Karlsruhe Institute of Technology

Large Scale Data Facility for Data Intensive Synchrotron Beamlines

R. Stotzka, W. Mexner, T. dos Santos Rolo, H. Pasic, J. van Wezel, V. Hartmann, T. Jejkal, A. Garcia, D. Haas, A. Streit

Institute for Data Processing and Electronics | Steinbuch Centre for Computing | Institute for Sychrotron Radiation

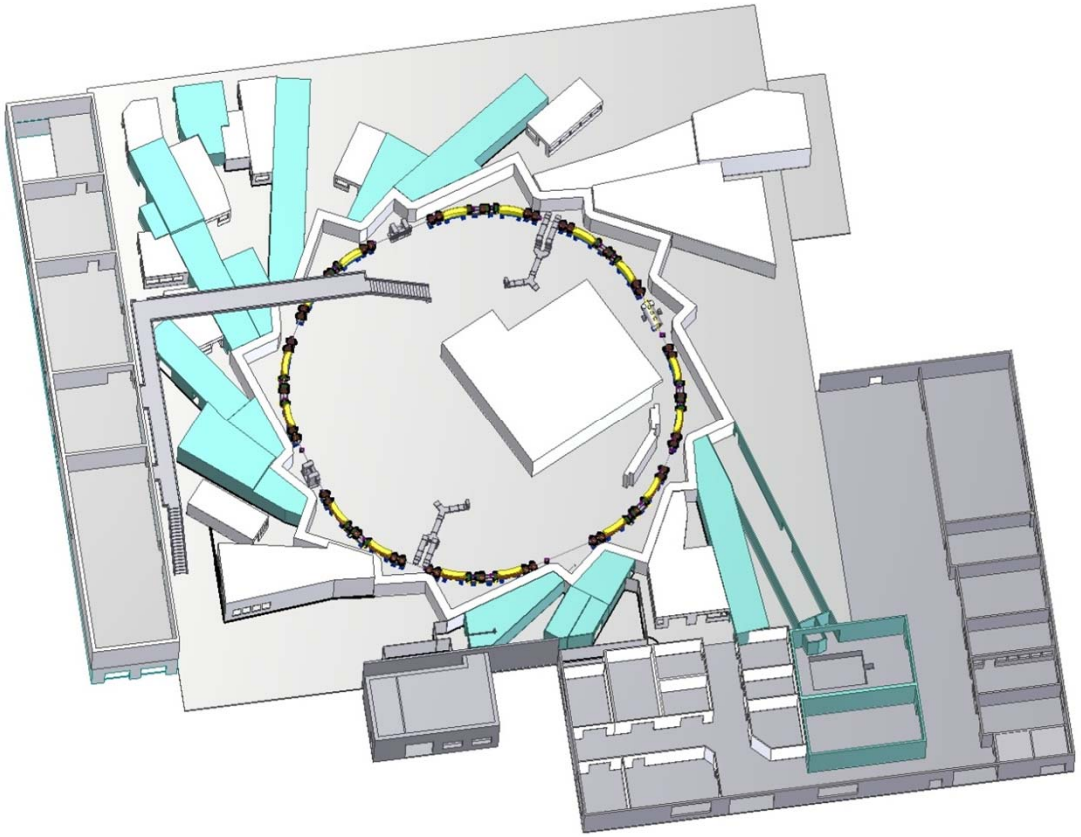
ANKA – Synchrotron Light Source at KIT

14 Beamlines operational

Far infrared – hard X-ray

Methods:

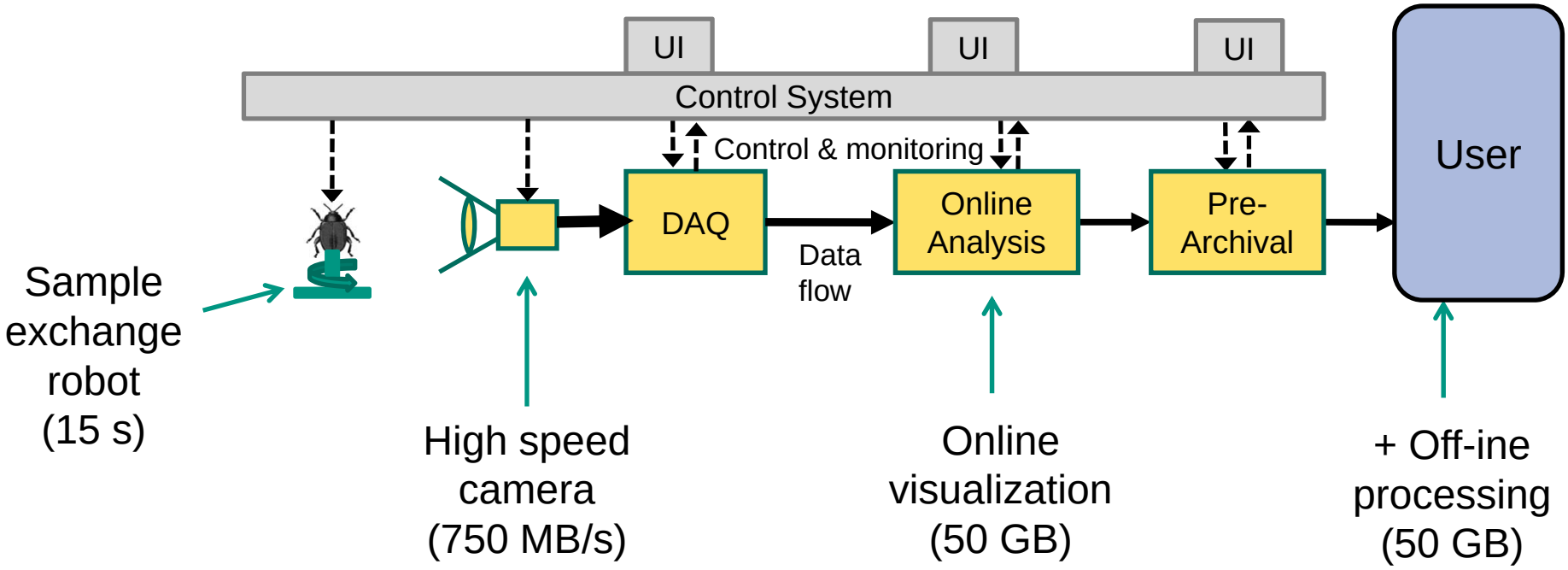
- **Topo-Tomo**
- Spectroscopy
- Fluorescence
- Diffraction
- Lithography



