CS3 2024 - Cloud Storage Synchronization and Sharing

Monday, 11 March 2024 - Wednesday, 13 March 2024

CERN

Book of Abstracts

Contents

Keynote / 159

Neutrality, Impartiality, and Independence of Humanitarian Action in the Digital Age at ICRC

Corresponding Authors: mmarelli@icrc.org, mvignati@icrc.org

Challenges and opportunities for Neutral Impartial and Independent Humanitarian Action in a world increasingly characterized by pervasive connectivity and connectivity denials, digital services, adverse cyber operations, debates around digital sovereignty, and global competition for digital infrastructure.

Services and Infrastructures / 154

CERNBox turns 10 years old: pitfalls and challenges of building CERN's cloud collaborative platform

Author: Hugo Gonzalez Labrador[1]

[1] *CERN*

Corresponding Author: hugo.gonzalez.labrador@cern.ch

CERNBox is CERN's flagship cloud collaborative storage platform and it has turned 10 years old. In this talk we'll delve into the main pillars that made this platform a success and the pitfalls and lessons learned of running a multi-petabyte platform satisfying the heterogeneous needs of thousands of scientists.

Services and Infrastructures / 162

Working with sensitive research data across borders and institutions

Author: Anne Bergsaker[1]

[1] *University of Oslo*

Corresponding Author: a.s.bergsaker@usit.uio.no

Working with sensitive research data is a challenge, along all steps of the research lifecycle. At the University of Oslo, we are developing and integrating solutions for data capture, storage, analysis and dissemination, to make it a little easier for researchers to work safely with sensitive data. The collection of services are what make up our internally developed research platform, called Educloud research. The goal is to make the platform with all the solutions it contains user friendly and self service based, while still maintaining high levels of security. The platform itself provides flexibility and allows for collaboration, both with colleagues working at the university, but also with external collaborators situated across the globe.

Services and Infrastructures / 139

Towards data sharing service for Physical Sciences Data Infrastructure

Authors: Jonathan Bathe[1]; Vasily Bunakov[None]

[1] *Scientific Technology Facility Council*

Corresponding Authors: jonathan.bathe@stfc.ac.uk, vasily.bunakov@stfc.ac.uk

PSDI (Physical Sciences Data Infrastructure) https://www.psdi.ac.uk/ is a part of Digital Research Infrastructures programme in the UK, aiming to accelerate research in physical sciences by connecting various data and computation systems researchers currently use. The need of a data transfer and data sharing service was identified in the early stages of PSDI, followed by a service design exercise and technology trials of EFSS solutions. We are going to present the current state of this effort and the expected directions for the development of data sharing service, collecting the audience feedback and opinions.

Services and Infrastructures / **147**

EFSS on a truly grand scale: Experience with the ByCS school cloud in Bavaria, Germany

Author: Nick Wilson[1]

Co-author: David Walter [1]

[1] *ownCloud*

Corresponding Authors: dwalter@owncloud.com, nwilson@owncloud.com

ByCS Drive is a cloud service based on ownCloud Infinite Scale built for up to 5 million users in 6300 schools.

This talk will give a look at the architecture, focusing on storage, data security and scalability to assure performance under extreme peaks and growing data requirements. A short demo will showcase special compliance measures, the new user interface, and integration with other software popular in education.

EFSS Products / **128**

Nextcloud. State of the nation

Author: Frank Karlitschek[None]

Corresponding Author: karlitschek@gmail.com

This talk will give an overview of the Nextcloud developments and improvements in the last 12 month. Several noteworthy things happened in the last Nextcloud releases. From architectural improvements to changes on APIs and the sync engine, to usebility and functionality. This Talk will give a full overview.

EFSS Products / **135**

What's new in Seafile for 2023

Author: Jonathan Xu[None]

Corresponding Author: xjqkilling@gmail.com

Seafile is a popular open source file syncing and sharing solution. Its features include robust and efficient file syncing, cross-platform virtual drive clients, efficient usage of server resources, and encrypted libraries. It is used by many European educational institutions, such as HU Berlin, the Max Planck Digital Library, and PSNC.

In 2023, we expanded Seafile's feature set into some cutting-edge areas, aiming to provide more value to our users. Directions include:
* SeaDoc –a collaborative document server
* Semantic search –improved search accuracy with AI
* UI modernization

We will share the outcomes of our 2023 work and what we are heading towards in the future.

EFSS Products / **149**

ownCloud update and roadmap for the CS3 Community

Authors: Holger Dyroff[None]; Holger Dyroff[None]

Co-author: Reinhard Schüller

Corresponding Authors: hd@owncloud.com, hdyroff@owncloud.com

How ownCloud assures the focus on Higher Education and Science/Research to drive the product development.

Services and Infrastructures / **174**

Community Site Report – Summary

Corresponding Author: steiger@id.ethz.ch

Services and Infrastructures / **143**

The S3 Object Storage Service on INFN Cloud

Authors: Ahmad Alkhansa[1]; Alessandro Costantini[2]; DIEGO MICHELOTTO[3]; Daniele Spiga[4]; Diego Ciangottini[5]; Federico Fornari[None]; Giada Malatesta[None]; Jacopo Gasparetto[6]; Massimo Sgaravatto[7]; Stefano Stalio[None]

[1] *INFN - CNAF*

[2] *INFN-CNAF*

[3] *INFN - National Institute for Nuclear Physics*

[4] *Universita e INFN, Perugia (IT)*

[5] *INFN, Perugia (IT)*

[6] *CNAF*

[7] *Universita e INFN, Padova (IT)*

Corresponding Author: alessandro.costantini@cnaf.infn.it

Backed by the 20 years of successful development and operation of the largest Italian research e-infrastructure through the Grid, the Italian National Institute for Nuclear Physics (INFN) has been running for the past three years INFN Cloud, a production-level, integrated and comprehensive cloud-based set of solutions, delivered through distributed and federated infrastructures.

INFN Cloud provides a large and customizable set of services, ranging from simple IaaS to specialized SaaS solutions, centered through a PaaS layer built upon flexible authentication and authorization services, offered via INDIGO-IAM, and optimized resources and services orchestration.

Since its beginning, INFN Cloud has offered its users and collaborations an S3-based Object Storage service for data archiving.Such S3 buckets can be accessed via a web ui or programmatically. They can also be mounted as volumes in a semi-posix fashion, providing a sort of "geographic home directory", which can be accessed remotely e.g. on Jupyter Notebooks.

Key features of the S3 service are (i) OIDC authentication via the Indigo DataCloud IAM service, (ii) the use of Open Policy Agent for fine grained authorization, (iii) full integration with other INFN Cloud services, and (iv) data replication over two data centers 400km away.

