Running the scientific data archive

Costs, technologies, challenges

Jos van Wezel
KIT computer centre and big data archives

- University computer centre services
  - for KIT students (20,000+) and employees (9000+)

- HPC (T2 level)
  - for KIT and universities in the state of Baden-Wuerttemberg

- GridKa – LHC Tier 1
  - support for (particle) physics computing
  - stakeholders: 4 LHC experiments and several others
  - using gLite grid middleware (storage management dCache, xrootd)

- Large Scale Data Facility (LSDF)
  - support for data intensive computing and data storage for principally all sciences
  - stakeholders: Helmholtz Association, state of Baden-Wuerttemberg
Scientific archives at KIT data centre

- GridKa/LHC Germany 10 PB (+2 PB/a)
- BioQuant / Univ. Heidelberg 3 PB (+1 PB/a)
- HLRS Stuttgart (+1 PB/a)
- State of Baden-Wuerttemberg 0.5 PB (+0.5 PB/a)
- KIT 2.5 PB (+0.5 PB/a)
  - colourful bouquet of data sources
- EUDAT and Human Brain projects (+? PB)

- KIT has been designated as central site for long time digital data storage in the State of Baden-Wuerttemberg
  - scientific as well as public, heritage data
- 16 PB today, expect 5 PB/a growth after 2014
- Single copy! Dual copy needed for archived data
- Long time storage on tape
- Use of IBM and Oracle enterprise tape technology
Storage media density

Archive corporation
QIC24 60 MB Drive, ~1985
Maybe even no disks at all
Preservation  Long term access of LHC data

- Even at a certain cost one can argue that there is a price we have to pay to record past investments in science and society.
- The costs of finding the Higgs Boson at the Large Hadron Collider at CERN are roughly 10 billion (10^9) €. Taking into account hardware, media, power and maintenance, the cost for storing the raw data of would be in the order of 2 to 3 million € per year
- KIT currently calculates ~300 € /yr/TB disk+tape copy
- The bulk of the LHC data today is stored on magnetic tape, not magnetic disk
- CERN uses home written SW for bit preservation
- Archival is heavily discussed

![Cost per period, breakdown by category](chart.png)

courtesy: CERN/Shiers
Archives and archive storage

- Communities take care of selection, ordering, ingesting of data
- Community or ideally archivists representing the community, takes care of curation, deletions, pruning, formats,…
- **Content Preservation, digital preservation**

- The data centers, i.e. KIT, take care of **bit preservation**
- Contrary to popular belief bit preservation is not an open and shut case
- Access is not standardised
- The bits that go in are the bits that come out
  - After 10, 100, 1000 years?
  - Integrity check for an exabyte?
big archives, big efforts

- Archive is of no use if you cannot access the data.
- Information technology has alleviated the problems of finding and sorting paper.
- And it also facilitated the explosion of data creation.
- And the exponential growth of the problems associated with managing, sorting, protecting, and archiving data.
- Libraries and museums will be confronted with large data sets.
- KIT/SCC research for data life cycle management and support.
Reliability

- Common practice is to store 2 copies (data stored on 2 media)
  - Some sites even don’t
  - Possibly on 2 different locations
  - Hardliners argue the copy should be made with totally different software and hardware

- Theory, theory
  - if we have no measurement, then how can we improve?

- The only available number is the Bit Error Rate
  - Estimated! Average!
  - Tape enterprise is $10^{19} - 10^{20}$ (enterprise disk is $10^{15}$)
  - Prefer something like the mean time to data loss (MTTDL)
  - Even then: loose a PB on average or a kB?
  - Around 2030 we’ll store a PB on a single cartridge
Access to data in deep archives

- Data interfaces assume disk interface i.e. on-line storage
  - Archivematica, Fedora, Invenio, Fixity etc.
  - RDF and NoSQL DBs
  - Interfaces still keep with POSIX

- No consideration with regard to off-line, or replicated, storage
  - Assume different storage is handled behind the scenes
  - No one waits for information to become on-line
    - Unless you know its will come on line

- Requirements from the data center perspective
  - How to handle asynchronous data access
  - How to do efficient offline check summing (with user provided method)
    - Applications should not DOS the (tape) storage systems
  - An upstream reference
    - how to get from content to the ‘owner of the data’
  - Fast data center exchange
    - changing resource providers, disaster recovery
Archive interface

- work in progress of the RADAR and bwArchive projects
- revisit the 2008 SNIA XAM standard
  - Asynchronous access to data on off-line storage
- Comes with an emulator / reference implementation
- Allows modular adaptation to different back-ends through Vendor Interface Modules (VIM)
Challenges

- **Determining the rate of bit rot and taking countermeasures**
  - we have no metric for archive (bit preservation) quality
  - this somehow should be part of a site certification

- **Intelligent selection of candidates for caching**
  - getting data from a deep (i.e. cost effective) archive takes time
  - on line caches in front of archives for fast access
  - what goes to and what data stays in the cache
  - Notions: trends, markets, relationships: maybe google can help?

- **Uniform interface for “cold” storage**
  - API that goes beyond classic POSIX but takes longer access times
  - Exchange data sets or data volumes with peers for redundancy or for contract changes: asynchronous operations
Summary

- costs
  - use tape if archive storage grows beyond ~5 petabytes

- reliability
  - very good, however need deterministic numbers

- interface
  - must be ready before KIT stores 100 PB
Thank you for your attention

Contact: jos.vanwezel@kit.edu
Tape media costs at KIT
Generic Archive Interface (in development)

Access Control
IO & caching

Administration
Bit Preservation

Disk
Metadata

Tape
Tape

High Performance Storage System (HPSS)

KIT
What is an Archive? SCC view

Definition of an Archive:
“A long time file repository containing a massive collection of digital information”

- No loss of data or degradation of quality for an agreed amount of time
  - SCC ensures proper integrity
- Does not depend on proprietary middleware or storage formats
  - Migration should be possible in some number of years and must be independent from content or owner
- Data is stored in (self-defining) containers
  - Containers are not files, they are “bitfiles” with a checksum, they may contain one or more files.
  - Reference to the container is fixed for the lifetime of the container
- The operator (SCC) is not responsible for passing content on to the next generation of owners
  - Data ownership has a lifetime that must be known in advance, but can be renewed
  - Data is stored after consent of a data provider and SCC. It is effectuated in a service level agreement for archives.

The above intentionally excludes searching, finding and global referencing.
Main SCC archive leads

- Installation and deployment of the HPSS system for managing large scale disk and tape resources. The system will replace TSM for long time storage tasks.
- Description and implementation of a common interface to archival storage to include asynchronous access to storage, data integrity assurance and coupling to IDP based authentication schemes. Implementation of prototype tools and applications using the common model.
- Development of service level agreements, archive storage classes, cost models, data protection practices and participation in archive certification qualifications
- Deployment of archive services for scientific data and data from heritage institutions