FIDIUM Overview

https://fidium.erumdatahub.de/

Status update of developments in the FIDIUM project

Kilian Schwarz
Analysis Facilities Workshop, Garching, June 18 2024

HELMHOLTZ

Table of contents

01 Introduction

02 Topic Area 1
- COBalD/TARDIS
- Usage of opportunistic resources
- AUDITOR

03 Topic Area 2
- Data lake monitoring
- Data management workflows
- Dynamic data caching
- Data lake prototypes

04 Topic Area 3
- Tests, optimisation, adjustments, support

05 FIDIUM prolongation
- TA1
- TA2
- TA3

06 summary and outview
- Summary
- Outview

FIDIUM introduction

FIDIUM Overview

Introduction

application

- Common application of 10 universities, 3 Helmholtz centres & CERN, 3 communities (KET, KHuK, KAT), submitted autumn 2020
- Funding period Q1 2021 – Q3, 2024

target

- Experiment overarching R&D in order to be ready to deal with the challenges of HL-LHC era
- 3 Topic Areas
  - TA1: including heterogeneous resources
  - TA2: federated data infrastructures
  - TA3: optimising and testing

Föderierte Digitale Infrastrukturen für die Erforschung von Universum und Materie (FIDIUM)

Gemeinsamer Antrag von Gruppen aus den Bereichen Elementarteilchenphysik, Hadronen- und Kernphysik und Astroteilchenphysik

- Rheinisch-Westfälische Technische Hochschule Aachen, Prof. Dr. Alexander Schmidt[1]
- Rheinische Friedrich-Wilhelms-Universität Bonn, PD Dr. Philip Bechtle
- Goethe Universität Frankfurt am Main, Prof. Dr. Volker Lindenstruth
- Albert-Ludwigs-Universität Freiburg, Prof. Dr. Markus Schumacher
- Georg-August-Universität Göttingen, Prof. Dr. Arnulf Quadt
- Universität Hamburg, Prof. Dr. Johannes Haller
- Karlsruher Institut für Technologie, Prof. Dr. Günter Quast
- Johannes Gutenberg-Universität Mainz, Prof. Dr. Frank Maas
- Ludwig-Maximilians-Universität München, Prof. Dr. Thomas Kuhr
- Bergische Universität Wuppertal, Prof. Dr. Christian Zeitnitz

Assoziierte Partner sind

- CERN, Dr. Markus Elsing
- DESY, Prof. Dr. Volker Gülzow
- GridKa, Dr. Andreas Petzold
- GSI Helmholtzzentrum für Schwerionenforschung, Dr. Kilian Schwarz[2]

[1] Sprecher des Verbundes
[2] Stellvertretender Sprecher des Verbundes

FIDIUM
Topic Area 1:
Tools for including heterogeneous resources

FIDIUM

Topic Area 1

WP 1: including opportunistic resources efficiently

KIT, Bonn, Frankfurt, Freiburg, Göttingen, GSI, DESY, GridKa

- Ongoing development and adjustment of resource manager COBalD/TARDIS
- Dynamic job scheduling
- Automatic scaling
- Usage of Tier-2 and Tier3 resources

WP2: accounting & controlling

KIT, Freiburg, Wuppertal, GridKa

- Tools for accounting of opportunistic resources
- Tools for efficiency monitoring

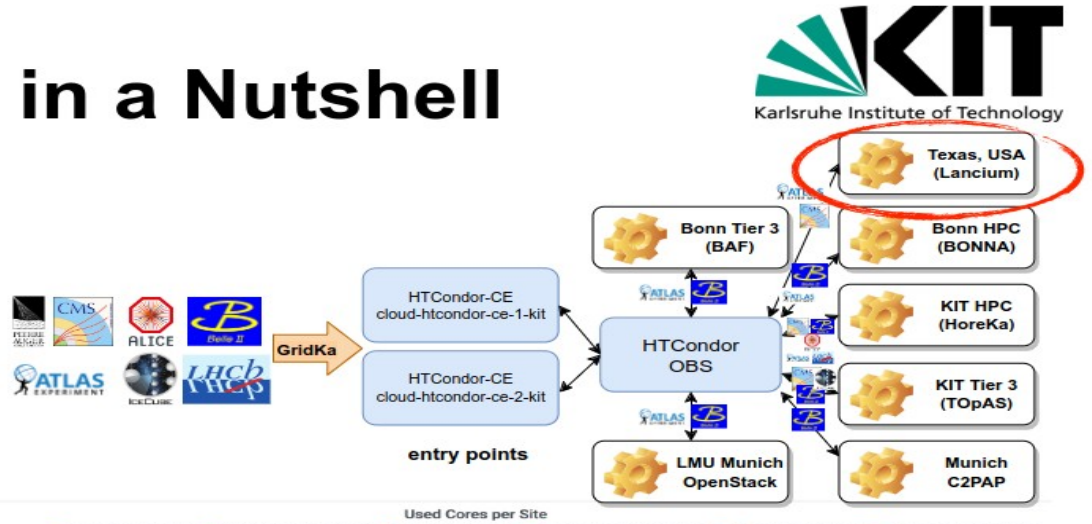
FIDIUM

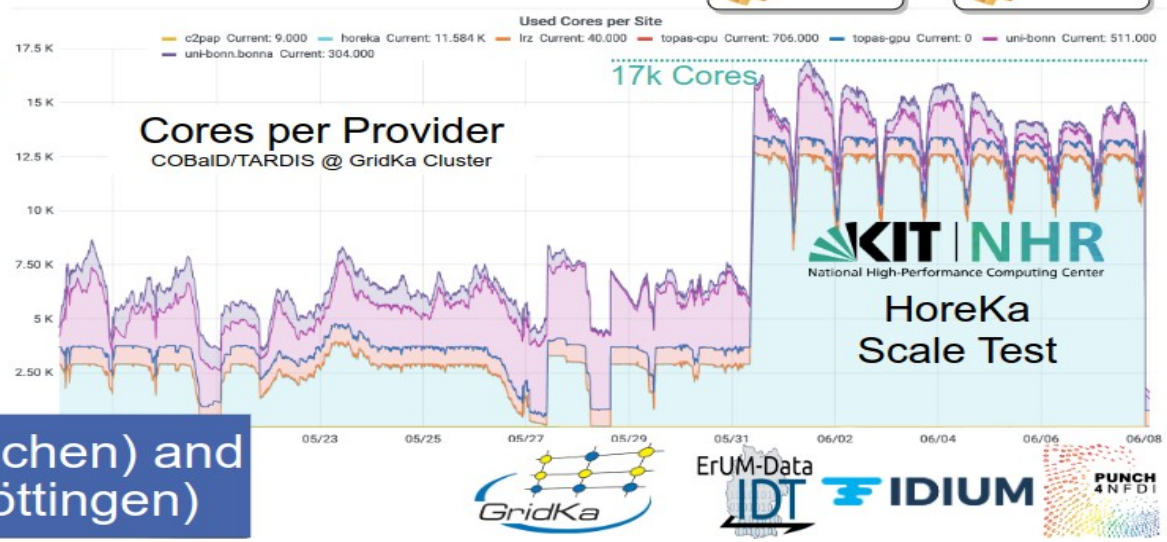
TA1/WP1/COBalD-TARDIS

Opportunistic Compute @ FIDIUM in a Nutshell

Simplify provisioning and utilization of third-party compute resources for the various communities:

