Advancing data management and analysis in different scientific disciplines

M. Fischer, M. Gasthuber, A. Giesler, M. Hardt, C. Jung, J. Meyer, F. Rigoll, K. Schwarz, R. Stotzka and A. Streit

CHEP 2016

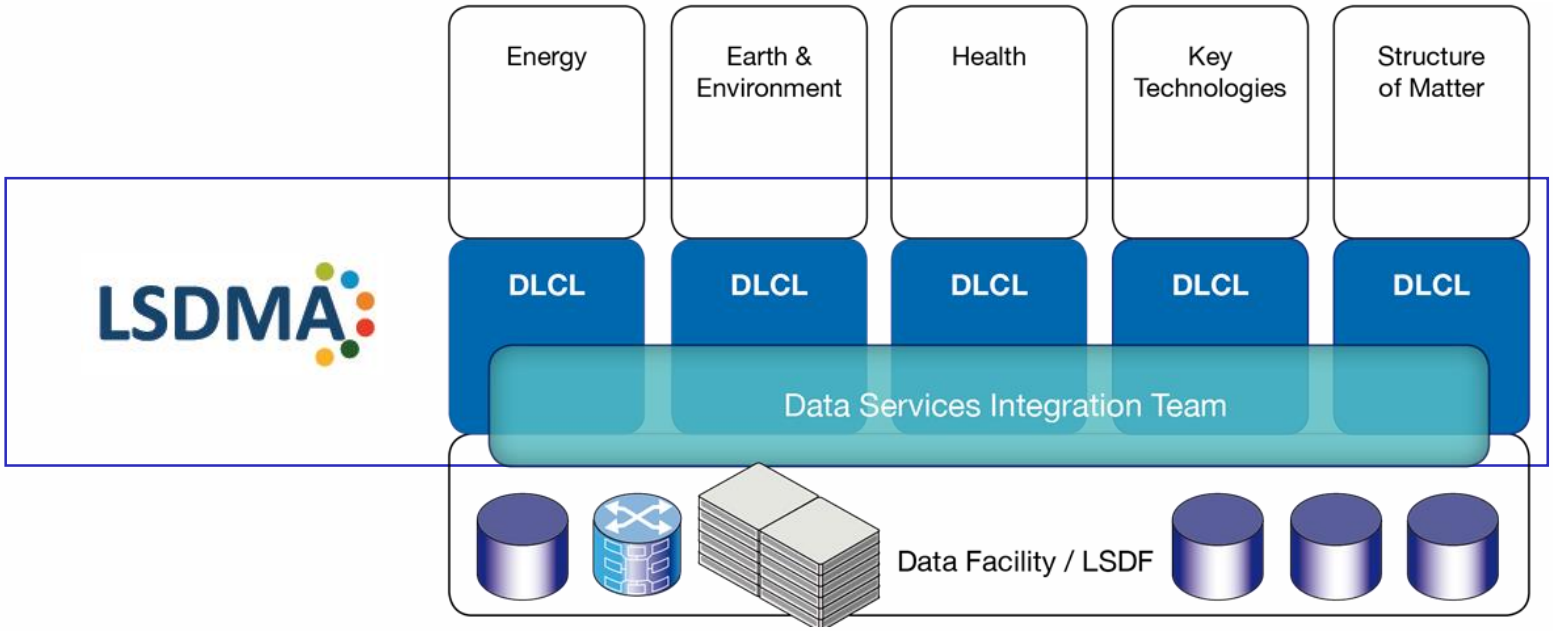
Overview

LSDMA Overview

Selected Highlight Activities

Experiences and Summary

LSDMA: Dual Approach



Data Life Cycle Labs

Joint R&D with scientific user communities

- Optimization of the data life cycle
- Community-specific data analysis tools and services

Data Services Integration Team

Generic methods R&D

- Data analysis tools and services common to several DLCLs
- Interface between federated data infrastructures and DLCLs/communities

