Long-term preservation of analysis software environment

Challenge : How to reprocess experiment data in 25 years, using old software version?

Virtual Machine

- Linux distribution based on SL5.
- Supports all popular hypervisors.
- Small footprint. *(300 MB)*
- Comes in four editions targeting different use cases. *(Batch, Head node, Desktop and Basic)*
- Flexible contextualization.

CernVM Filesystem

- Read-only, globally distributed file system optimized for software distribution.
- Based on standard protocols *(ex. HTTP)*.
  - Highly scalable, redundant and reliable via a multi-Tier infrastructure.
  - Already used in production by LHC experiments.

CernVM - based data preservation

Assuming you have preserved your data files and recorded in the bookkeeping database the version string of CernVM you used, you will be able to recreate the same CernVM image, appropriate for a future virtualization technology and reprocess the data in the same way 25 years later,

- CernVM uses Conary software repository for automated, strict component versioning.
- You need only the CernVM version string to rebuild CernVM image on demand.

- Minimal cloud middleware is required to recreate virtual cluster for data reprocessing.
- CernVM can be contextualized using a small subset of EC2 API that allows it to be deployed on public or private clouds *(OpenNebula, OpenStack, Eucalyptus etc..)*.

Bookkeeping

Private Cloud