- Dynamic, transparent and on-demand integration via COBalD/TARDIS (in-house development)
- Provide community-overarching unified entry points to a variety of resources (HPCs, Clouds, …)
- Demonstrated production scale operation during scale test together with HoreKa (KIT HPC cluster)
- Production deployment across HEP institutes & HPC resources coordinated by KIT/GridKa
- Central building block of the Compute4PUNCH infrastructure within PUNCH4NFDI





Similar setup deployed at CLAIX HPC (RWTH Aachen) and on-going deployment at Emmy (University of Göttingen)

FIDIUM

TA1/WP1/COBalD-TARDIS

The Entire COBalD/TARDIS Ecosystem

container-stacks

Container images to provide dedicated job environments

Available containers

Container	Environment provided
wlcg-wn	Provides a standard environment to run all jobs of VOs supported by WLCG
htcondor-wn	Provides a standard htcondor enabled workernode configurable using ansible

cobald Public

Cobald is an Opportunistic Balancing Deamon

Python 11 9

tardis Public

Transparent Adaptive Resource Dynamic Integration System

Python 15 17

H **HTCondor_configs**

Project ID: 3523 Leave project

236 Commits 3 Branches 0 Tags 278 KiB Project Storage

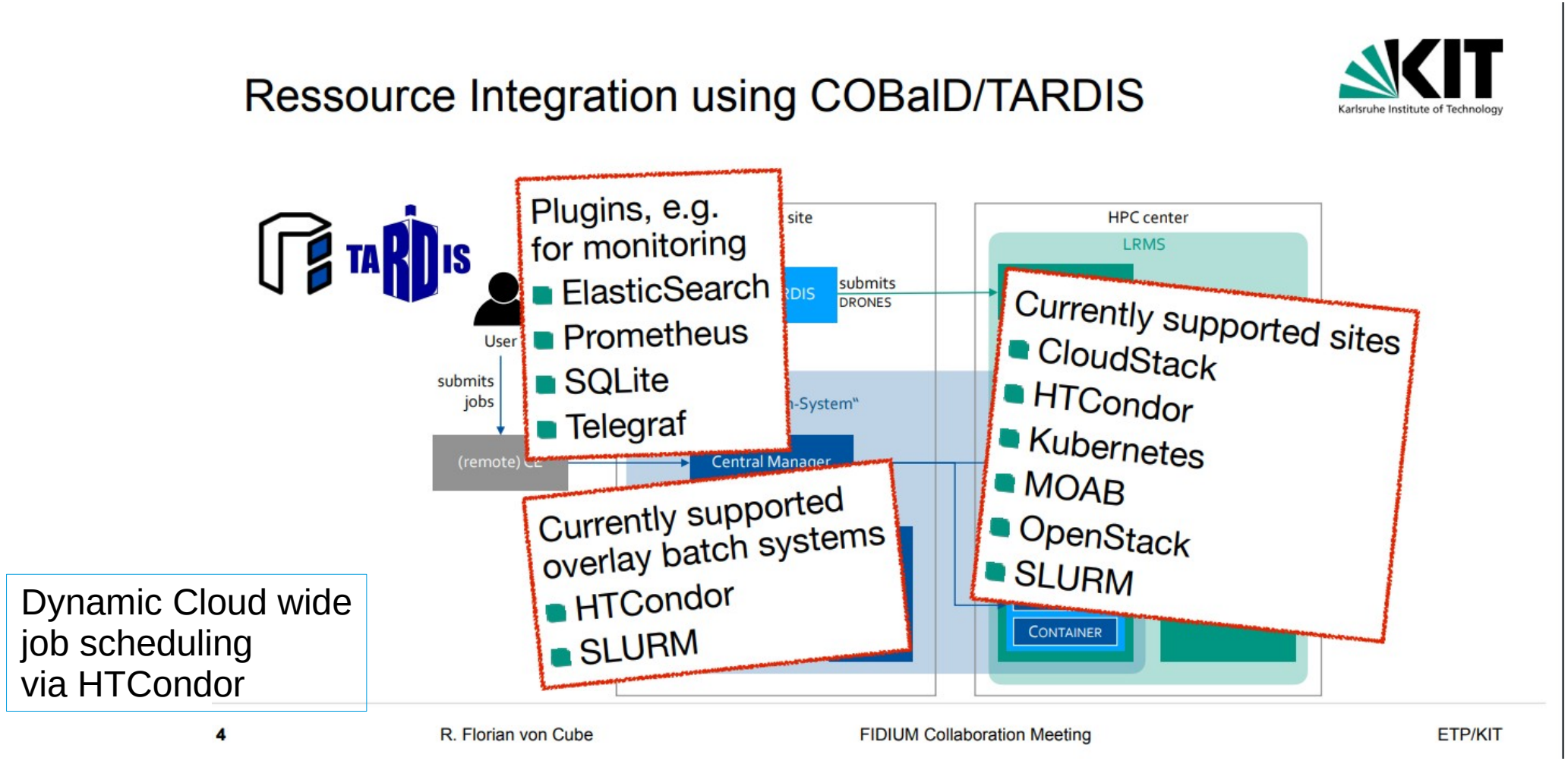
HTCondor configs for each site

condor-git-config 0.1.5

```
pip install condor-git-config
```

FIDIUM

TA1/WP1/COBalD-TARDIS



Dynamic Cloud wide job scheduling via HTCondor

FIDIUM

TA1/WP1/integration of opportunistic resources

Integration of HPC resources

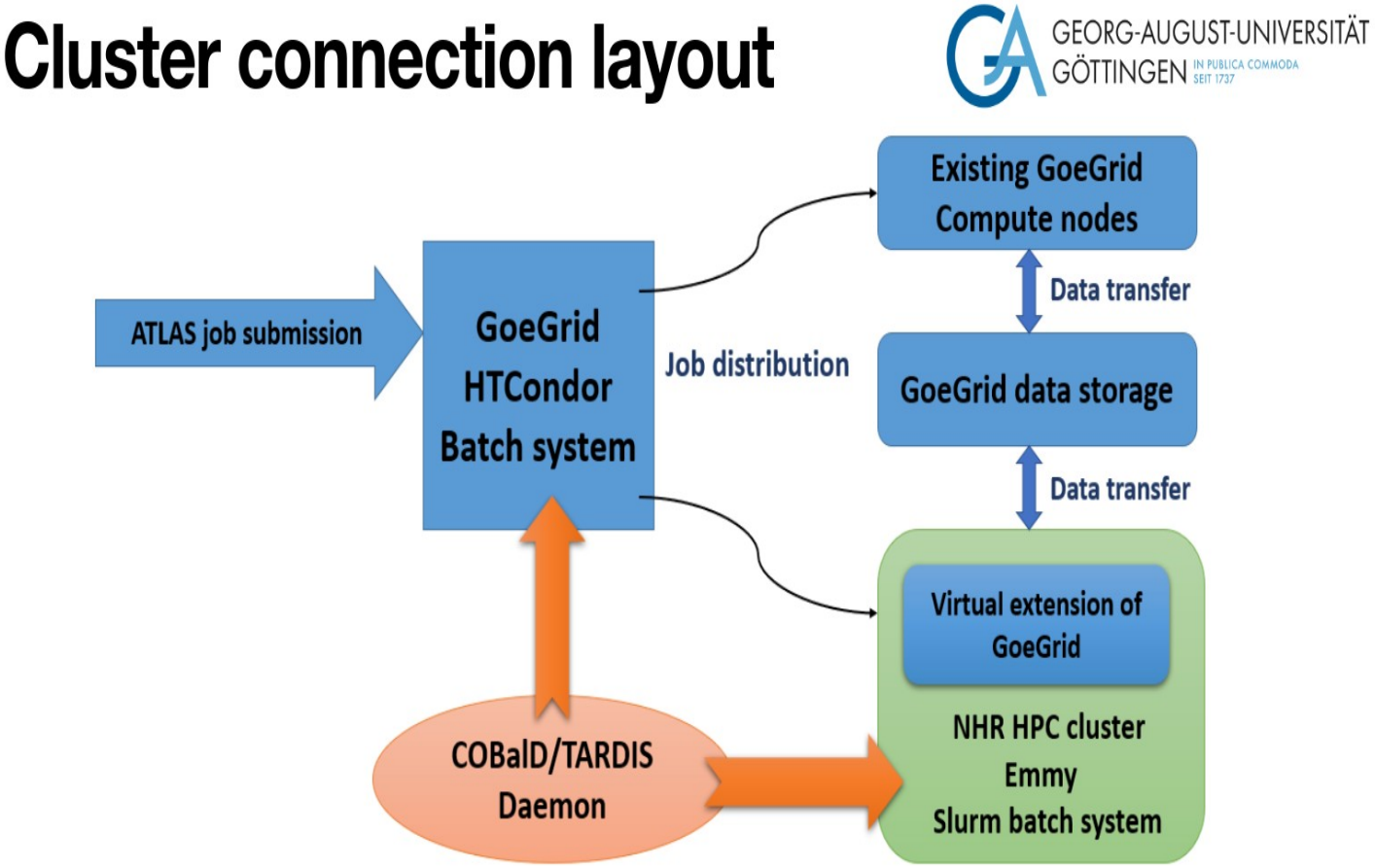
- e.g. integration of HLRN Emmy resources (HPC) into GoeGrid (HTC) via COBalD/TARDIS

Integration of desktop machines

- Integrating desktop machines using backfilling jobs via HTCondor extensions at Uni Bonn

Integration of GPU resources

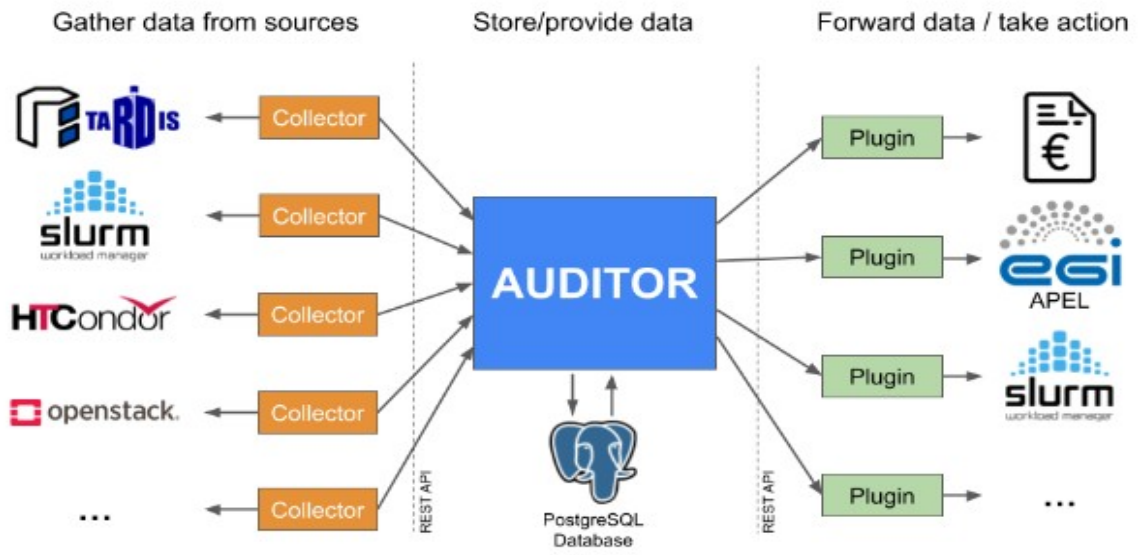
- Generalised vector and scalar version of CBM tracking, GPU based parallelisation, containerisation at U Frankfurt



- The solution is to virtually extend the GoeGrid batch system into Emmy using containers turning HPC nodes into virtual nodes with own job scheduling.