LSDMA Facts & Figures

Initial duration: 2012-2016

- Project is a Helmholtz portfolio extension → inclusion of activities into Helmholtz program-oriented funding in 2015, cross-program initiative

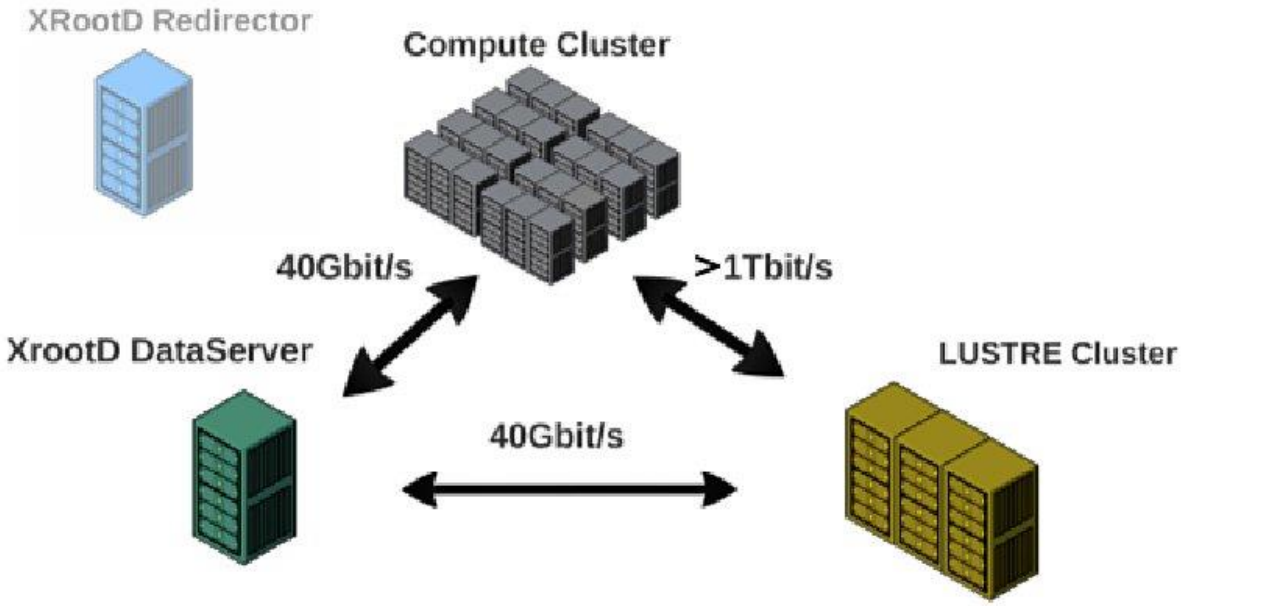
Partners:

- Helmholtz Association: KIT, DESY, FZJ, GSI
- External: DKRZ, U-Heidelberg, U-Ulm, TU-Dresden, U-Hamburg, HTW-Berlin, U-Frankfurt

