T 17: Data, AI, Computing 2 (analysis tools)

Time: Monday 16:00–18:00 Location: Geb. 30.34: LTI

T 17.1 Mon 16:00 Geb. 30.34: LTI

Resource-aware Research on Universe and Matter: Call-to-Action in Digital Transformation — •Kilian Schwarz — Deutsches Elektronen-Synchrotron (DESY), Notkestraße 85, 22607 Hamburg

Given the urgency to reduce fossil fuel energy production to make climate tipping points less likely, we call for resource-aware knowledge gain in the research areas on Universe and Matter with emphasis on the digital transformation. A portfolio of measures is described in detail and then summarized according to the timescales required for their implementation. The measures will both contribute to sustainable research and accelerate scientific progress through increased awareness of resource usage. This work is based on a three-days workshop on sustainability in digital transformation held in May 2023.

T 17.2 Mon 16:15 Geb. 30.34: LTI

PUNCH4NFDI und XRootD für Storage4PUNCH: Towards a National Federated Computing Infrastructure — Philip Bechtle, Oliver Freyermuth, Michael Hübner, •Simon Thiele, and Luka Vomberg — Physikalisches Institut, Universität Bonn

The NFDI (**N**ationale **F**orschungs**D**aten **I**nfrastruktur, engl. National Research Data Infrastructure) has the aim to systematically connect valuable data from science and research and make it available. PUNCH (**P**articles, **U**niverse, **NuC**lei and **H**adrons) is a consortium in the NFDI, dealing with the titular fields of particle, astro, nuclear and hadron physics. PUNCH4NFDI is organised into several task areas focusing on data management and transformation, the access to computing resources, challenges related to data irreversibility, as well as education and outreach. One important aspect of data management is the transfer of data in a federated storage and computing infrastructure and allowing for the preservation of access rights between individuals and groups.

XRootD is a software framework to manage the access to data and a prime candidate for fulfilling the above role. It consists of software with a scaleable architecture and a communication protocol. Several plug-ins are available to extend the functionality for various specific purposes, such as the use of tokens for authorization.

In this talk I will firstly introduce the PUNCH4NFDI project, then go more into detail on the goals of task area for data management and present my work on the use and modification of XRootD for PUNCH to fulfill all requirements needed for the federated storage infrastructure.

T 17.3 Mon 16:30 Geb. 30.34: LTI

Opportunities for Open Science through Artificial Intelligence — •Jutta Schnabel for the ANTARES-KM3NET-ERLANGEN-Collaboration — ECAP, FAU Erlangen-Nürnberg

The KM3NeT collaboration is currently constructing neutrino detectors in the Mediterranean Sea, with petabytes of high-quality scientific data to be analysed and shared in the years to come. Targeting both particle and astroparticle research, the opportunities for cooperative data analysis are manifold. The KM3NeT collaboration has implemented an Open Science program to coordinate the efforts in providing FAIR data and scientific workflows in accordance with the developing standards in the respective communities. In the years to come, employing artificial intelligence in discovery, integration and accessibility of scientific data will gain importance. In this contribution, the consequences of these developments will be explored in the context of the KM3NeT Open Science program.

T 17.4 Mon 16:45 Geb. 30.34: LTI

HELIPORT: An overarching Data Management System at HZDR — •Stefan E. Müller[1], Thomas Gruber[1], Oliver Knodel[1], Jeffrey Kelling[1], Mani Lokamani[1], David Pape[1], Martin Voigt[1,2], and Guido Juckeland[1] — [1]Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany — [2]Technische Universität Dresden, Dresden, Germany

Researchers at the Helmholtz-Zentrum Dresden-Rossendorf rely on a variety of systems and tools when it comes to administer their research data. Processes involving research data management include the project planning phase (proposal submission to the beamtime proposal management system, the creation of data management plans and data policies), the documentation during the experiment or simulation campaign (electronic laboratory notebooks, wiki pages), backup- and archival systems and the final journal and data publications (collaborative authoring tools, meta-data catalogs, software and data repositories, publication systems). In addition, modern research projects are often required to interact with a variety of software stacks and workflow management systems to allow reproducibility on the underlying IT infrastructure. The "HELmholtz ScIentific Project WORkflow PlaTform" (HELIPORT), which is currently developed by researchers at HZDR and their collaborators, tries to facilitate the management of research data and metadata by providing an overarching guidance system which combines all the information by interfacing the underlying processes and even includes a workflow engine which can be used to automate processes like data analysis or data retrieval.