Under construction:

- High-resolution X-ray diffraction
- High-resolution **IMAGE**

Data Intensive Beamlines at ANKA
Topo-Tomo and IMAGE



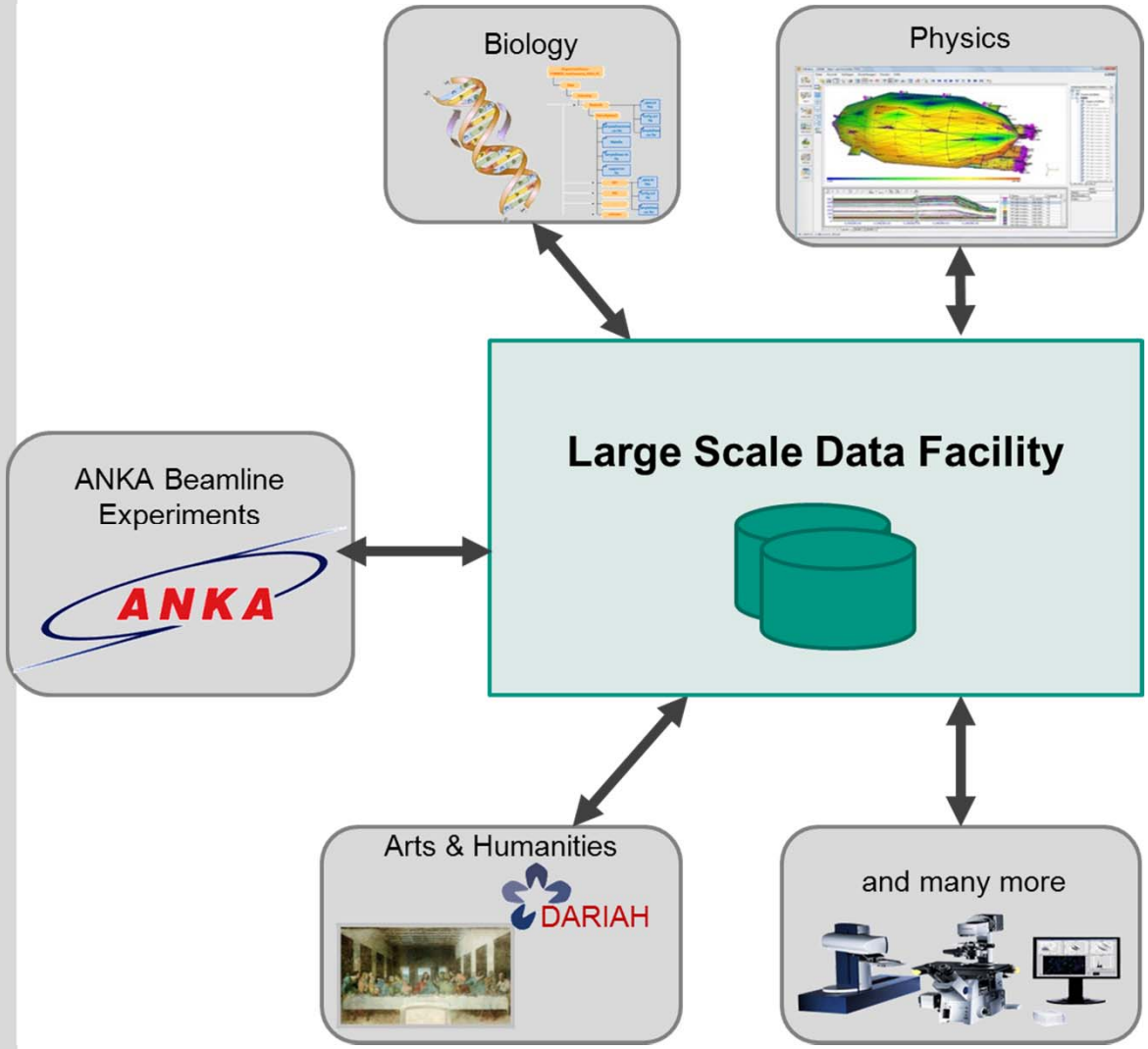
Storage requirements: several PB/a

Now and the Future

	Now	Near future
Sample exchange	Manual	Automated
DAQ	Automated	Automated
Pre-processing	Automated	Automated
Online visualization	Manual	Automated
Offline processing	Manual	Manual/automated
Data management	Manual	Automated
Meta data	Incomplete	Full description
Storage and archival	Manual	Automated

- Fast changing experiment environments
- Need for a flexible automated data management

Large Scale Data Facility (LSDF)