Coordination: KIT

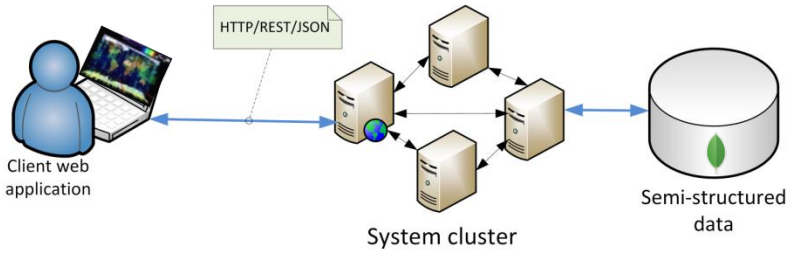
DLCL Structure of Matter: FAIR@GSI

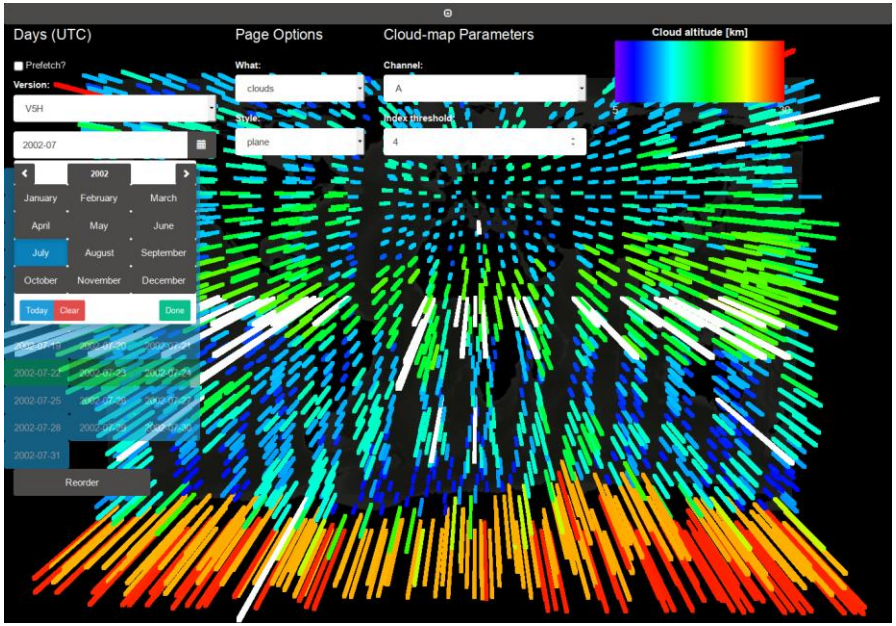
- GSI hosts the German ALICE T2 centre, providing 7% of the ALICE T2 resources
- ALICE T2 jobs run in a multi - purpose HPC centre. Data to/from the HPC environment are tunneled through an XrootD forward proxy
- Storage Element: XrootD data servers on top of Lustre file system - an XrootD client plugin can provide direct access to Lustre for jobs running at GSI.
- experience gained in context of ALICE T2 provide an important guideline for the forthcoming distributed computing environment of FAIR



DLCL Climatology

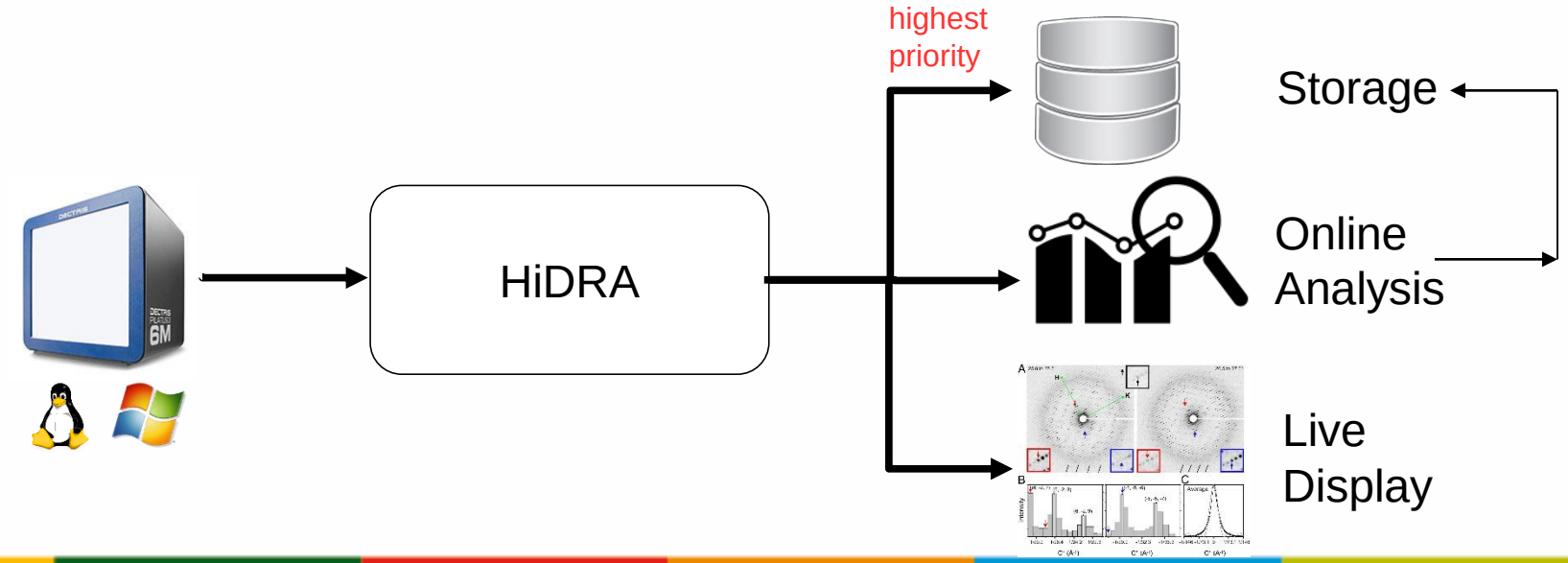
- browser application for visualization of large amount of unstructured data (MIPAS)
- 3D-visualization: WebGL
- MEAN stack
- web service API for MongoDB and predefined use cases
- node.js cluster to perform preprocessing of MongoDB data
- framework with frontend/backend workers, controller, scheduler