Recently, the service has been migrated from a Minio Gateway on top of a distributed OpenStack Swift cluster to a multisite CEPH RGW infrastructure supported by a web user interface that has been developed in house.

We will describe the main features of our setup, focusing on the authentication/authorization model and the most prominent use cases, comparing the pros and cons of the new solution we adopted in respect to the one we abandoned.

Services and Infrastructures / **175**

Publication of open research data with sync&share storage

Authors: Andreas la Roi[1]; Madeleine Fritschi[1]

Co-authors: Gianluca Caratsch [1]; Tilo Uwe Steiger

[1] *ETH Zürich*

Corresponding Authors: andreas.laroi@library.ethz.ch, gianluca.caratsch@id.ethz.ch

The publication of open research data (ORD) is becoming an integral part of publicly funded research. Scientific publishers, funders and research institutions often require researchers to publish the data and code that is relevant for their articles. This data needs to be citable and published in a repository that follows the FAIR principles. For scientific domains that generate large datasets institutional repositories reach their limits.
This presentation introduces the service Libdrive, an extension of the ETH Research Collection, ETH Zurich's institutional repository. This service as a use case of sync&share storage allows researchers to publish datasets of several terabytes in accordance with the ORD guidelines of the various stakeholders.

We provide an overview of the service, it's infrastructure and how it's integrated with the ETH Research Collection and central storage services. Subsequently we show the trends in its use and the challenges we've encountered in implementing this service.

Services and Infrastructures / **140**

Sunet Drive - Sweden - Community Site Report

Authors: Magnus Andersson[1]; Micke Nordin[1]; Richard Freitag[None]

[1] *SUNET*

Corresponding Authors: mandersson@sunet.se, freitag@sunet.se

Sunet Drive is Sweden's national data storage solution, and part of the ScienceMesh. It is a federated solution consisting of 54 nodes, one for every Swedish institution, including one node for external users. We will give an up-to-date overview of of Sunet Drive, including

- User and storage development
- New customer on-boarding and customizations
- Updates and incidents
- Extension to a third data center
- Implemented and planned features

Special focus of the community report will lie on the plan to develop Sunet Drive into a sovereign academic toolbox, capable of FAIR data handling and data analysis. This includes our efforts in developing Secure Zones and Step-up-Authentication, as well as the integration of RDS and the development of a new connector for the Swedish National Dataservice system DORIS. In addition, we will briefly talk about "Scalable JupyterHub", funded through GN5-1 - GÉANT Project Incubator, and planned to be integrated into Sunet Drive.

Services and Infrastructures / **181**

Federated Sync&Share

Author: Enrico Signoretti[None]

Co-author: Marco Moschettini [1]

[1] *Cubbit*

Corresponding Author: enrico.signoretti@cubbit.io

In this talk, we focus on a new paradigm to create a federated sync and share platform by integrating Next Cloud and Cubbit.

In this paradigm, NextCloud acts as the sync engine and front-end application for collaborative sharing, while Cubbit provides the scalable geo-distributed infrastructure on which data are stored in a disaster-proof, redundant geo-distributed scheme. The goal of the project is to build a federated cloud spread across European research centers and universities, where single institutions contribute computing and capacity resources and receive back highly resilient cloud storage at a very competitive price. This service is delivered in two different modes: Local or Global. The first is aimed at single departments and small teams, while the latter is aimed at international organizations and larger teams working across different countries. The combination of Next Cloud with Cubbit offers a unique solution centered on key aspects of modern data management and sharing: data sovereignty, security, resiliency and mobility, and cost.

Services and Infrastructures / **144**

Combining NextCloud with Direct Access to dCache at DESY

Authors: Christian Voss[None]; Peter van der Reest[1]; Tigran Mkrtchyan[2]; Tim Moeller[None]

[1] *Deutsches Elektronen-Synchrotron DESY*

[2] *DESY*

Corresponding Author: christian.voss@desy.de

The DESY Sync&Share Service is based on NextCloud and dCache as underlying storage system. It currently offers several `PiB` to customers at DESY and many other laboratories within the Helmholtz Association. Since DESY Sync&Share is also used to store and share scientific data a better integration into the scientific infrastructure is desirable.

Using dCache as backend storage allows for convenient direct access. The Sync&Share dCache can easily serve as a Grid storage element or Rucio storage element. Through different protocols such as NFS, XrooD or WebDAV data shared by other scientists can easily be analysed on the DESY compute clusters. An integration with services such as FTS is possible.

The challenge is the integration with NextCloud. Data not written through NextCloud directly is unknown to NextCloud. Through file-scans externally written data can be imported into NextCloud. However, dCache offers access and billing data about every file transfers. By passing these through an event streaming platform it is easy to trigger the registration.

In the talk we introduce the use cases at DESY, show the current setup and discuss limitations.

Cloud Interoperability: Handling Data (GEANT SIG-CISS) / **186**

Introduction to the GÈANT Community Programme

Corresponding Author: dawn.ng@geant.org

In this presentation the goals, strategy and structure of the GÈANT Community Programme will be presented.
Examples of community involvement will be given, with the aim to inform researchers and institutions about the concrete
possibility of support from the GÈANT Community to innovative projects. Examples will be provided of community engagement
initiatives and plans”

Cloud Interoperability: Handling Data (GEANT SIG-CISS) / **187**

GÉANT cloud services and research

Corresponding Author: david.heyns@geant.org

A brief update on GÉANT cloud activities which will discuss a few community case studies and highlight the status of the OCRE 2024 tender.

Cloud Interoperability: Handling Data (GEANT SIG-CISS) / **188**

The GRNET Cloud and the GRNET approach to hybrid

Corresponding Author: louridas@grnet.gr

Support of commercial cloud services has become fundamental to many of the NRENs, given that 40 NRENs participate in the OCRE cloud framework, and that the consumption of commercial cloud services by the NREN constituency increases by up to 100% year on year in many countries. Looking into the future, OCRE 2024 is looking at a framework value of €1.5 Bn. The easiest, and most flexible way for a National Research and Education Network (NREN) to allocate public cloud resources to the community is to allocate budgets to specific projects. Users are then at liberty to use the cloud resources they need, when they need them, and as they need them; the NREN does not need to predict, product, and allocate resources of specific types. To do that, however, it is necessary to ensure that projects remain at the approved budget. That runs contrary to the pay-as-you-go model of public cloud providers, where consumers of services pay retrospectively for consumed resources. Finding a way to create and enforce budgetary limits upfront is therefore fraught with significant technical tasks. We report how GRNET has worked for a technical solution to the problem, the lessons learned, and the way ahead.

Cloud Interoperability: Handling Data (GEANT SIG-CISS) / 184

HEANet Development of a national shared storage service for active research data in Ireland

Author: Roberto Sabatino[1]

[1] *HEAnet*

Corresponding Author: roberto.sabatino@heanet.ie

We are developing the specification of a national service for storage of active research data. Technically, what we're doing is well known to this community as it is based on nextcloud and EUDAT's B2Drop.