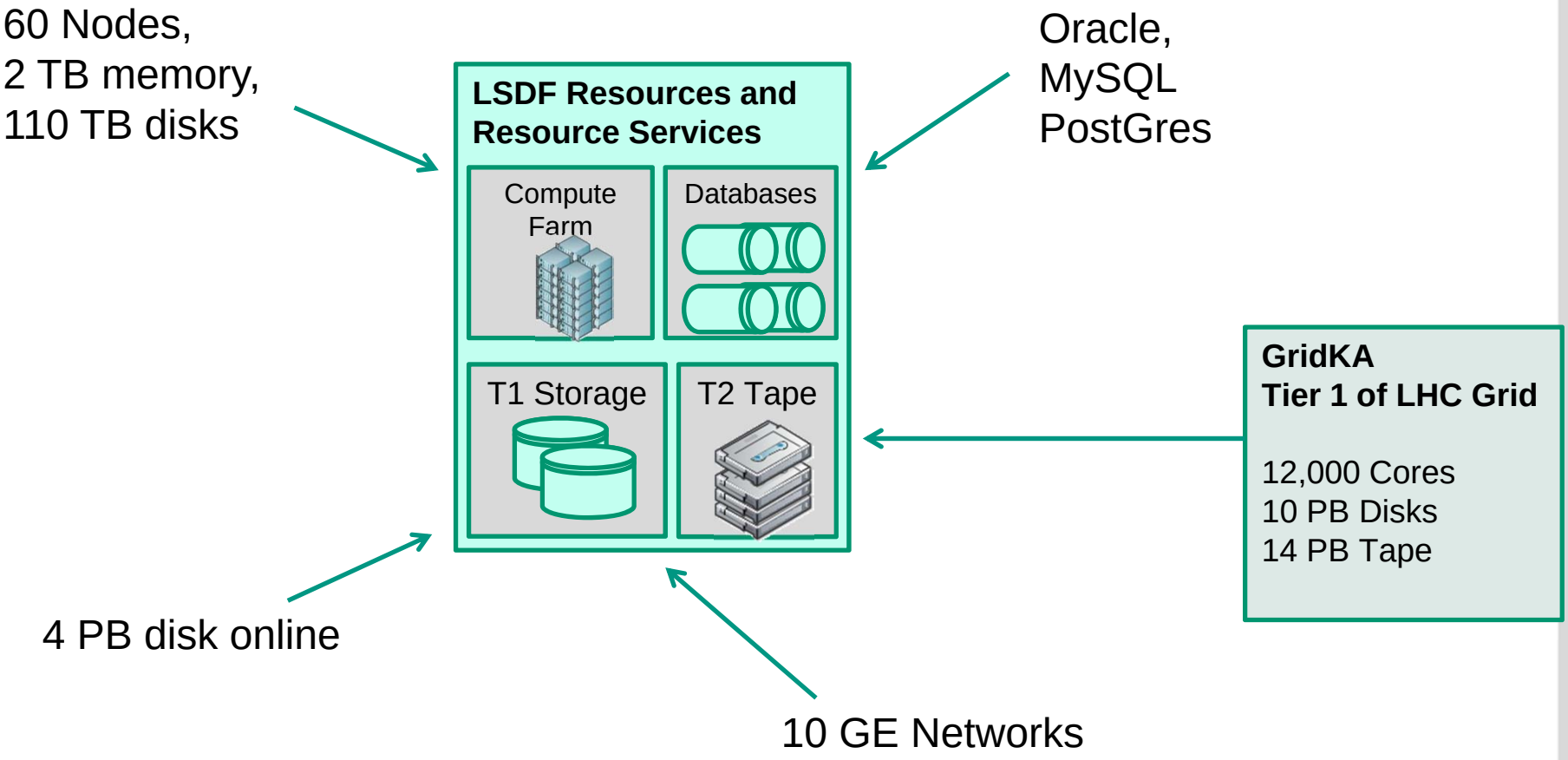
Research Programme **SuperComputing** of the Helmholtz Association

Aims:

- Provide storage
- Provide archives
- Provide data services
- Provide data management
- Provide support
- Provide data intensive computing
- Provide data analysis