DLCL Structure of Matter: HiDRA – Petra III & Flash @DESY

- Problem:
  - data has to be drained from the detectors fast enough (>30Gb/sec)
  - experimental conditions have to be monitored and analyzed in close to real time to prevent the collection of unfavorable data, which also helps with preserving the valuable sample (online analysis)
- HiDRA is a generic tool set for high performance data multiplexing with different qualities of service based on Python and ZeroMQ



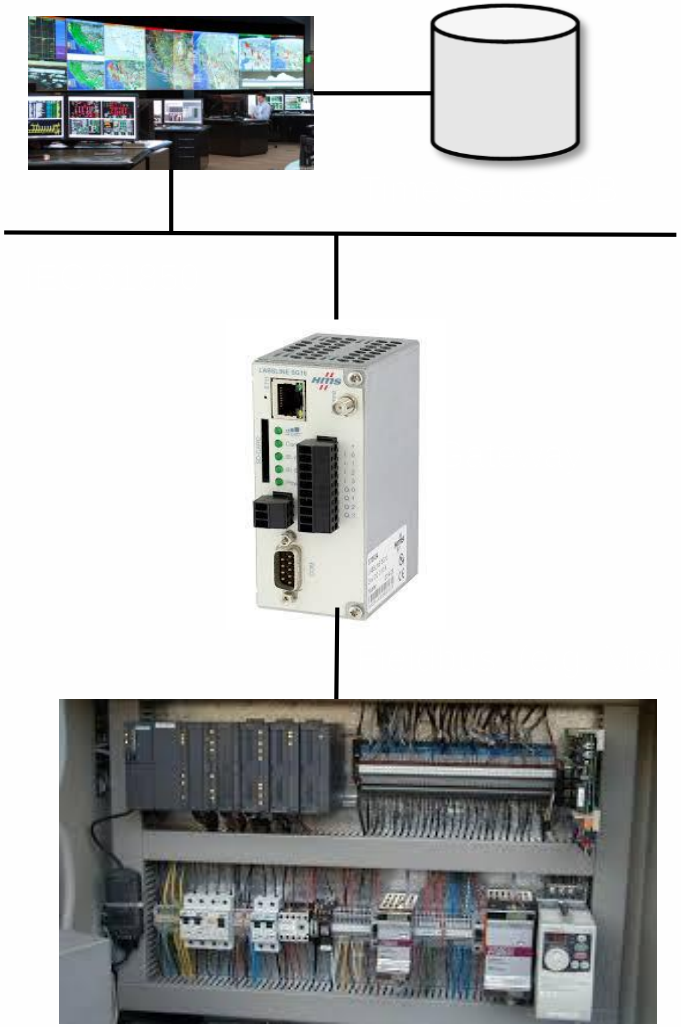
DLCL Energy: Energy Lab 2.0

Energy Lab 2.0: An energy grids simulation and analysis laboratory incl.