In this presentation we will elaborate on the national context in which this work is taking place, with focus on the challenges such as organisational structures and resourcing within institutes, national policies and strategies in respect of research Infrastructures and of course funding models. Some of the issues presented may resonate with other NRENs, and we welcome a dialogue on these points.

Cloud Interoperability: Handling Data (GEANT SIG-CISS) / 190

Bringing users'data to the (super)computers at CSC

Corresponding Author: kalle.happonen@csc.fi

The needs of the user data management has grown constantly in the last years.
With the current generation of supercomputers CSC did some restructuring on how user data is brought to CSC's services.
The current workflow centers around Allas - an S3/SWIFT object storage service, which is central to the data management at CSC. What has the user experience been like after some years of using this approach?
What have the users been missing? What are the main pain points, and what are our development targets?

Keynote / 160

From Data To Knowledge: Computing For High-Energy Physics Experiments

Corresponding Author: mario.lassnig@cern.ch

Modern large-scale sciences have become increasingly complex and face unprecedented data & compute challenges. The number of data-intensive instruments generating substantial volumes of data is growing and their accompanying internal and external workflows are becoming more complex every day. Their storage and computing resources are usually heterogeneous and are distributed at numerous geographical locations belonging to different administrative domains and organisations. In this presentation, we will try to give an insight how we solved these problems in High Energy Physics: from data taking, distributed processing, infrastructure, and user analyses, with specific examples from the ATLAS Experiment.

Collaborative Data Science and Visualisation / **129**

Jupyterhub on Kubernetes as a platform for developing secure shared environment for data analysis at MAX IV

Author: Andrii Salnikov[None]

Co-authors: Zdenek Matej ; Dmitrii Ermakov ; Jason Brudvik

Corresponding Author: andrii.salnikov@maxiv.lu.se

MAX IV Laboratory has operated as a user facility since 2016 and continuously evolving the IT infrastructure to facilitate data collection and enable end-user data analysis possibilities. Jupyterhub running on the bare-metal Kubernetes cluster is one of the primary environments at MAX IV premises aimed to address the challenge of providing secure and shared service, while optimizing access to compute and GPU resources for scientific data analysis.

An initial key objective was the development of a fully unprivileged container environment that operates seamlessly with existing user credentials. This approach aims to enhance security without compromising accessibility of the scientific data. The goal of achieving this without direct modification of the notebook container image is challenging but solved via the introduction of helper services that sync data from the user database (LDAP) to the Kubernetes objects and mounting the overlay inside the container.

In order to achieve enhanced resource visibility within the containers, the project integrates LXCFS [1] into the platform. This integration provides a comprehensive view of available resources, a pivotal feature for efficient data analysis and management when it comes to running parallel tasks e.g. via OpenMP.

The project also emphasizes robust GPU sharing support, covering both V100 and A100 GPUs, including dedicated and shared MIG partitions on the A100 GPUs. Management of GPU memory within Kubernetes is facilitated through MortalGPU [2] - an in-house developed fork of MetaGPU. This solution allows for overcommitment and limitation of GPU memory, akin to RAM, providing fine-grained control and container-scoped visibility of GPU resources. Prometheus exporter of MortalGPU shows GPU resource utilization, offering usage statistics in Grafana and insights into resource availability during instance spawning.

To further extend capabilities, JupyterHub hooks are utilized to introduce “Compute Instance profiles” support to offer end-users to choose between shared GPU partition or dedicated resource options. Moreover, hooks are used for enabling precise control over profiles and images access restrictions through RBACs defined in the IdP.

Additionally, extra containers within the Pods with Jupyter Notebooks are conditionally deployed to enforce features like walltime restrictions for notebooks or utilize JupyterHub as an OIDC client, providing JWT tokens to CLI tools (such as in the Nordugrid ARC [3] client case).

Running the setup on Kubernetes platform brings benefits of re-using the infrastructure to deploy several environments. Currently we are running production deployment, test deployment (for minor updates testing before production) and the "next"deployment for major further developments. We are working towards establishing another deployment exclusively for providing EOSC service as Open Data analysis platform. Moreover the same infrastructure is used for Jupyter notebooks CI testing as well [4].

This project exemplifies successful usage of JupyterHub on Kubernetes as a base platform for developing a secure, shared environment for data analysis at MAX IV. By addressing resource management, security, and extensibility, it delivers collaborative scientific data analysis platform.

The contribution will provide implementation details of the platform and example use-cases running at MAX IV.

[1] LXCFS: https://linuxcontainers.org/lxcfs
[2] MortalGPU - Kubernetes device plugin implementing the sharing of Nvidia GPUs between workloads: https://gitlab.com/MaxIV/kubernetes/mortalgpu
[3] "Advanced Resource Connector middleware for lightweight computational Grids". M.Ellert et al., Future Generation Computer Systems 23 (2007) 219-240.
[4] Brudvik, J., Schoen, S., Matej, Z., & Barty, A. (2021). ExPaNDS Testing and Validation Framework (1.0). Zenodo. https://doi.org/10.5281/zenodo.5718671

Collaborative Data Science and Visualisation / 150

Develop data-centric web apps in Jupyter with Voilà and VOIS

Author: Davide De Marchi[1]

Co-authors: Armin Burger [1]; Pierre Soille [1]; Pieter Kempeneers [1]

[1] *European Commission - Joint Research Centre*

Corresponding Author: davide.de-marchi@ec.europa.eu

After more than four years of experience in developing dashboards with Jupyter and Voilà [1] and the development of a library that simplifies the creation of user interfaces for compelling interactive visualization, we can say that yes: it is possible to use Jupyter as an advanced development environment for the creation of complex web applications, centred on data and equipped with a simple but modern user interface and perfectly capable of supporting interactive and integrated exploration on multiple datasets.
What we have done in this period at the Joint Research Centre of the European Commission, within the on-premise cloud infrastructure called **BDAP** [2] [3] (Big Data Analytics Platform), was to support our fellow scientists and researchers with advanced storage services for big data, parallel and GPU processing, data analysis and Machine Learning/Deep Learning.
In the final phase of many of the collaborations we have established with the scientific units of the JRC, the communication of research results has gradually taken a more important role. Whether it is to inform colleagues and collaborators, as well as to communicate results to the political DGs of the Commission, to publish in scientific journals or to present research work to an audience of non-experts, over time the creation of interactive tools that allow autonomous exploration of the data and results generated by research have met with growing success.
We started from simple dashboards where basic graphs allowed us to summarize the concepts being researched, and then gradually added geo-spatial visualizations with the aim of geographically framing the phenomena studied, to finally arrive at complex applications that can be used via the web for interactive communication and the exploration of multiple and interconnected datasets. Geographical, textual and numerical data integrate with customized charts and visualizations created ad-hoc to best describe the object of study and the results obtained. Very often, these applications accompany and complete the publication of the research in scientific journals, providing a tool, open to all, which allows for the interactive and autonomous verification and exploration of the study results. Moreover, we recognise the importance of Jupyter in this process. As it is the standard environment