FIDIUM

TA1/WP2/accounting & monitoring



Modular accounting ecosystem

- **Collectors**
  - Accumulate data
- **Core component**
  - Accept data
  - Store data
  - Provide data
- **Plugins**
  - Take action based on stored data

Documentation and code

https://github.com/ALU-Schumacher/AUDITOR

AUDITOR: AccoUnting Data handlIng Toolbox for Opportunistic Resources

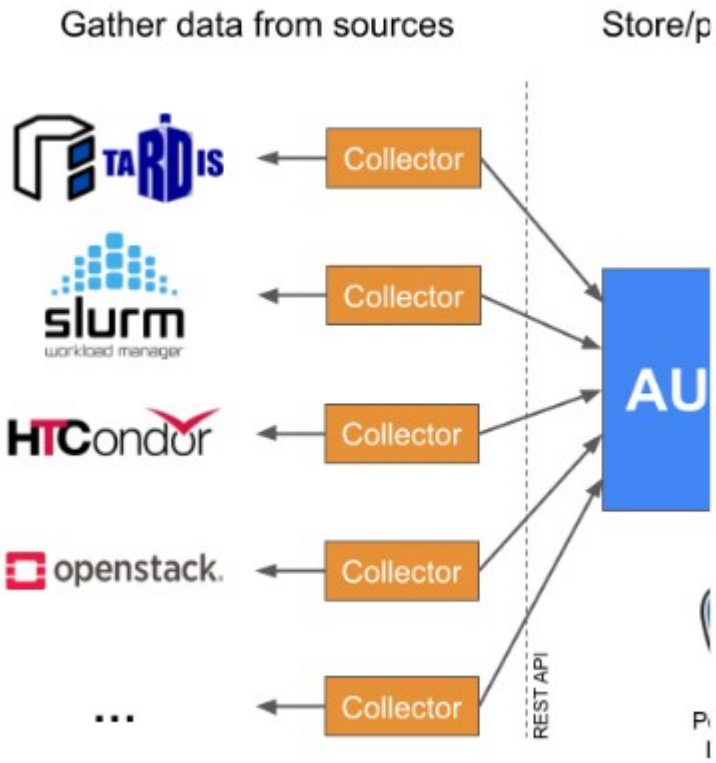
Presented at 26th International Conference on Computing in High Energy & Nuclear Physics (CHEP)
May 8-12, 2023, Norfolk, VA, USA: AUDITOR: Accounting for opportunistic resources

Michael Böhler | October 26th 2023 5 /11

FIDIUM

TA1/WP2/AUDITOR

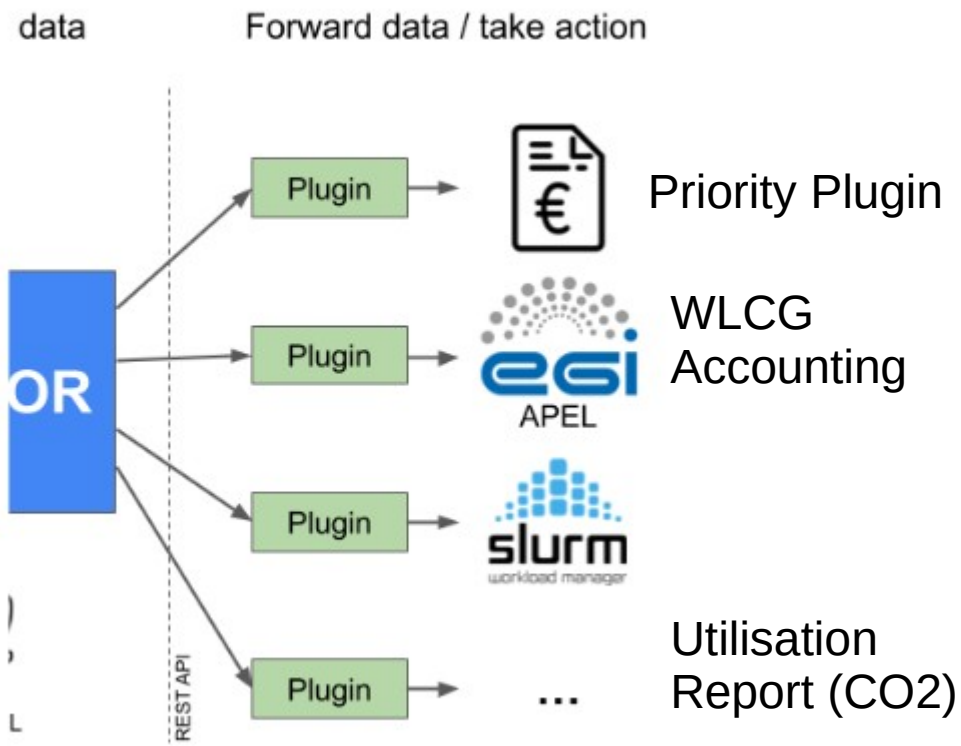
Data Collectors:



Core Component:

- access via REST
- data in PostgreSQL
- completely stateless

Data Forwarders :



FIDIUM
Topic Area 2:
Data Lakes, Distributed Data Caching

FIDIUM

Topic Area 2

WP1: monitoring

Wuppertal, GSI

- Load on data lake components
- Data access patterns

WP2: caching

KIT, Frankfurt, Mainz, Munich, Hamburg, GSI, DESY, GridKa

- Development of data caches
- Including data caches
- Parallel ad-hoc HPC file systems

WP3: workflows

KIT, Mainz, Hamburg, Göttingen, GSI, CERN, DESY, GridKa

- Replication and placement mechanisms
- Data management
- Efficient data access

WP4: prototypes

Frankfurt, Mainz, Munich, GSI, CERN, DESY, GridKa

- Data lake prototypes
- Including users, centres, data sources
- QoS

FIDIUM

TA2/WP1/monitoring

XrootD based monitoring system

- A test system has been setup by Wuppertal and GSI
- Close collaboration with Frankfurt, Mainz, and Munich foreseen

FIDIUM

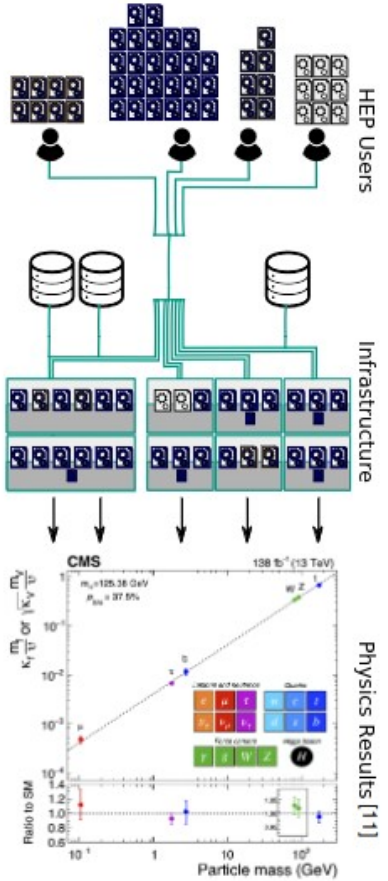
TA2/WP2/cache simulation

DCSim: Implementation of (HEP) Extensions and Simulator

- Code publicly available [8], presented at CHEP [9, 10]