For heterogeneous research communities with heterogeneous requirements

LSDF Resources

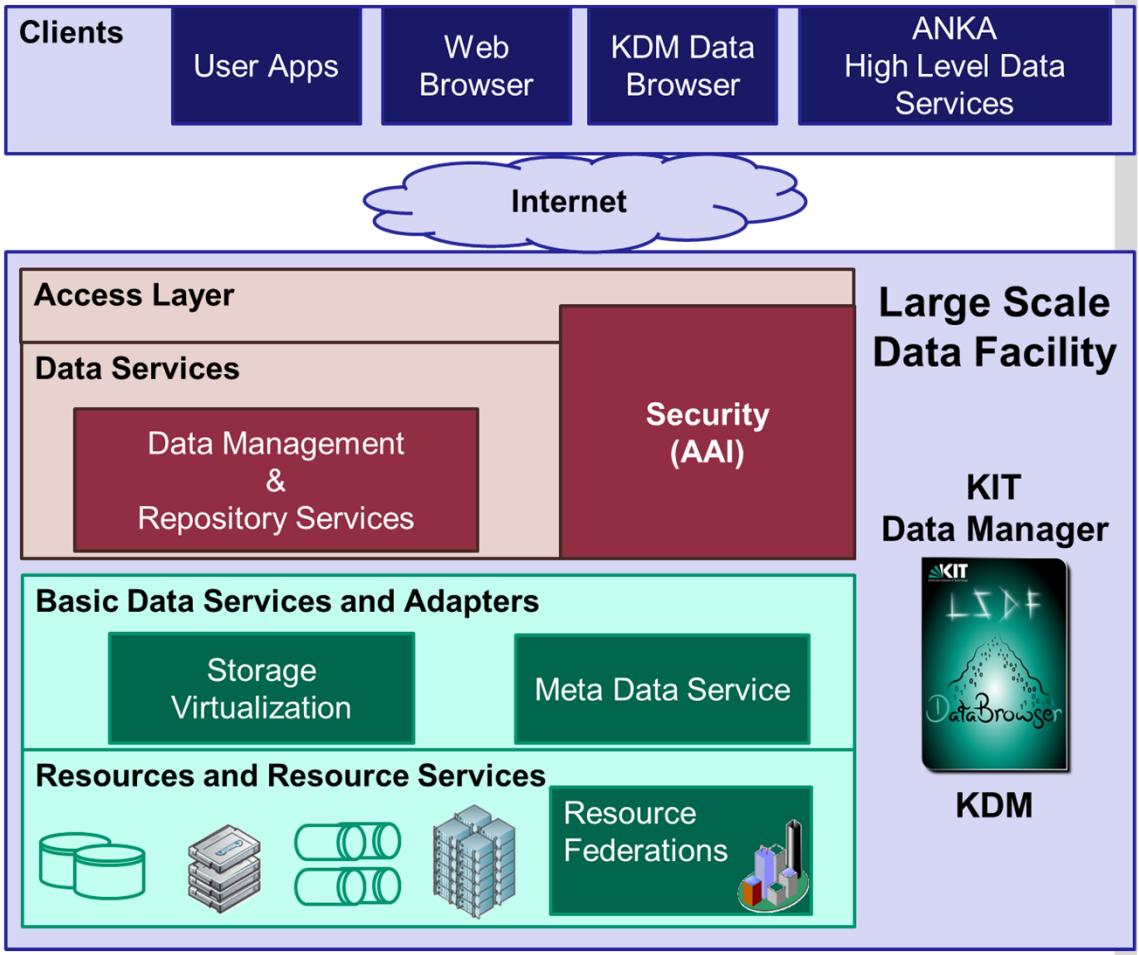


Software and Service Development

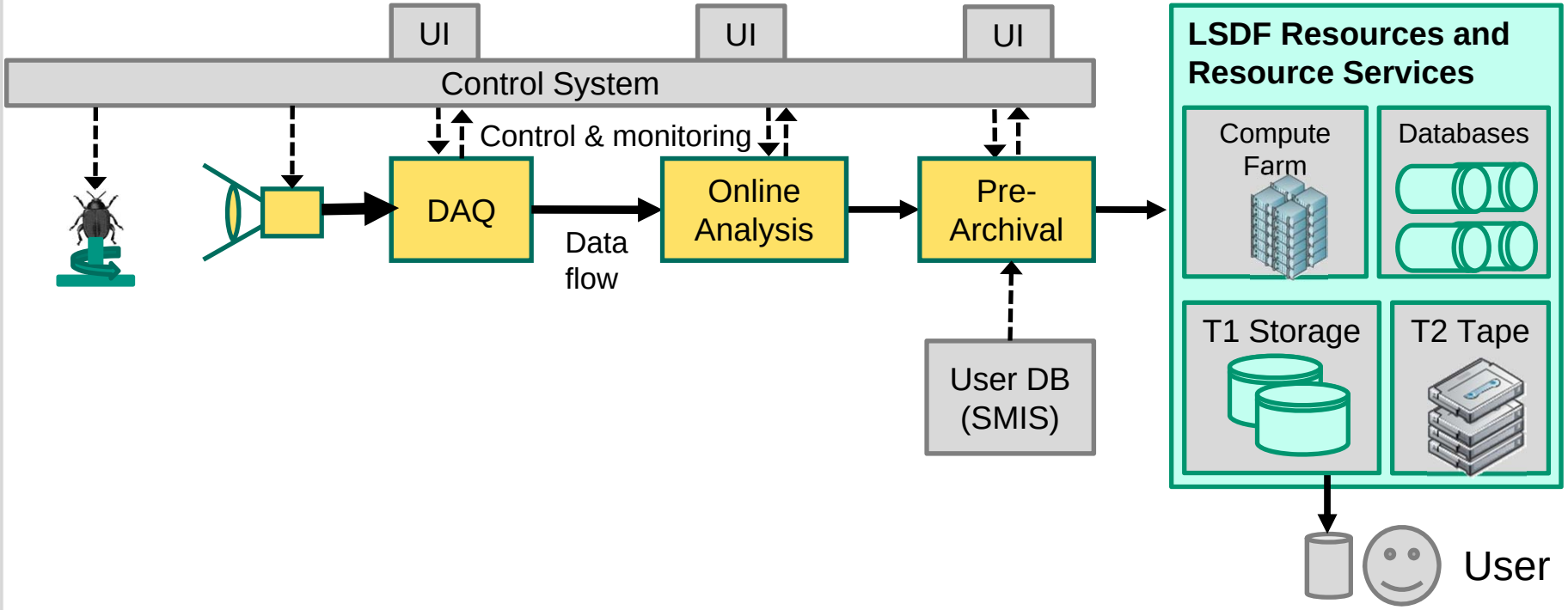
KIT Data Manager (KDM)
Software and service infrastructure

KDM DataBrowser
- Graphical user interface
- Data ingest
- Data access
- Meta data management
- Processing launch

LSDF Execution Framework for Data Intensive Applications
- Development
- Deployment
- Ingest