vastly used by data scientists and researchers from all over the world for the interactive exploration and analysis of data, being able to use the same framework also for the communication of research results, it undoubtedly guarantees enormous advantages in terms of continuity and simplicity. Researchers, data scientists and software developers can freely move from the focus of the research to its final communication while always remaining within the same work environment.
The creation of Voilà in 2019 enabled Jupyter notebooks to be automatically transformed into interactive dashboards, opening up the possibility of using Jupyter as an application development environment. The availability of widget libraries (ipywidgets [4] initially, then followed by ipyvuetify [5] and others) made it possible to insert graphic elements into dashboards that were directly connected to the data and allowed guided exploration. In this context, with the experiences made in BDAP, we began to collect code snippets that could be generalized and reused, until they became the core of a Python library. This is how the **VOIS** (VOIlà Simplification library) library was born, which a few months ago we managed to make Open Source and publish in the code.europa.eu repository of the European Commission [6] [7].
Through the VOIS library, we aim to simplify the development of data visualization applications and support the rapid creation of the application's visual infrastructure by making available a very vast catalogue of reusable and customizable graphic widgets. The library has been enriched over time with tools to:

- manage responsiveness in a simplified way (i.e. to automatically adapt the display to the different graphic resolutions of the screens, targeting smartphones, tablets and desktop screens of any resolution),
- create multi-page applications (where the user can navigate through different pages that alternate on the screen and present distinct and complementary functions),
- add modal dialog boxes that allow forms and graphic views to be superimposed on the background page allowing for a modern interaction similar to desktop applications.

All this is accompanied by specific modules to manage the subdivision of space on the page, functions that allow for the download/upload of data between the user's PC and the application that runs on the cloud, a gallery of interactive charts created in SVG [8] to fill any gaps present in the open source charting libraries, a centralized management of colours and themes that allows for easy control of the graphic coherence in all parts of an application.
All this has allowed us to move from the creation of dashboards to the creation of real web applications, with multiple windows and complex integrated functions, to provide the user with advanced visualizations and highly interactive experiences.
It is also interesting to note that the recently introduced VaaS service (Voilà as a Service) enables BDAP users to autonomously create and deploy their dashboard in production, through an automated procedure based on Gitlab repositories. This new service, together with the intensive training on the usage of the VOIS library, is contributing to the spreading of Voilà dashboards usage by many research groups in the JRC.
With this presentation, we intend to provide a review of the main applications that we have developed in recent years using Jupyter, Voilà and the VOIS library and which touch on a great variety of different scientific fields, from citizen science to earth observation, from the analysis of European agriculture to air quality monitoring. Moreover, in showing these applications, we also want to talk about the difficulties encountered, the obstacles overcome and the lessons learned, providing ideas on how to best use Jupyter as a development environment for complex web applications. Jupyter was not born with a focus on software development, but on data analysis. Nevertheless, with the experiences that we have had and the guidelines we have derived from it, we want to demonstrate that it also performs reasonably well as a development and testing environment for data-centric web applications.
The development of the VOIS library was partially funded by the JRC's participation in the Horizon2020 project called **CS3MESH4EOSC** and led by CERN which ended in 2023 [9].

[1] https://voila.readthedocs.io/

[2] P. Soille, A. Burger, D. De Marchi, P. Kempeneers, D. Rodriguez, V.Syrris, and V. Vasilev. "A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data". Future Generation Computer Systems 81.4 (Apr. 2018), pp. 30-40.
https://doi.org/10.1016/j.future.2017.11.007.

[3] D. De Marchi, A. Burger, P. Kempeneers, and P. Soille. "Interactive visualisation and analysis of geospatial data with Jupyter". In: Proc. of the BiDS'17. 2017, pp. 71-74. https://zenodo.org/record/3248741#.XeDvSuhKg2w.

[4] https://ipywidgets.readthedocs.io/en/latest/

[5] https://ipyvuetify.readthedocs.io/en/latest/

[6] https://vois.readthedocs.io/

[7] https://code.europa.eu/jrc-bdap/vois

[8] https://developer.mozilla.org/en-US/docs/Web/SVG

[9] https://cs3mesh4eosc.eu/

Collaborative Data Science and Visualisation / **169**

Evolving SWAN through simplification

Author: Diogo Castro[1]

[1] *CERN*

Corresponding Author: diogo.castro@cern.ch

SWAN stands for Service for Web-based ANalysis, also known as CERN's Jupyter service.
The project has undergone a transformative evolution in response to - and to align with - the changes in the upstream Jupyter project. This evolution prompted a simplification of our customizations, enhancing the project maintainability and facilitating deployments beyond CERN.
In this presentation, we will focus on what has changed, from the developers and deployers point of view, while also giving an update on the integrations with CERNBox, CS3 APIs, GPUs and the future towards an Analysis Facility at CERN.

OpenCloudMesh Campfire / **168**

OCM State of the Art

Corresponding Author: giuseppe.lopresti@cern.ch

This presentation will set the scene for the Campfire session.

The current state of the specification and its latest evolution will be presented, with reference to the ScienceMesh infrastructure and the CS3 community at large.

A number of questions will be raised, which can be discussed in the Panel discussion that will follow the topical lightning talks.

OpenCloudMesh Campfire / **145**

Federated groups via OCM

Author: Tom Wezepoel[1]

[1] *SURF*

Corresponding Author: tom.wezepoel@surf.nl

At SURF, we have a large number of cloud environments. But how can you optimally collaborate when users are spread across multiple environments.
Creating a share with a group is much easier, than with several individuals. Our story how we solve this with federated groups.

OpenCloudMesh Campfire / **158**

Trusted servers and MFA with OCM

Author: Micke Nordin[1]

[1] *SUNET*

Corresponding Author: kano@sunet.se

When data is sensitive, it is valuable to know who is accessing it. Multi factor authentication (MFA) aims to solve this by raising the level of assurance of an identity. We can implement MFA for a single EFSS system, but being able to signal requirements of MFA to other trusted systems, and having them honor these requirements, would be very useful.

We can implement this in two parts, one part is by having the EFSS system refuse to share to a non trusted EFSS system, this will be implementation dependent and not in scope of this discussion.

The other part is to add a capability to the OCM specification regarding MFA as well as a separate permission. The combination of these two additions to the specification will allow two EFSS systems to signal to other systems that they will honor the MFA requirements on a share, and conversely signal that a share can only be accessed by a multi factor authenticated user.