1. Define workloads of jobs with:
   - Number of operations to execute
   - Memory
   - Input-datasets and sizes of outputs
   - Submission time
2. Define platform (network & hosts) with
   - properties: N_{core}, CPU-speed, RAM, disk, bandwidth
   - roles: worker, storage, data cache, scheduler, ...
3. Instantiate initial deployment of files on storage systems
4. Start the simulation!
   - Jobs are scheduled and run
   - Input-files are streamed and cached
   - Caches evict files if necessary

   → Job dynamics are monitored

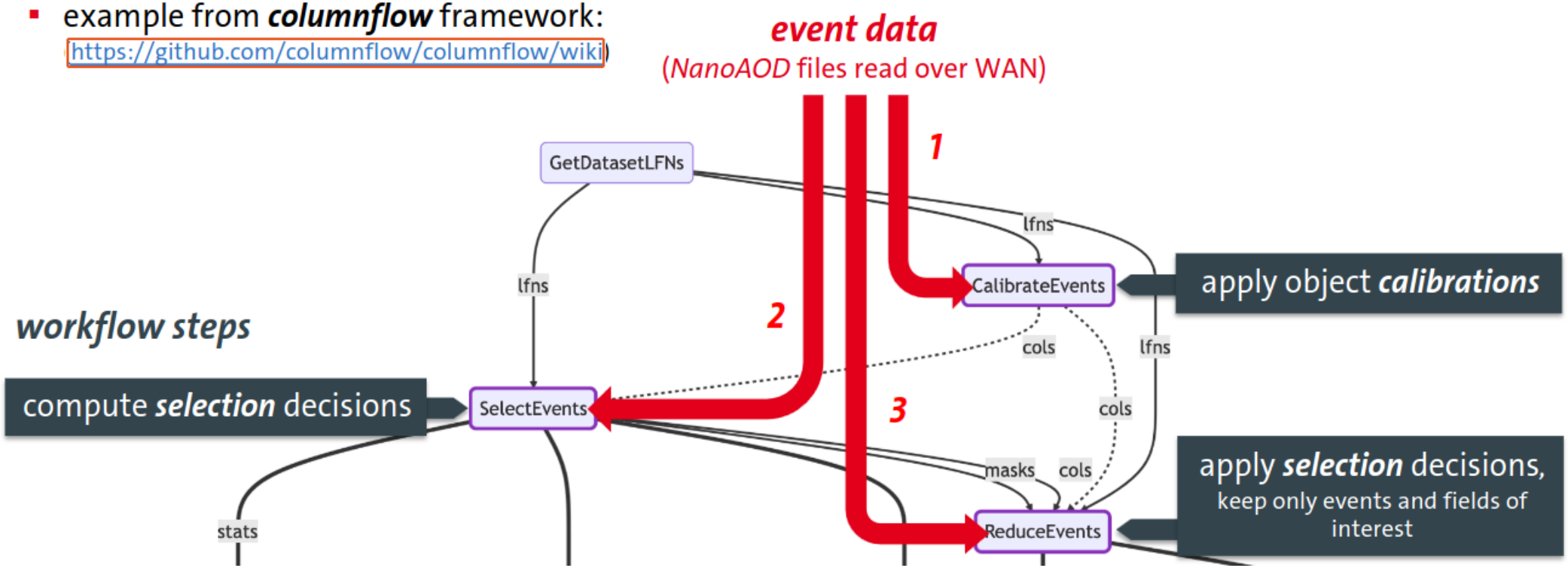


- Planning the future computing infrastructure for HEP
- Interplay between T1 und T2 with cache
- Validate tuned simulator with measured data

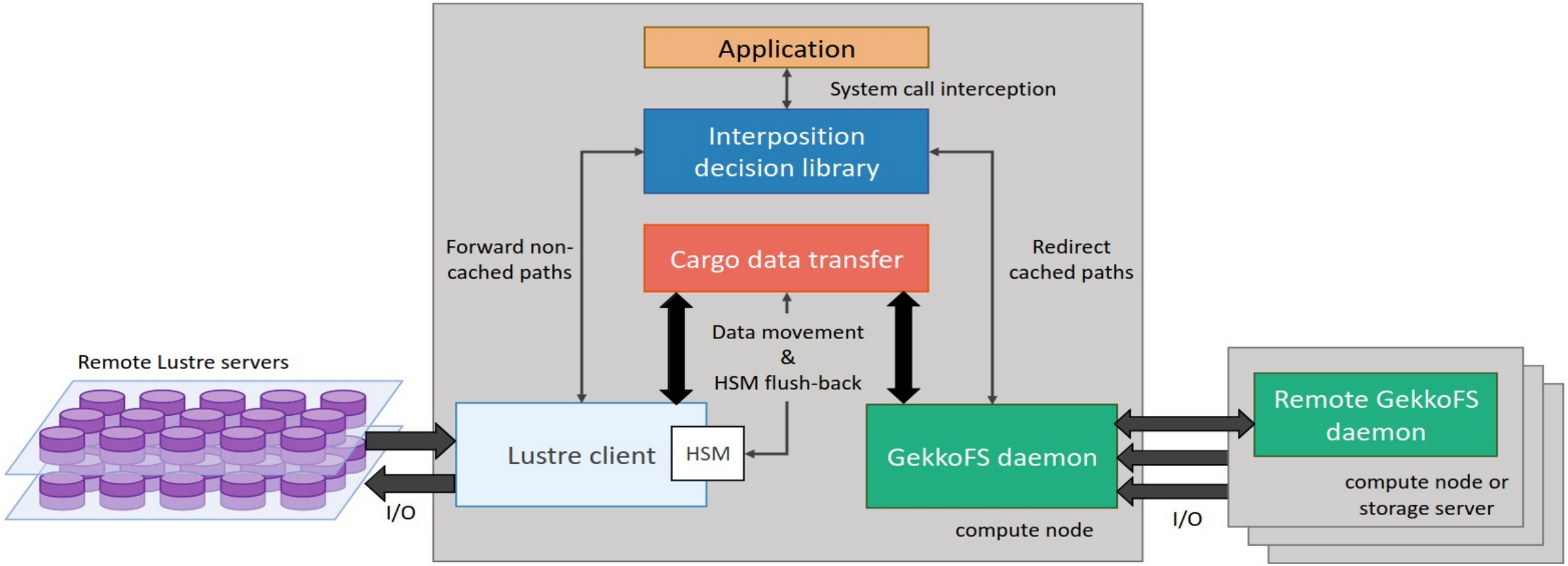
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Caching advantage for orchestrated workflows