- Simulation platform
- Microgrid - test field
- Power-hardware-in-the-loop
- KIT Energy Smart Home Lab
- FZI House of Living Labs

Main tasks of the DLCL Energy in the Energy Lab 2.0:

- Energy data / time series management
- Standardized data access



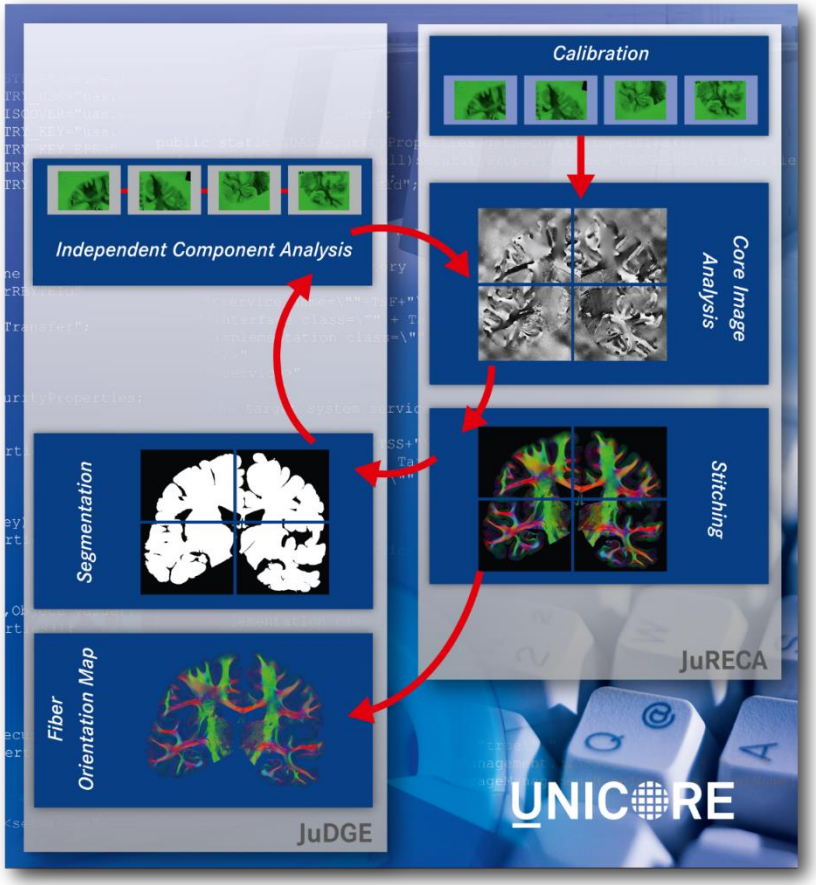
DLCL Neuroscience: 3D Polarized Light Imaging Workflow

Goals and Methods

- Generation of high-resolution, large-scale brain models describing nerve fiber architecture
- Acquisition of imaging data by Polarized Light Imaging (3D-PLI) technique
- 2500 serial images of histological sections per human brain
- ~750 GB data processed per section and 3D-PLI workflow execution
- Complex chain of image processing tools
- Integration of tool chain in UNICORE workflow system

Results

- Significant speedup of execution time by level of automation (reduced from days to hours)
- Abstraction - Enabled users to perform complex data analysis without deep knowledge of tools
- Large-scale distributed supercomputing and data sharing across various platforms