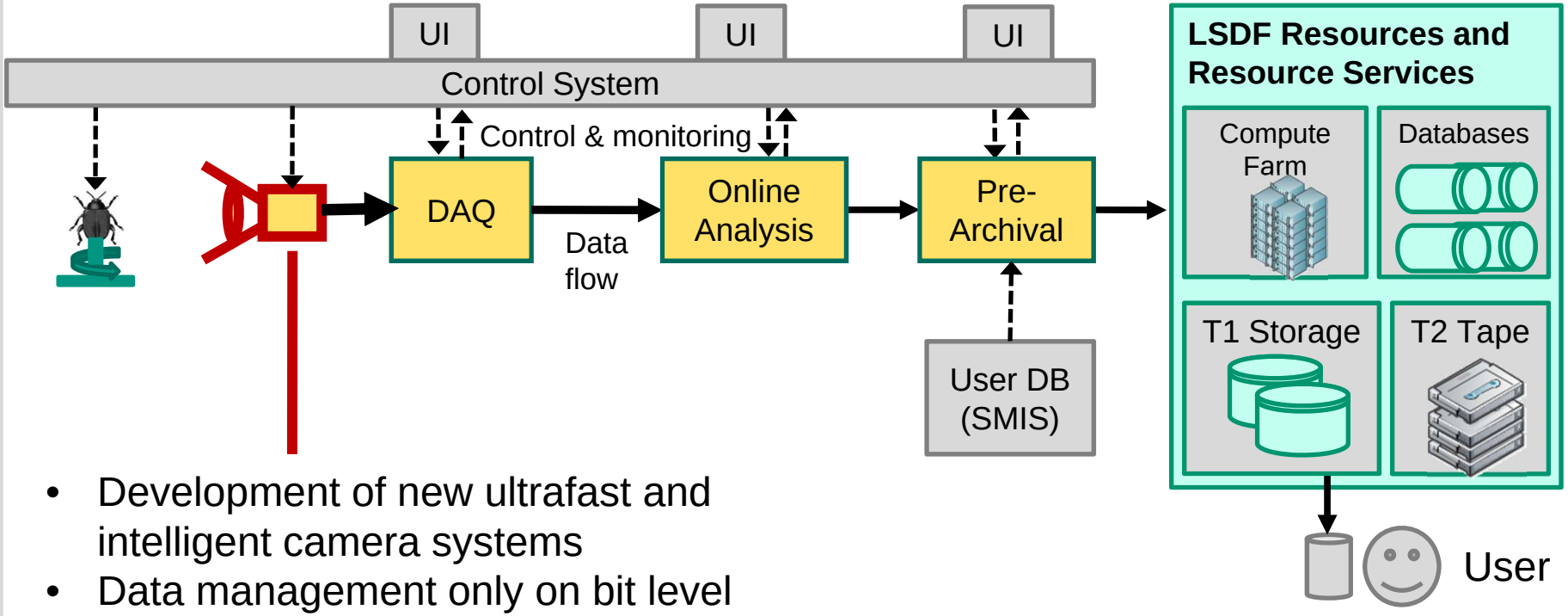


Imaging Beamline Data Management



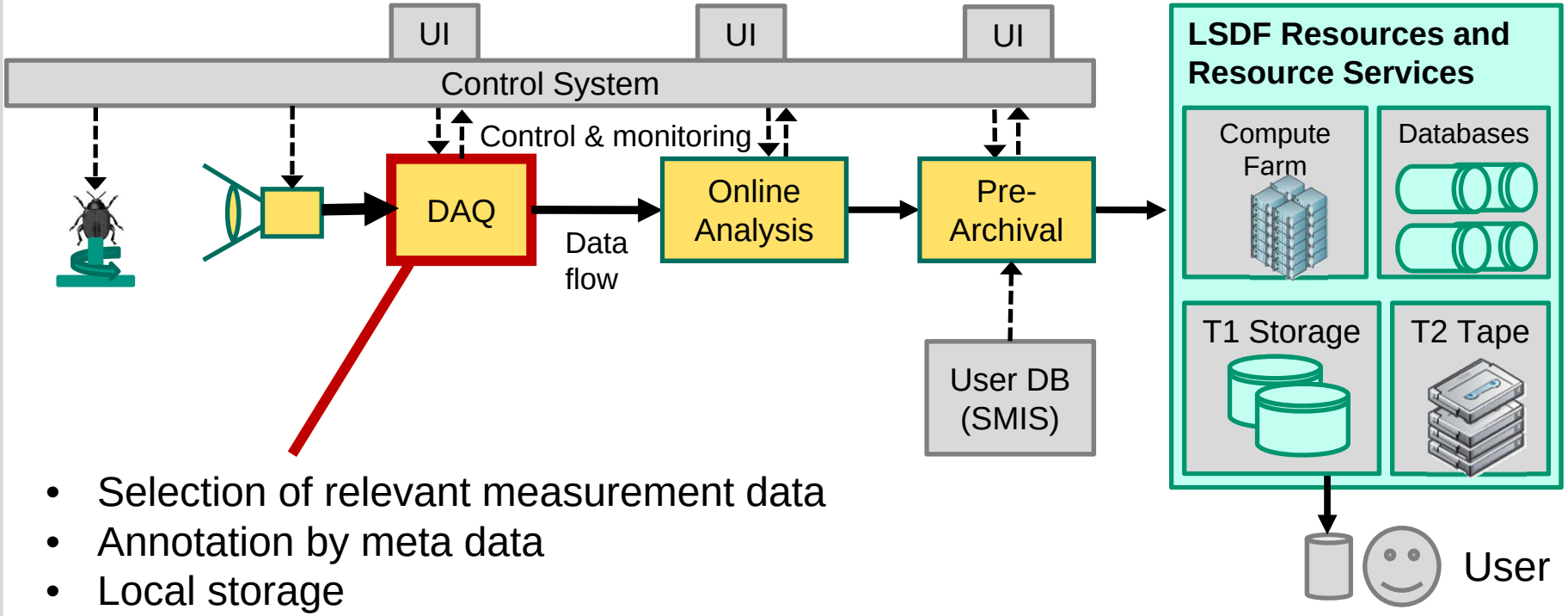
- Fast changing experiment environments
- Need for a flexible automated data management