- modern workflows typically orchestrated
  - workflow steps defined with high granularity, optimized for fast analysis turnaround
  - same files read at multiple stages in workflow → greatly boost performance through local caching on first read
- example from ***columnflow*** framework:
  https://github.com/columnflow/columnflow/wiki



Lustre client-side caching with ad-hoc file systems



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

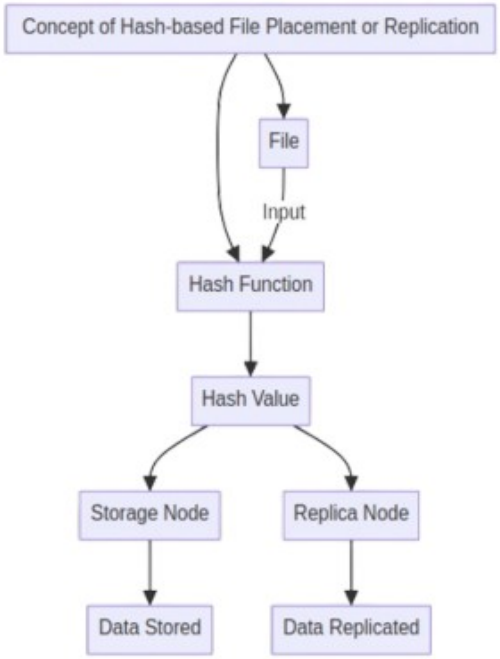
Work for hash-based file placement/replication

First focus on efficient integration of **dynamic caching via XRootD/XCache** into relevant workflows - utilizing high-bandwidth WAN between Goethe HLR and GSI

Now work for integration of **hash-based file placement/replication** mechanisms:

- Data files are hashed, generating a unique value determining their storage node in a HPC cluster
- Actual storage can be geographically spread
- **Hash value** can also identify replica nodes, ensuring data availability even in case of local failures

Geographically distributed, hash-based system → quick, reliable data access potentially accelerating workflows of data analysis in FIDIUM



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Tests for data lake prototypes

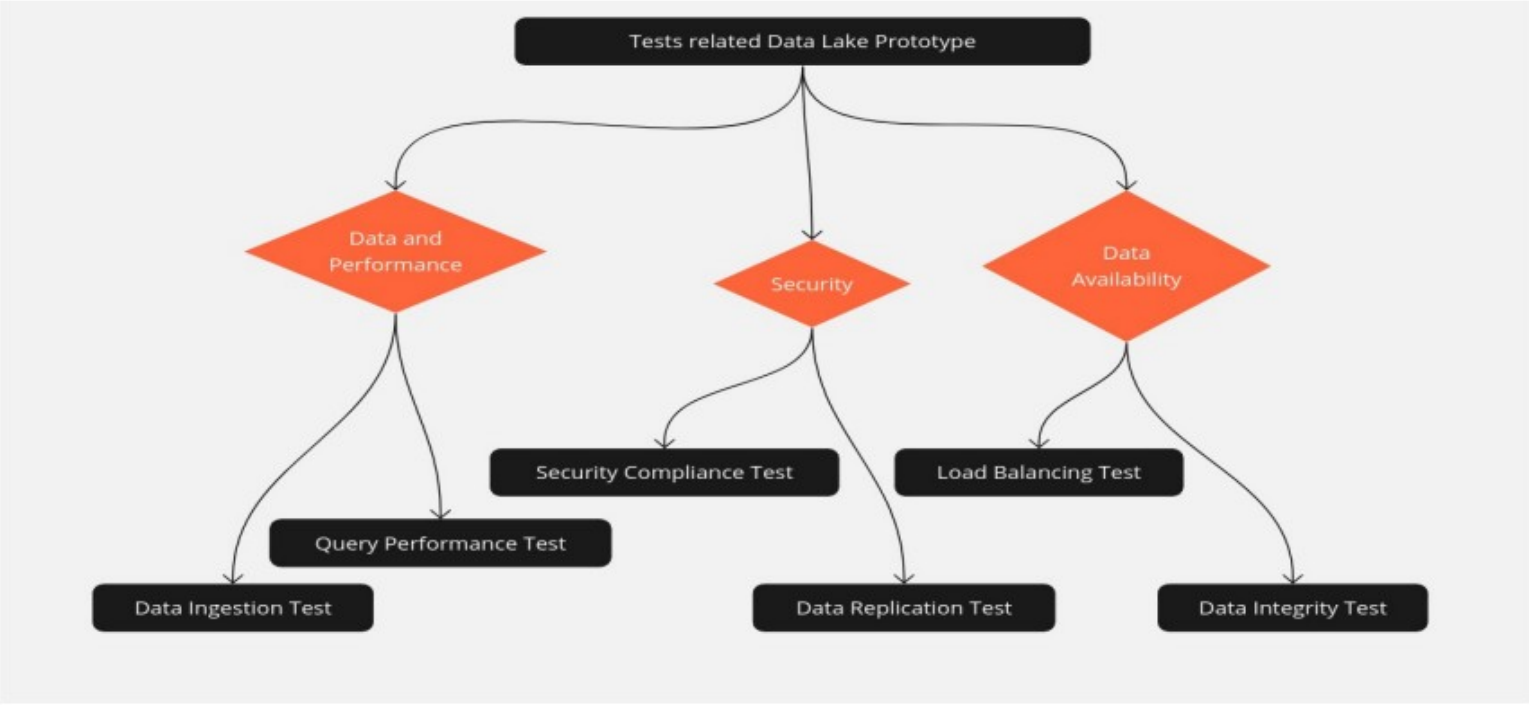
Testing all basic operations on prototypes:

oidc-token-based authentification

File placement/replication

Some tests also with davix file management

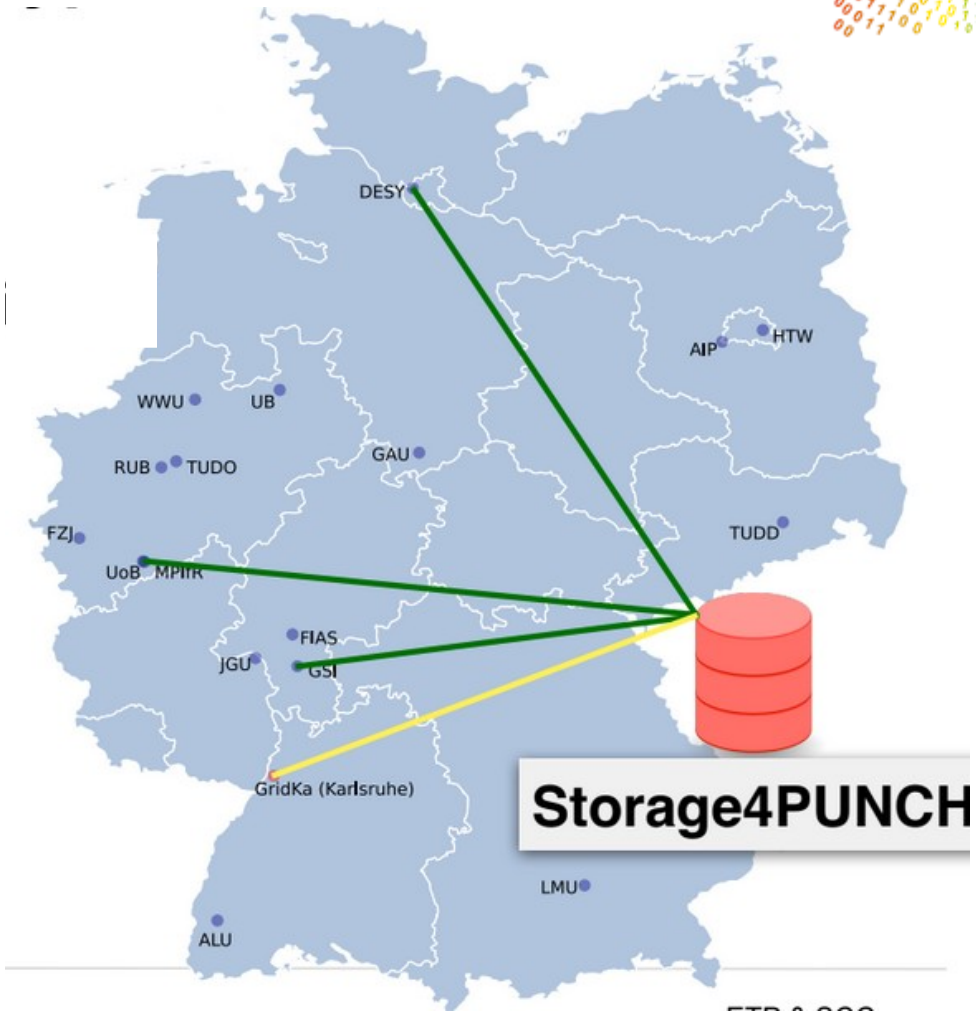
Final setup XRootD-based, coordination also with GSI group



FIDIUM

TA2/WP4/prototypes

- Common testbed together with PUNCH4NFDI planned
- Based on storage middleware XrootD (Bonn, GSI) and dCache (DESY, KIT)
- Token based access



FIDIUM
Topic Area 3: adjustment, tests and optimisation

FIDIUM

Topic Area 3

WP1: tests, optimisation, deployment

Aachen, KIT, Mainz, Wuppertal, Frankfurt, Munich, Freiburg, Hamburg, Göttingen, GSI, DESY, GridKa

- Integration of components
- Functional tests
- Integration in production environments
- deployment

WP2: adjustments

Aachen, KIT, Frankfurt, Munich, Hamburg, DESY, GridKa

- Optimisation for specific workflows
- Optimisation for parallel analysis

WP3: support

KIT, Freiburg, GSI, DESY, GridKa

- Site overarching support infrastructure

FIDIUM

Topic Area 3

WP1 & WP2 & WP3

- Freiburg: AUDITOR has been integrated into operations
- Freiburg: Test suite for AUDITOR use cases including support
- Hamburg and Wuppertal: integration of local clusters
- Göttingen: integration of NHR+HLRN cluster Emmy
- KIT: scaling test – intergration of remote and wind powered CPU cores in federated infrastructure
- FIDIUM solutions are being applied by DARWIN for federated infrastructure for Darkmatter use case
- Aachen: cache aware scheduling
- KIT: large scale usage of GPU systems for ML training at local TopAS cluster
- Munich: Analysis Grand Challenge analysis
- Mainz: Ad-hoc file system (GekkoFS) management scripts finalized for cluster integration
- Mainz: Testing of transparent staging between the parallel file system and GekkoFS during a compute job

FIDIUM
Prolongation
up to Q3/2025

FIDIUM prolongation

TA1

WP1: including opportunistic resources

- Göttingen: regular production on NHR HPC cluster, dynamic load balancer for HPC resources
- KIT: dynamic integration of GPU resources
- Munich: open data based ML applications on standard batch systems, integration of GPUs in parallel Dask analysis environment
- Bonn: energy charts based data, optimisation algorithm, evaluation of hardware cost/ energy consumption / energy cost
- Freiburg: finalisation and validation of AUDITOR Utlilisation Report Plugin
- Frankfurt: high performance optimisation for parallel tracking code, user friendly configuration

WP2: accounting

- Freiburg: AUDITOR Stresstest Collector Plugin, optimisation of collector data base schema

FIDIUM prolongation

TA2

WP1: monitoring

- Mainz and Wuppertal: adjusting of monitoring system to requirements of experiments

WP3: workflows

- Mainz: HSM optimisation, hash based methods for distributing data and metadata in data lakes, adjustment of Slurm for using ad-hoc resources
- Göttingen: preparing Rucio for data lake scenario
- Wuppertal and DESY: comparing Xcache and dCache, preparing easy installation methods, development of cache awareness in dCache and Xcache in connection with RUCIO

WP2: caching

- Mainz: client side caching for Lustre, performance measurements
- KIT: development of prototype of hash based Xcache for HPC
- Aachen: example implementation of HPC Xcache prototype
- Wuppertal: preparing dCache for use in federations
- DESY: preparing dCache for use as dynamic data cache at opportunistic resources

FIDIUM prolongation

TA3

WP1: tests, optimisation, deployment

- DESY, Hamburg: prototype dCache instance connecting DESY NAF and PhysNet Uni Hamburg
- DESY, Wuppertal: federated dCache instance between DESY and Wuppertal, performance measurement, tests for cache awareness of dCache and XCache
- Freiburg: optimisation of AUDITOR data base, adjusting AUDITOR to new requirements
- KIT: HTCondor Collector for AUDITOR
- Aachen: production use of Aachen NHR centre for CMS
- Mainz: connecting Mainz NHR centre to data lake