DLCL Key Technologies

Data repositories

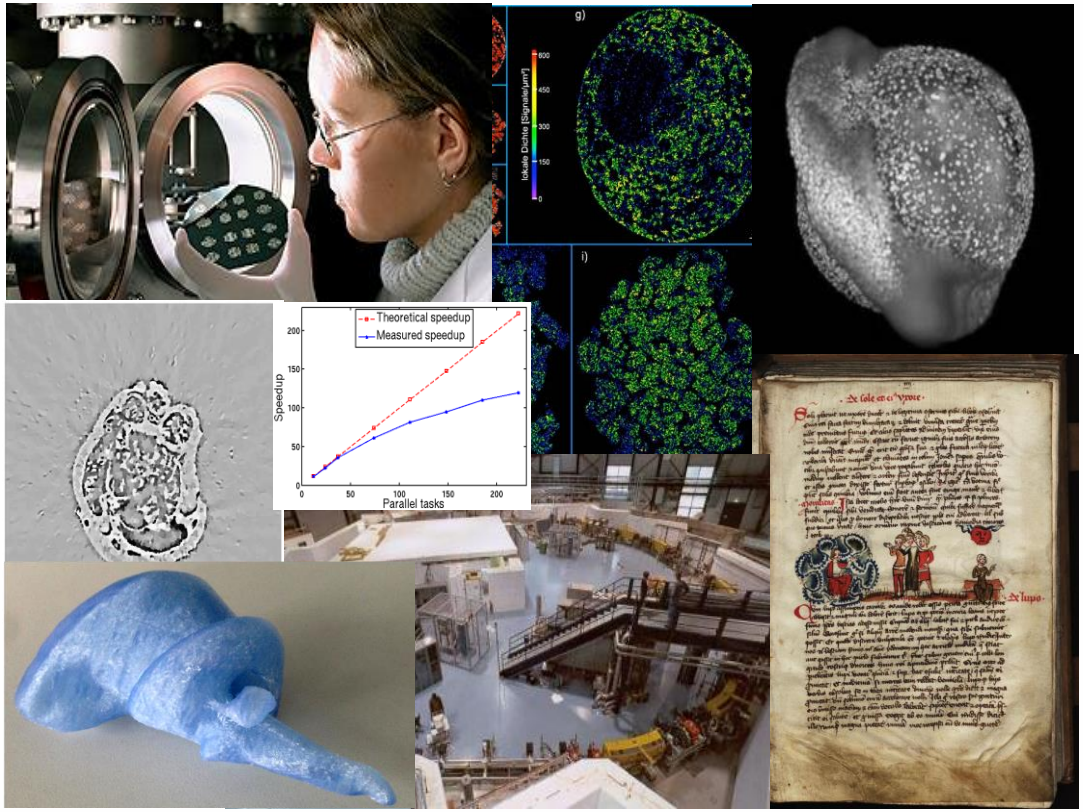
- Distributed
- Special-purpose

Data description for re-use (metadata): data provenance

Data analytics:

- Automatic feature (metadata) extraction
- Tools

Algorithms



DSIT Highlight: align AAI with European context

- Authentication and Authorisation Infrastructure:
  - Concepts and Pilots:
    - => AARC (Authentication and Authorisation for Research Communities)
  - Infrastructure development and deployment:
    - ⇒ INDIGO-DataCloud

- Building the bridge between X.509, SAML and OpenID-Connect (OIDC)
  - Users are in SAML-home IdPs
  - Services may be in OIDC, X.509 and SAML

DSIT: More Highlight Activities

Unicore

- Extended to support REST (not just SOAP)
- Extended to support map-reduce
- Generic interface for metadata management

Extensive performance measurement and tuning of Lustre / ZFS

- Adaptive Data Reduction
- Parallel Snapshots

KIT-Data-Manager: a repository system for large data released as open source

dCache

- CDMI Support
- QoS for Storage support

Experiences and Summary

Experiences:

- Needs of communities vary immensely, even within research areas
- Communities‘ interest in new tools and methods fueled by need and by new research potential
- Automation of procedures still very often a quick win
- Interoperable AAI well on its way
- Policies (e.g. Open Data) and legal regulations (e.g. data privacy) are additional challenges

Summary

- For five years, LSDMA has advanced selected scientific communities in their data management and analysis
- Communities highly profited from this work
- Dual approach was a great choice