This lightning talk will highlight these ideas further and discuss these additions to the OCM specification.

OpenCloudMesh Campfire / **155**

OCM discoverability through DNS

Author: Hugo Gonzalez Labrador[1]

[1] *CERN*

Corresponding Author: hugo.gonzalez.labrador@cern.ch

The OCM protocol has recently introduced a groundbreaking feature known as the “invitation workflow,” designed to enhance the discovery of users across diverse institutions. This innovative approach, while effective in facilitating discoverability, is currently dependent on a singular directory of trusted sites.

Drawing inspiration from the proven practices of email protocols over the past decades, this presentation will delve into the realm of possibilities for elevating discoverability. Our exploration will focus on leveraging DNS records, not only for trusted networks but also for public OCM workflows. We aim to redefine and optimize the user discovery process, ensuring a more robust and adaptable solution for seamless inter-institutional communication.

OpenCloudMesh Campfire / **176**

Evolving the OCM test suite to ease implementations' compliance

Corresponding Author: kontakt@mesterheide.net

We give an update on the latest development of the OCM test suite. This includes the switch from Puppeteer to Cypress as a testing framework and the ability to execute tests in CI pipelines. We also present current test coverage and vendor support.

OpenCloudMesh Campfire / **173**

Standardizing Open Cloud Mesh as an open standard

Corresponding Author: michiel@unhosted.org

We received funding from NLnet and created a W3C community group. The protocol is now versioned and we are spending effort to document it properly with a specification that every vendor can follow. This will increase the value of OCM to end users, and we hereby invite vendors to get involved in the evolution of the protocol.

OpenCloudMesh Campfire / **177**

OCM panel: where do we go from here?

Corresponding Authors: giuseppe.lopresti@cern.ch, michiel@unhosted.org, frank.karlitschek@nextcloud.com, kfreitag@owncloud.com, xjqkilling@gmail.com, hugo.gonzalez.labrador@cern.ch, ron.trompert@surf.nl

After a round table from the main stakeholders (Nextcloud, ownCloud, Seafile), we open the debate to talk about the future of OCM and its governance

FAIR Data Management / **178**

Digital repositories for FAIR data management

Author: Lars Holm Nielsen[1]

[1] *CERN*

Corresponding Author: lars.holm.nielsen@cern.ch

The presentation will give an overview over existing and upcoming FAIR-enabling features in Zenodo and InvenioRDM. Zenodo has through the collaboration with Plazi built up the Biodiversity Literature Repository as a prime example of FAIR data management with domain specific metadata in a general-purpose repository. Zenodo will further soon launch a Zenodo-community together with the European Commission with will add further FAIR-enabling features into Zenodo.

FAIR Data Management / **142**

SND Doris and Sunet Drive - FAIR and sovereign data publication in a federated world

Authors: Stefan Jakobsson[1]; Richard Freitag[None]

Co-author: Juri Hößelbarth

[1] *Swedish National Dataservice*

Corresponding Authors: stefan.jakobsson@gu.se, freitag@sunet.se

FAIR data management has come a long way since its first publication of guiding principles for scientific data management and stewardship in 2016. Many universities and funding bodies have adopted FAIR as a de facto standard for their data management processes, and many publicly available systems have been established to support scientists in their goal of achieving compliance with the guidelines.

Public repositories such as Zenodo, based on InvenioRDM, or OSF come with their own policies and retention times, and can generally only be used for fully open and public research data. Setting up own repositories can be cumbersome and time-consuming, while simultaneously implementing custom access control, workflows, and separation of data from metadata can be hard to achieve. This is where the Swedish National Dataservice and DORIS comes into play. One of the new functions in DORIS is that both researchers and staff in the units for research data support (DAU) in Swedish HEIs can update and edit existing data descriptions. In the previous, separated, system, editing was possible for SND staff only. Another new function is that the local DAU can now enter who reviews the uploaded research data, which makes it clearer for researchers as well as for DAU staff who is doing what. There is also a possibility for local DAU staff to issue pre-booked DOIs.

Combined with Sunet Drive, a federated data storage solution and a member of the ScienceMesh, DORIS enables universities to implement sovereign publication processes. Research data resides under the governance of the university, while the publication of metadata is handled by the SND. This includes cases where metadata should be published, while the access to the data can be requested via DORIS. Large data sets can also be handled more easily, since the data resides in Sunet Drive, the file storage solution managed by the university.

Design considerations and challenges during the development of the solution will be discussed, as well as future developments, giving other universities, institutions, and NRENs an overview of the expected effort in case similar solutions are planned to be implemented. This includes developing a connector for the underlying RDS framework, but also administrative tasks required to maintain a curated and metadata driven publication platform

SND Doris in combination with Sunet Drive is an important that can be used by organizations and research groups in achieving FAIR data management, while retaining data sovereignty and control over their published research data.

FAIR Data Management / **151**

Seamless Integration of Data Sharing Repositories with High-Performance Computing Simulation Platform

Authors: Karol Zając[1]; Taras Zhyhulin[1]

Co-authors: Jan Meizner [1]; Maciej Malawski [2]; Marek Kasztelnik [3]; Marian Bubak [4]; Piotr Nowakowski [3]; Piotr Połeć [3]

[1] *Sano Centre for Computational Medicine*

[2] *AGH University of Krakow (PL)*

[3] *ACC Cyfronet AGH*

[4] *AGH Krakow*

Corresponding Author: t.zhyhulin@sanoscience.org

Scientific advancements increasingly rely on complex computational models fueled by diverse datasets. However, the collaborative sharing of these datasets poses significant challenges, hindering progress of research within scientific communities. This paper addresses the pivotal issue of efficient data sharing among scientists engaged in advanced simulations and computational modeling.

The proposed solution integrates the Model Execution Environment (MEE) with widely used data repositories such as Dataverse or Zenodo for seamless connection to data. This integration empowers scientists to access external data securely and in accordance with predefined rules directly within their workflow templates on the simulation platform. Notably, this approach eliminates the need for users to allocate local disk space, as the High-Performance Computing (HPC) system fetches the required data directly to the executing job directory.

A compelling use case scenario illustrates the efficiency of this solution: a research team collaboratively working on a computational model relies on external data stored in repositories. Seamless integration features allow team members to execute simulations without utilizing storage on their devices, streamlining the research process. Furthermore, the valuable research data generated during the simulation is being positioned for storage on the data sharing platform. The research team can control access to this data, ensuring that it remains within the consortium and preventing public dissemination until establishing a publication agreement.

This publication is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No 857533. This publication is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. This publication is (partly) supported by the European Union's Horizon 2020 research and innovation programme under grant agreement ISW No 101016503. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016227.