WP2: adjustments

- DESY, Hamburg: COBalD/TARDIS between DESY NAF and PhysNet Uni Hamburg, performance measurements
- Munich: comparing data formats on SSD and HDD, scaling of parallel application when accessing data on remote servers
- Frankfurt: energy consumption of parallel tracking code on multi core CPUs and GPUs
- Aachen: real time simulation of dynamic energy supply for steering local cloud resources, resource usage in real time simulation

WP3: support

- Working support, tutorial for including NHR resources, Slurm interface of AUDITOR plugin

FIDIUM
Summary and outview

FIDIUM

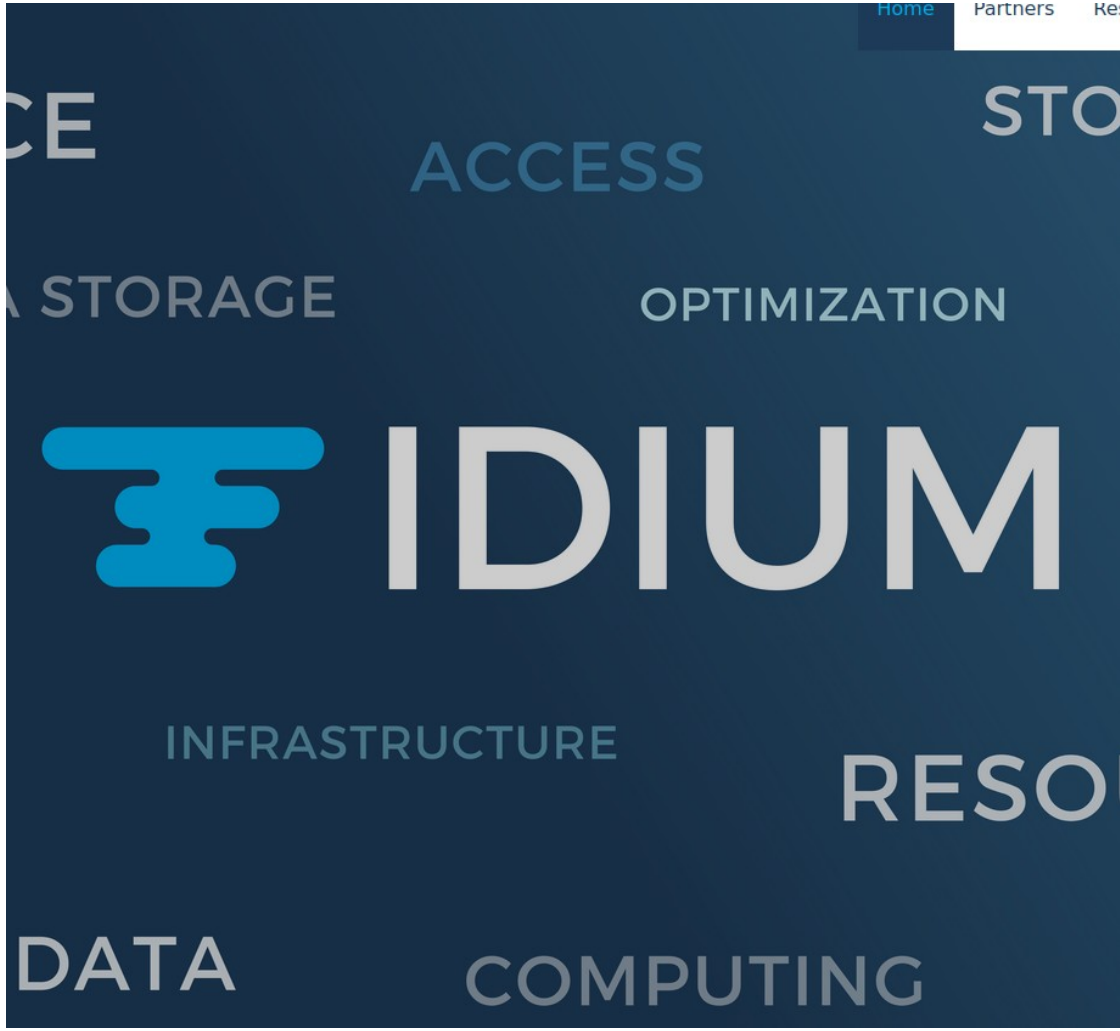
Summary

Consortium

- Common application of 10 universities, 3 Helmholtz centres & CERN, 3 communities (KET, KHuK, KAT), submitted autumn 2020
- Funding period Q1 2021 – Q3, 2024, prolongated up to Q3 2025
- See https://fidium.erumdatahub.de/

Mission

- Successfully adressing the mission to develop experiment overarching software so that the experiments are ready to face the challenges of HL-LHC era
- Important software tools for distributed computing, federated storage, monitoring, and accounting are being provided by FIDIUM



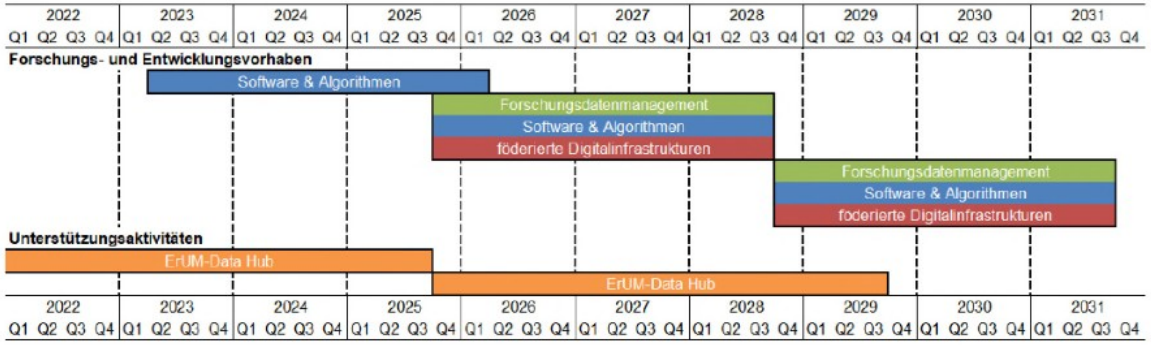
FIDIUM

Outview

ErUM-Data

- We will most probably have 3 follow up projects in the upcoming ErUM-Data call
  - Sustainable Federated IT Infrastructures for ErUM
  - Federated Storage Infrastructures for ErUM
  - Analysis Facilities (this workshop)

ErUM-Data: Timeframe and implementation



- Prisma strategy meeting on January 23rd/24th, 2024 in Hamburg

Thank you

Contact

Deutsches Elektronen-
Synchrotron DESY

www.desy.de

Kilian Schwarz
IT/Scientific Computing
Kilian.schwarz@desy.de
040 8998 2596