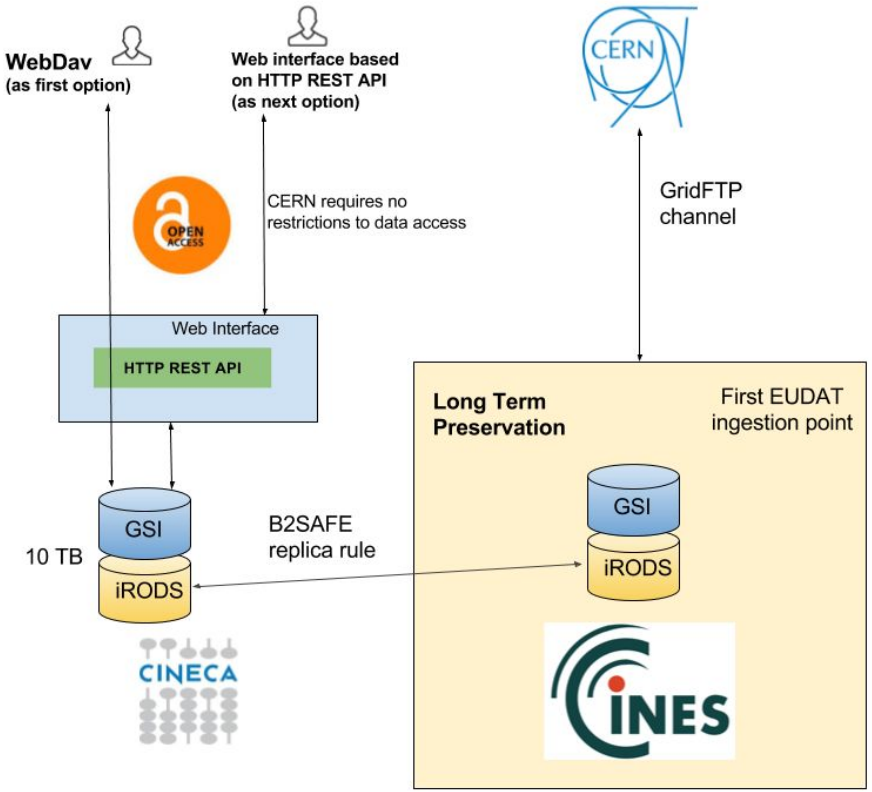
Service	HEP	EOSC
Trustworthy Digital Repository (TDR)	CERN Castor+EOS	EUDAT TDR (part of CDI)
PID/DOI systems		EUDAT B2Handle
Digital Library	CERN Document Server	EUDAT B2Share (Zenodo)
Software + Environment	CVMFS + CernVM	CVMFS + CernVM Tested on EGI FedCloud

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

Software and Environment - Solution



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
- Trustworthy Digital Repository:
  - 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.

Conclusions - so far...

General:

- Most of the pieces put in place
- Good example of a use-case for EOSC (not just technically)
- We've learned quite a bit too, especially w.r.t the archiving of data

Are we being FAIR?

- Very much so (w.r.t open and re-usable):
- Data is open, CernVMs are open, software is open via CVMFS
- No portal like CERN for finding data based on metadata (yet… B2FIND?)

Outlook

- Already took many positive steps!
- 3 months left….
- Complete Data management solution:
  - Ingest of data into EUDAT TDR
  - Replication of data to B2SAFE node with WebDAV
- Integration - end to end test
  - Create VM, download data, run basic analysis
- Wrap up:
  - Look at what worked well and what didn’t work too well
  - Feedback and suggestions - from both sides
  - Write final report

Thanks For you time
Any Questions..