Imaging Beamline Data Management



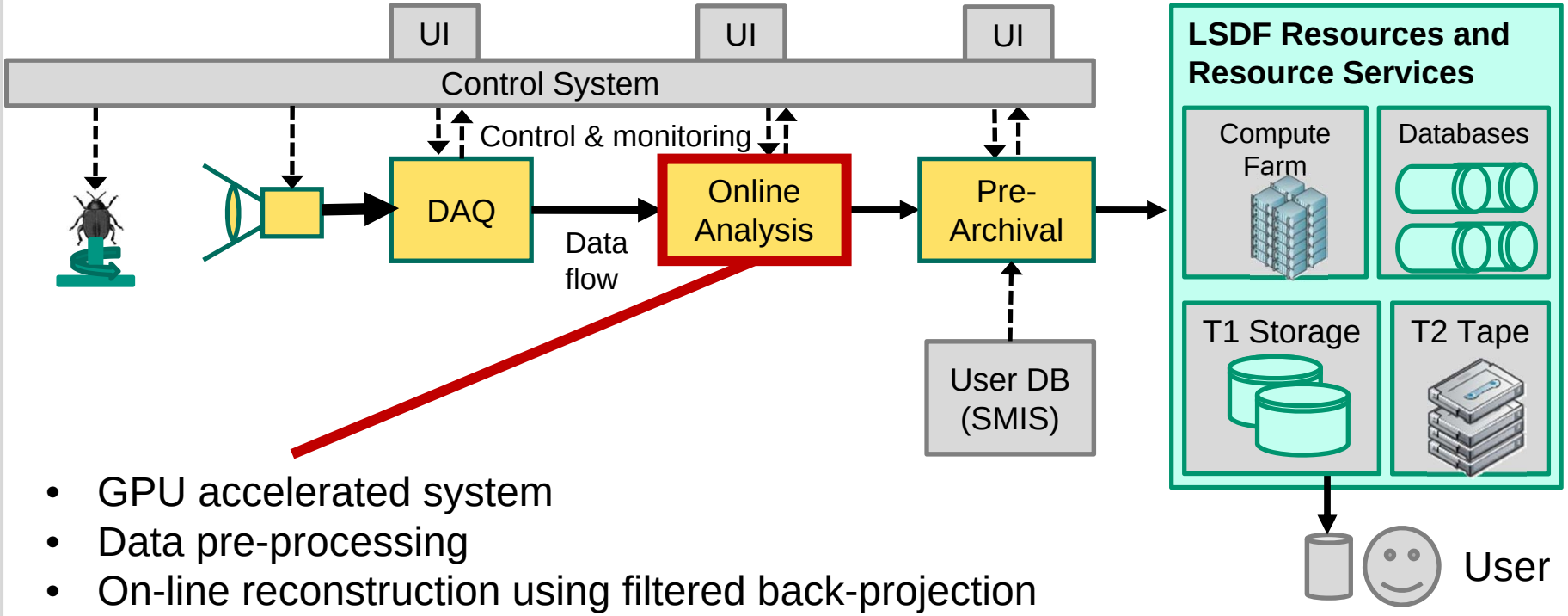
- Development of new ultrafast and intelligent camera systems
- Data management only on bit level

Imaging Beamline Data Management



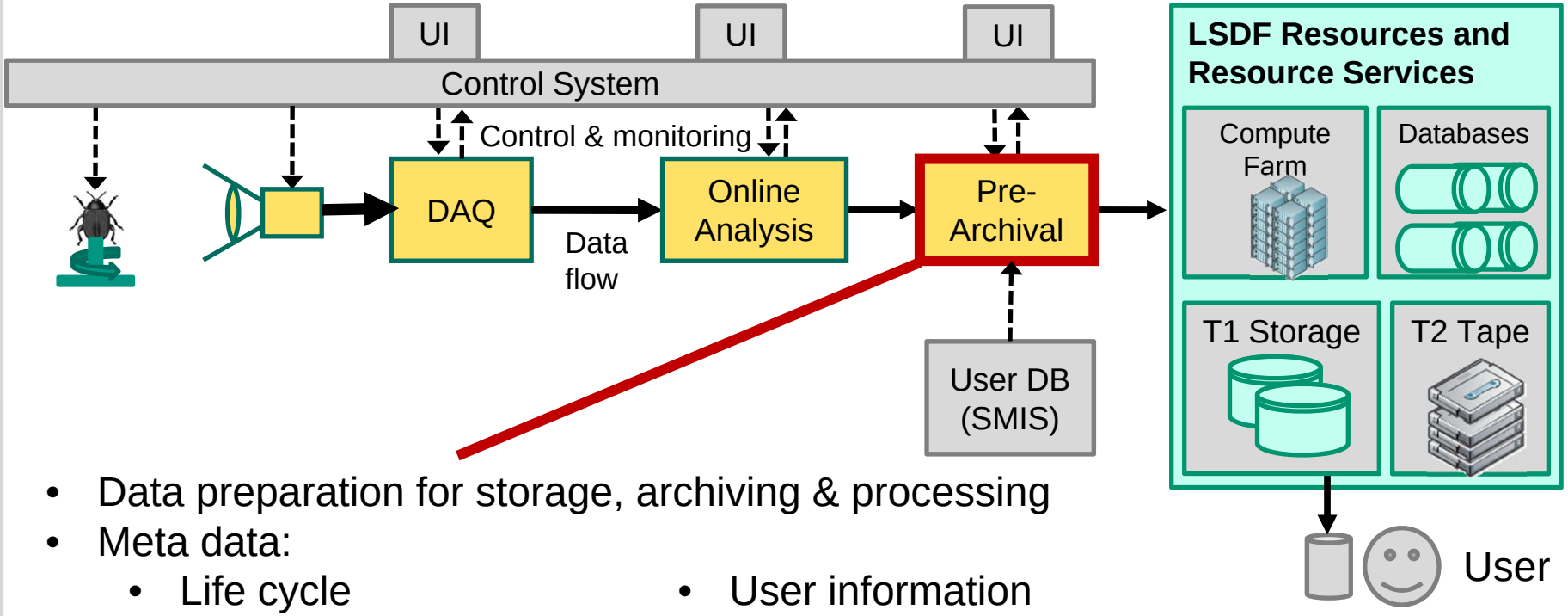
- Selection of relevant measurement data
- Annotation by meta data
- Local storage
- Unique data format