FAIR Data Management / **179**

CERN Open Data

Authors: Jose Benito Gonzalez Lopez[1]; Pablo Saiz[1]; Tibor Simko[1]; Zacharias Zacharodimos[1]

[1] *CERN*

Corresponding Author: pablo.saiz@cern.ch

The goal of CERN Open Data is to make experiment data available to the general public in a way that can be easily used and to preserved for the long-term. This comes with a set of challenges:
* The most important is to make sure that the data follows the FAIR principles: it should be Findable, Accessible, Interoperable, and Reusable.
* That also includes the the long preservation of the data, and ensuring that it is immutable.
* On top of that, the volume of data can also be challenging. Currently, more than 5 PB are accessible through the portal.

This talk will present the current status of the CERN Open Data and its plans, focusing on the storage and access of the data.

FAIR Data Management / **152**

Utilizing RDataFrame for Data Preservation and Open Publishing Data and Analyzes Software for HEP

Authors: Pawel Kruczkiewicz[1]; Leszek Grzanka[1]; Valentina Avati[1]; Kamil Krzysztof Burkiewicz[2]; Maciej Malawski[1]

[1] *AGH University of Krakow (PL)*

[2] *AGH University of Science and Technology (PL)*

Corresponding Author: pawel.jan.kruczkiewicz@cern.ch

CERN produces, analyzes and archives vast amounts of data. To conduct an analysis a lot of software in the form of scripts and code is produced. As the time goes by and new approaches supersede the old ones, the aforementioned artifacts may become hard to understand and setting up and running them can be challenging. This may be a crucial concern when trying to publish the data in an open repository like CERN OpenData. Furthermore, an old code cannot leverage new technological advancements which could potentially enhance its performance.
To address this issue an effort to restore data and analysis from LHC Run 1 has been conducted. This work describes the process of transforming data regarding the analysis and code from the LHC Run 1 at the TOTEM experiment. It utilizes RDataFrame –a modern data processing tool –to transcribe C++ scripts into a form of a comprehensible Jupyter notebook. As a result, the number of lines of code has been greatly reduced, thus enhancing the readability. In addition, the notebook can be run on a novel serverless engine architecture.
The process described in this work shows potential applicability for further data preservation and publication efforts.

Collaboration Products / 166

Indico: an Open Source event management system

Authors: Adrian Mönnich[1]; Pedro Ferreira[1]

[1] *CERN*

Corresponding Author: adrian.moennich@cern.ch

We will be presenting Indico, an event-management system born out of the collaborative spirit at CERN. Initially developed more than 20 years go, to meet the unique demands of the world's largest physics lab, Indico has since evolved and transcended its origins, becoming a globally adopted solution for the organization events of all scales, used in more than 300 organizations world wide, including the United Nations, and integrated in a lively user community. We will be focusing on the main features of the tool, its role at CERN, integration with organizational tools (e.g. CERNBox), as well as the work we have been doing in community building, and our plans and ideas for the future.

Collaboration Products / 134

Unleash the Power of COOL: Seamless Integration, Cutting-Edge Features and Empowered Collaboration for your Documents

Author: Michael Meeks[1]

[1] *Collabora Online*

Corresponding Author: michael.meeks@collabora.com

Join us for an exciting update from the world of Collabora Online (COOL). Let us show you how users and integrators benefit from using a security focused, truly open-source, online office suite.
In this session we'll show you why File Sync & Share and LMS provisions are integrating Collabora Online into their products. Hear about the work we've done over the past year to improve integration APIs, and make COOL an even better experience for its users. See what makes Collabora Online so feature rich and powerful.
For administrators seeking cutting-edge solutions, we have interesting new Kubernetes deployment options that are easy to install and automatically-scale. From seamless deployment and integration of external tools, to easy font management APIs and tooling to enhance compatibility across platforms –COOL makes your job a breeze.
But that's not all! We have infused the underlying LibreOffice technology with a host of essential features for users, building on our foundation of strong interoperability and collaborative editing. Hear about the latest features including improved UX, better document navigation, management of change tracking, easy barcode and QR code creation, multi-page floating tables, and preservation of compact pivot tables, as well as how we have improved performance across the board. We have taken significant strides to make COOL more accessible for users with impairments adding screen reading support, dark mode, keyboard navigation for forms –as well as per-view settings for various rendering features.
Learn how Collabora Online brings scalable, secure, on-premise document editing to everyone – allowing integrators to provide extra functionality to their offering, and users to stay in control of their data.

Collaboration Products / **136**

Refining document collaboration with ONLYOFFICE: when flexibility matters

Authors: Galina Goduhina[None]; Oleksiy Ivanov[1]

[1] *OnlyOffice*

Corresponding Author: aleksey.ivanov@onlyoffice.com

Flexibility is an essential skill for teamwork in general, especially in dynamic and challenging situations. Flexibility factor is also of great significance for document collaboration which nowadays is a must for everyone.

Every day we work with numerous office files together with colleagues, team members, various external users, etc. It is important to be able to collaborate on files from different locations and time zones (especially for distributed teams), from different devices and environments (e.g. when it's needed to integrate a solution into the existing corporate infrastructure), as well as to be able to extend the solution functionality at any time you need it, both internally and externally.

In our presentation, we'll cover the following aspects:

- How ONLYOFFICE allows effectively working with all popular formats of office files: docs, sheets, slides, forms, and PDFs.
- Role of flexibility and its types —what integration options are available for ONLYOFFICE Docs (API, WOPI, plugins, connectors).
- How to organize secure teamwork on sensitive files and research papers using the room-based collaboration environment ONLYOFFICE DocSpace.

Collaboration Products / **132**

Status Update of the no-code platform SeaTable

Author: Christoph Dyllick-Brenzinger[None]

Corresponding Author: cdb@datamate.org

SeaTable is the world leading self-hosted no-code platform. SeaTable enables you to develop and build efficient business process in the shortest possible time. You can easily design your database structure, store any kind of data, define access rights for your team or externals and visualize your data with various charts. Automations help to streamline your work. Digitalization or creation of business processes can be done by everybody without writing one line of code.

In this presentation, I will give an overview of the improvements that happened in SeaTable in the last year.

Collaboration Products / **165**

Exploring Nextcloud Tables

Author: Marcel Scherello[None]

Corresponding Author: marcel.scherello@nextcloud.com

Managing structured data seamlessly alongside files is crucial, especially in research and collaborative projects.

This presentation introduces you to Nextcloud Tables, a tool to blend spreadsheet functionality with database management. With its user-friendly interface, Nextcloud Tables caters to a wide array of professional needs without requiring advanced coding skills.

Discover how Nextcloud Tables revolutionizes data handling by allowing the creation of customizable tables, supporting various data types from simple text to complex progress bars and date/time fields. Learn how it enhances team productivity with real-time collaboration, sharing and integration.