T 17.5 Mon 17:00 Geb. 30.34: LTI

dCache: Inter-disciplinary storage system — •Tigran Mkrtchyan[1], Krishnaveni Chitrapu[3], Dmitry Litvintsev[2], Svenja Meyer[1], Lea Morschel[1], Paul Millar[1], Marina Sahakyan[1], and Kilian Schwarz[1] — [1]Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany — [2]Fermi National Accelerator Laboratory (FNAL), Batavia, USA — [3]National Supercomputer Center, Linköping University, Sweden

The dCache project provides open-source software deployed internationally to satisfy ever more demanding storage requirements. Its multifaceted approach provides an integrated way of supporting different use-cases with the same storage, from high throughput data ingest, data sharing over wide area networks, efficient access from HPC clusters and long-term data persistence on tertiary storage. Though it was originally developed for the HEP experiments, today it is used by various scientific communities, including astrophysics, biomed, life science, which have their specific requirements. With this contribution, we would like to highlight the recent developments in the dCache regarding integration with CERN Tape Archive (CTA), advanced metadata handling, bulk API for QoS transitions, RESTAPI to control interaction with tape system and the future development directions.

T 17.6 Mon 17:15 Geb. 30.34: LTI

Metadata curation efforts at Kascade Cosmic-Ray Data Centre — •Victoria Tokareva, Andreas Haungs, Doris Wochele, Jürgen Wochele, and Donghwa Kang — Karlsruhe Institute of Technology, Institute for Astroparticle Physics, 76021 Karlsruhe, Germany

Metadata curation plays a pivotal role in advancing the machine-actionability of data, thereby facilitating the fulfillment of FAIR data principles. The KASCADE Cosmic-ray Data Centre (KCDC) has served as a repository for data generated by the high-energy astroparticle physics experiment KASCADE, along with several other research projects, since 2013. Additionally, it functions as an information platform in the field of high-energy astroparticle physics for both the astroparticle community and the general public.

The platform provides users with a diverse set of digital objects. While some of them possess rich sets of metadata (including persistent identifiers such as DOIs and ISNIs), and allow machine access via REST API, efforts to integrate with the broader PUNCH4NFDI Data Portal underscore the necessity of a unified and comprehensive approach to curate a wider range of granted resources. The contribution examines KCDC's digital resources as Digital Research Products and covers recent developments, with a focus on integrating our curation practices into data-intensive physics communities' endeavors and the national data infrastructure landscape.

This work is partially supported by the DFG fund "NFDI 39/1" for the PUNCH4NFDI consortium.

T 17.7 Mon 17:30 Geb. 30.34: LTI

AUDITOR: Accounting for opportunistic resources — •Raghuvar Vijayakumar, Michael Böhler, Benjamin Rottler, Dirk Sammel, and Markus Schumacher — Universität Freiburg

In response to the increasing demand for computing resources in High Energy Physics (HEP), we have integrated under-utilized computing resources using COBalD/TARDIS in an opportunistic manner. However, resource sharing requires robust accounting. We present AUDITOR (AccoUnting DatahandlIng Toolbox for Opportunistic Re-

sources), a flexible and extensible ecosystem tailored to address a broad range of use cases and infrastructures. AUDITOR employs specialised collectors that monitor, capture and record accounting data, subsequently stored in a database. The recorded data are then made accessible to plugins, which analyse the accounting information to perform specific tasks, such as computing the CO2 footprint and forwarding the accounting data to other accounting systems. AUDITOR is written in Rust, a modern programming language optimised for performance, type safety and concurrency. The python client pyauditor allows to utilize all AUDITOR functionalities via a python interface. In this talk, we present how AUDITOR works and how Pyauditor can be used to query and analyze large accounting data sets.

T 17.8 Mon 17:45 Geb. 30.34: LTI

Workflow Optimization for HEP Jobs on Opportunistic Resources with XRootD — •Robin Hofsaess, Manuel Giffels, Günter Quast, Maximilian Horzela, and Matthias Schnepf — Karlsruher Institut für Technologie

The WorldWide LHC Computing Grid (WLCG) is tailored to support High Energy Physics (HEP) workflows, offering an optimal environment for job execution. However, alongside the mostly homogeneous WLCG infrastructure, we increasingly rely on diverse 'opportunistic resources' such as High-Performance Computing (HPC) centers and cloud service providers. These resources are more heterogeneous and not inherently suited for HEP workflows that for example might require a high throughput of data. Since the future German HEP computing strategy foresees an increased share on such resources, we are currently working on concepts that not only aim to maximize computational throughput, but also align with the broader goals of sustainable resource utilization in an era of exponential data growth.

In this talk, I will present a first Proof of Concept implementation using an XRootD cache to increase the efficiency of HEP jobs on the HoreKa HPC cluster in Karlsruhe.