Imaging Beamline Data Management



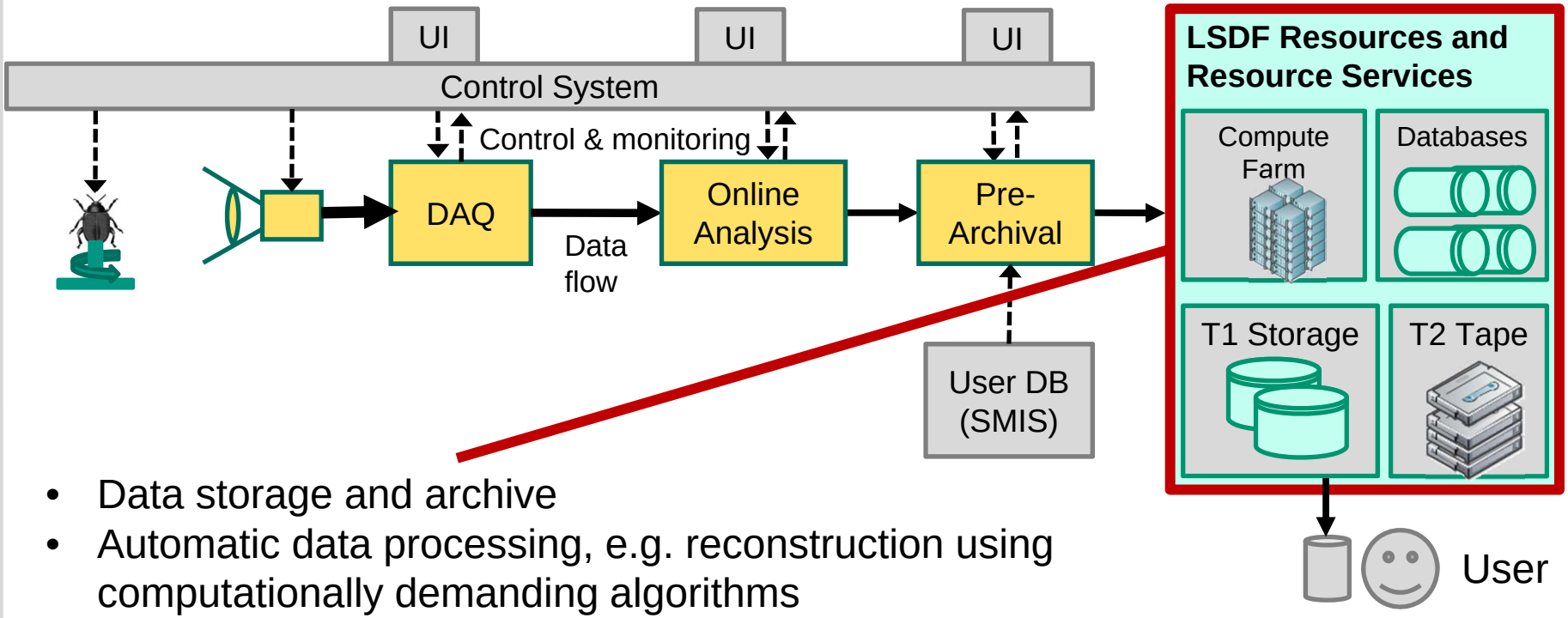
- GPU accelerated system
- Data pre-processing
- On-line reconstruction using filtered back-projection accelerated by GPUs (h→m)
- User selects promising data sets

Imaging Beamline Data Management



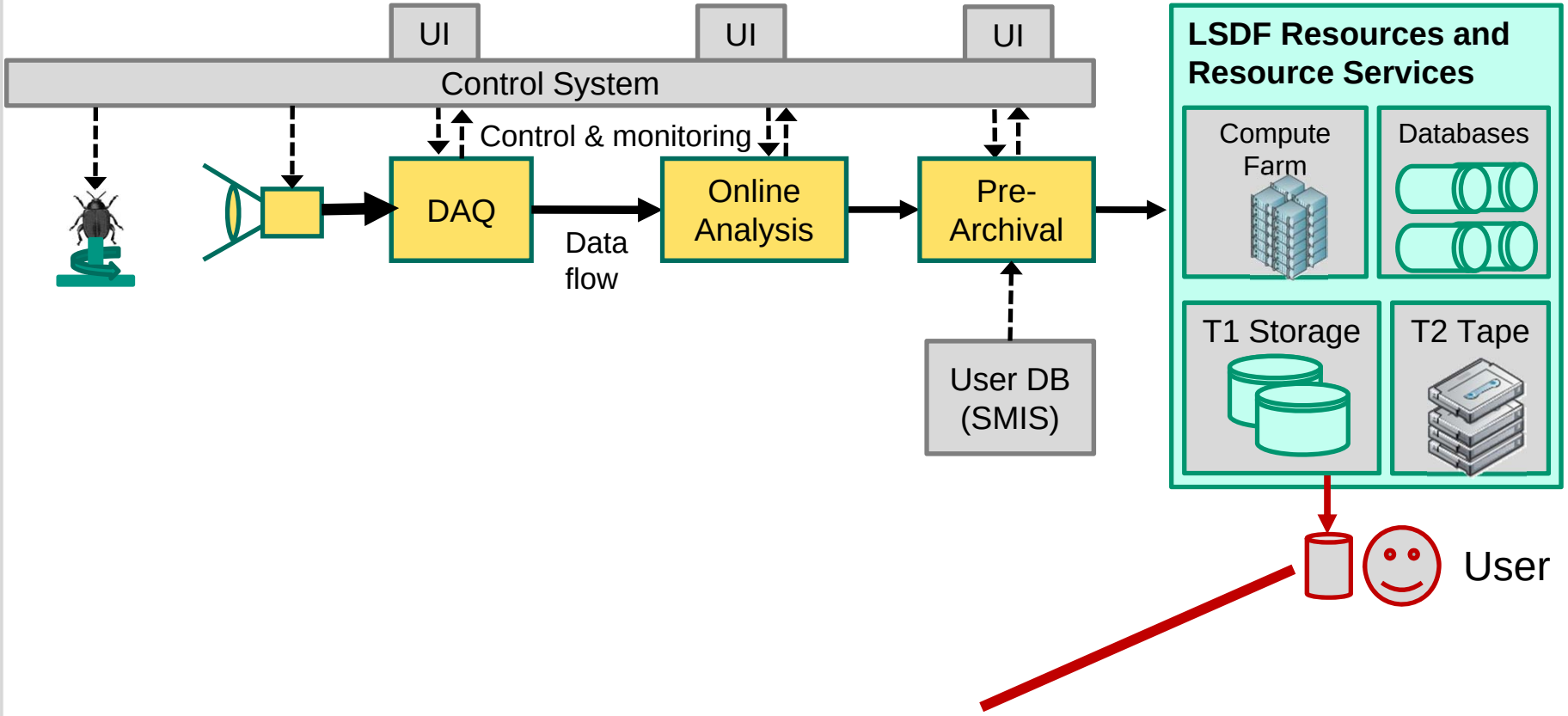
- Data preparation for storage, archiving & processing
- Meta data:
  - Life cycle
  - Access rights
  - Experiment description
  - Offline processing preparation
  - User information
  - Processing information
  - Data provenance
- Ingest