Keynote / **161**

European Strategy for Data and the Resulting Landscape

Presented by: European Commission

Topics of interest to the CS3 community:

Bird's eye view on structure, objectives and scope of Europe's Strategy for Data. The resulting the landscape of EC-supported incentives and projects is based on several pillars:

- Scientific and research community: research and innovation actions (RIA), EOSC
- Industry and public services: data spaces (DEP —Digital Europe Programme) for long-term standard operations
- Flagship initiatives with high-societal impact: Destination Earth, AI4Europe,…

How this translates into services? New platforms are being launched in 2024, such as:

- AI-ON-DEMAND
- Copernicus Data Space
- EOSC EU Node

Are these services supposed to be interoperable? What would be the role of SIMPL protocols?

What is the long term economic model and market incentives to ensure successful implementation of single market for data; B2G, G2B, B2B data sharing. How shall this happen?

What is the long-term strategy to ensure European competitiveness vis-a-vis established or emerging technology superpowers in other parts of the world? How can we be leveraging on the European assets and expertise to achieve this goal. Is European ICT infrastructure / technology / expertise also an asset (in addition to domain specific assets such as Weather Models)?

Panel discussion: EOSC Services & Federated Infrastructures / 131

EOSC EU node - data centric cloud services

Authors: Maciek Brzezniak[1]; Norbert Meyer[2]

Co-authors: Andrea Manzi ; Fredric Wallsten [3]; Guido Aben [4]; Holger Dyroff ; Lars Fischer [5]; Stefan Otto [6]; Zdenek Sustr [7]

[1] *PSNC*

[2] *Unknown*

[3] *Safespring*

[4] *SUNET*

[5] *NORDUnet*

[6] *Sikt*

[7] *Czech Technical University (CZ)*

Corresponding Authors: maciek.brzezniak@gmail.com, meyer@man.poznan.pl, sustr4@cesnet.cz

For those who track the development of EOSC, you'll remember the first five years of "building EOSC"were devoted to building the initial federation of existing research data infrastructures in Europe and design and implement the first EOSC Core service needed to build a web of FAIR data. All of this activity was conducted through grants calls; e.g., for the abovementioned EOSC Core service, the EOSC-Future project conducted a lot of development activities. At some point during this process, thinking inside the EC evolved to the point where they decided to depart from the hitherto used scheme of using grant calls to build EOSC. It seems a desire had grown to test the market for its ability to provide parts of the services needed to construct EOSC, and in May 2022 they did indeed publish a "prior information notice"stating the EC was about to go to tender for three lots to build an actual, functional pilot node for the imagined EOSC.
When that notice to tender (the "PIN") was published, many NRENs and similar entities realised they either were unable to bid (e.g., for incorporation or mandate reasons), or were not comfortable bidding (e.g., conservative legal advice, failure to secure board approval) or did not have the personnel on-board that could manage a bid procedure.
As a result, there was at the time considerable fear that external, commercial operators might swoop in and disrupt the ecosystem of not-for-profit R&E operators, with all kinds of fallout; fragmentation of the provider landscape, loss of training, loss of intra-R&E cohesion, loss of operator-customer relationships, disinvestment, etc.
The resulting tender procedure lasted for over a year and had a few bends and twists. Fortunately, at the time of writing of this submission, we can fast-forward to a good outcome for the community. A group of NREN operators, rounded out through the participation of two friendly commercial operators, has managed to win the two user-facing lots (#2 and #3) of this procurement, at a total contract value of ˜€15M for three years.
This gives the R&E community a seat at the table with the EC; this is a splendid chance to continue to influence the direction of EOSC; it also gives us a wonderful chance to continue to develop our own service portfolio in direct contact with a pan-European user base.
This presentation will cover the makeup of the bid team, the specs tendered for, the process of responding to the tender and the challenges in doing so, and the infrastructure built by the time TNC24 comes around; we will also touch upon future architectural plans and opportunities for the R&E community.

Panel discussion: EOSC Services & Federated Infrastructures / 171

ScienceMesh: Community Federation

Author: Jakub Moscicki[1]

[1] *CERN*

Corresponding Author: jakub.moscicki@cern.ch

Panel discussion: EOSC Services & Federated Infrastructures / 192

The value of Open Science collaborations in Scientific Computing

Corresponding Author: xavier.espinal@cern.ch

Panel discussion: EOSC Services & Federated Infrastructures / 172

Panel discussion: EOSC —what's in there for the CS3 community?

Technology Bricks: Testing and Resilience / 141

Continuous Testing at a Global Scale

Author: Richard Freitag[None]

Co-authors: Magnus Andersson [1]; Micke Nordin [1]

[1] *SUNET*

Corresponding Author: freitag@sunet.se

Sunet Drive is a national file storage infrastructure for universities and research institutions in Sweden. It is based on a Nextcloud Global Scale setup and is comprised of 54 nodes, one prepared for each institution. This setup ensures data sovereignty while being part of a larger federation, including the ScienceMesh for international collaboration. The setup is duplicated in a test environment which is also used for staging and development. In total, around 300 virtual servers are deployed, including nodes and applications setup using kubernetes.

One step in ensuring the functional stability of such a large setup is rigorous and automated testing, which has been identified as one of seven critical success factors (CSFs) that represent continuous practice. Strategic test automation where tests are continuously adapted and executed based on requirements complements traditional IT monitoring solutions, and provides a viable strategy for testing in an industrial-grade DevOps project. This covers basic tests of the status pages comparing expected values with actual values. Functionally, user life cycle management, app-consistency and WebDAV are tested, as well as more sophisticated Selenium testing for login, multi-factor authentication, document editing, and more. Automated testing generates around 30000 daily test points, ensuring the stability and consistency of the deployed solution. Scalability and load testing is done using a setup that can simulate the load from many users, and has been used to test for up to 6000 users.

The presentation covers all stages of the application life cycle. While you model your IT-system, integrate applications and components, you scale and expand it to a global scale. The fundamental question arising from this is "Does it really work all the time?" You start by writing tests that systematically cover core functionality promised to the end users. Automating these tests, e.g., by using pipelines, you continuously test your solution.
We will guide the audience through the whole test automation pipeline, including the technical setup required, such as Jenkins with multiple workers and X virtual frame buffer (Xvfb) for testing in virtual desktop environments. Challenges reducing the test execution times via multi-threaded test execution will be discussed, as well as trade-offs in test coverage. Eventually, you want to validate functional stability for your end users rather than replicating regression testing that should be covered by the software suppliers.

Continuous testing is an important tool that can be used as a complement to conventional infrastructure monitoring. It ensures functional consistency throughout the life cycle of hosted IT solutions.