Imaging Beamline Data Management



- Data storage and archive
- Automatic data processing, e.g. reconstruction using computationally demanding algorithms
- Extending data sets by processing results

Imaging Beamline Data Management



- User access via DataBrowser or Web interface
- Temporary accounts for external users

Discussion

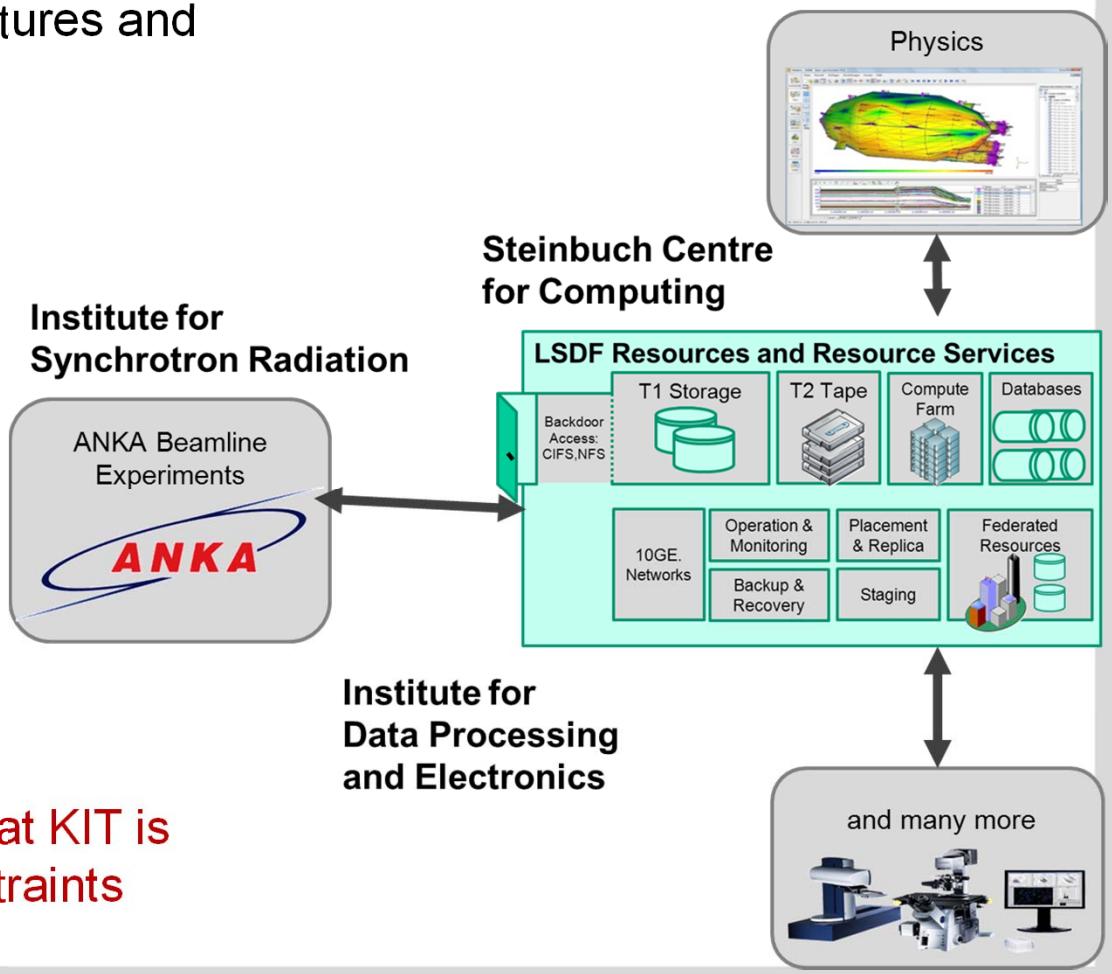
Imaging Beamline Data Management	
• DAQ	in production
• Online analysis	in production
• Pre-archival	in development
• Data format (NeXus)	in development
• Common Data Model tools (SOLEIL)	in preparation
Large Scale Data Facility	
• Data infrastructure and basic services	in production
• Data intensive computing	in production
• DataBrowser, execution framework	prototype, in production (biology)
• KIT Data Manager (KDM)	in development

Conclusion

KIT provides high level infrastructures and services

Large Scale Data Facility infrastructures:

- “Unlimited” resources
- Independent technology development
- Shared funding:
  - Research Programme SuperComputing
  - Communities



Cutting edge beamline research at KIT is not limited by infrastructure constraints