Technology Bricks: Testing and Resilience / **148**

ownCloud Tech Talk on Kubernetes Deployment, Performance, and Load Testing

Author: Klaas Freitag[1]

[1] *ownCloud*

Corresponding Author: kfreitag@owncloud.com

We share experiences running the microservices-based ownCloud Infinte Scale software with many instances in a highly scalable virtual architecture.
The second part covers motivation, architecture and results of load testing with K6.

Technology Bricks: Testing and Resilience / **127**

All good things come in threes

Author: Jean-Marie de Boer[None]

Corresponding Author: jean-marie.deboer@surfsara.nl

From the very beginning the sync&share services within the Online Data Services group at SURF have been designed with a view of moving them to a geo-distributed setup for extra resilience. In my talk, I will describe the initial design choices, the preparations and the actual steps taken to move a running service from a single datacenter to three datacenters without as much of one second of downtime. I will also touch on the ongoing challenges and areas where we can still improve.

Technology Bricks: Testing and Resilience / **164**

SCION based ScienceDMZ and fast file transfer for HPCCs

Author: Francois Wirz[1]

[1] *ETHZ*

Corresponding Author: wirzf@inf.ethz.ch

Today's research often relies on a high volume of data. While universities cannot always provide the computing resources to process a given amount of data, a high-performance computing cluster (HPCC) offers a cost-effective alternative for researchers.
With a SCION-based Science DMZ, the HPCC and each university operate as independent autonomous system (AS), managing their own cryptographic keys and enforcing their own network rules. Each AS also has its own LightningFilter deployed.
High-volume data transfers between a university and the HPCC are performed using dedicated, SCION-based systems, such as Hercules, which route their traffic through LightningFilter instead of the network's general-purpose firewall.
With LightningFilter, the ASes can control the amount of traffic received from the other ASes or from specific hosts.
For instance, the HPCC can enforce different rate limits for each university while guaranteeing a certain throughput for specific hosts. Such limits are important to protect the services offered from misbehaving hosts that consume more bandwidth than agreed and thus partially or completely block other hosts from reaching the service.

We will present the latest developments on SCION based ScienceDMZ and showcase some early deployments and PoCs.

Technology Bricks: advanced integration / 146

Open Data Lifecycle Management with Onedata

Authors: Michał Orzechowski[1]; Łukasz Opioła[2]; Bartosz Kryza[2]; Lukasz Dutka[2]

[1] *AGH University of Science and Technology, Academic Computer Centre Cyfronet AGH, Krakow, Poland*

[2] *AGH University of Kraków, Academic Computer Centre Cyfronet AGH, Krakow, Poland*

Corresponding Author: orzechowski.michal@gmail.com

Onedata[1] is a high-performance data management system with a distributed, global infrastructure that enables users to access heterogeneous storage resources worldwide. It supports various use cases ranging from personal data management to data-intensive scientific computations. Onedata has a fully distributed architecture that facilitates the creation of a hybrid cloud infrastructure with private and commercial cloud resources. Users can collaborate, share, and publish data, as well as perform high-performance computations on distributed data using different interfaces: POSIX-compliant native mounts, pyfs (python filesystem) plugins, REST/CDMI API, and S3 protocol (currently in beta).

The latest Onedata release line, 21.02, introduces several new features and improvements that enhance its capabilities in managing distributed datasets throughout their lifecycle. The software allows users to establish a hierarchical structure of datasets, control multi-site replication and distribution using Quality-of-Service rules, and keep track of the dataset size statistics over time. In addition, it also supports the annotation of datasets with metadata, which is crucial for organising and searching for specific data. The platform also includes robust protection mechanisms that prevent data and metadata modification, ensuring the integrity of the dataset in its final stage of preparation. Another key feature of Onedata is its ability to archive datasets for long-term preservation, enabling organisations to retain critical data for future use. This is especially useful in fields such as scientific research, where datasets are often used for extended periods or cited in academic papers. Finally,

Onedata supports data-sharing mechanisms aligned with the idea of Open Data, such as the OAI-PMH protocol and the newly introduced Space Marketplace. These features enable users to easily share their datasets with others, either openly or through controlled access.

Currently, Onedata is used in European projects: EUreka3D[2], EuroScienceGateway[3], DOME[4], and InterTwin[5], where it provides a data transparency layer for managing large, distributed datasets on dynamic hybrid cloud containerised environments.

Acknowledgements: This work is co-financed by the Polish Ministry of Education and Science under the program entitled International Co-financed Projects (projects no. 5398/DIGITAL/2023/2 and 5399/DIGITAL/2023/2)

REFERENCES

1. Onedata project website. https://onedata.org.
2. EUreka3D: European Union's REKonstructed in 3D. https://eureka3d.eu.
3. EuroScienceGateway project: open infrastructure for data-driven research. https://galaxyproject.org/projects/esg/.
4. DOME: A Distributed Open Marketplace for Europe Cloud and Edge Services. https://dome-marketplace.eu.
5. InterTwin: Interdisciplinary Digital Twin Engine for Science. https://intertwin.eu.

Technology Bricks: advanced integration / **138**

New iRODS APIs: Presenting as HTTP and S3

Author: Justin James[1]

Co-author: Terrell Russell [1]

[1] *iRODS Consortium*

Corresponding Author: jjames@renci.org

This year's releases of iRODS 4.3.1 as well as standalone APIs exposing iRODS systems via HTTP and S3 help new users use their existing, familiar tools to integrate with an iRODS Zone. This talk will cover the requirements, design, and initial releases of these new APIs.

Technology Bricks: advanced integration / **137**

Utilizing large language models with free and open source software in EFSS while protecting digital sovereignty for Universities.

Author: Micke Nordin[1]

Co-authors: Magnus Andersson [1]; Richard Freitag [1]

[1] *SUNET*

Corresponding Author: kano@sunet.se

The recent media buzz around so called "Artificial Intelligence", divides users of IT systems into two polarized groups. One group is vocal about wanting to use these new tools and points out that these features are missing from offerings outside of IT giants like OpenAI/Microsoft and Google. The

other group is worried about copy right and privacy issues that arise from having all your files and communications scanned by large corporations for generating and commercializing large language models (LLM) and other machine learning (ML) tools.

Luckily, as competent engineers, we can appease both of these groups at the same time using free and open source tools and ethically sourced models running on your own hardware. In this talk I will present how Local AI[0] can be integrated with Nextclouds ML tools[1,2,3]. Having access to GPU resources is helpful, but not necessary for decent results.

You would be forgiven for thinking that large enterprises such as Google and OpenAI have access to resources not available to the general public, and while that is certainly true, it is true for marketing resources more than anything else. In fact machine learning has long been the domain of academia and the free and open source movement, and the contribution of the large corporations in the space is mostly doing the work at scale and defining expectations of users.

There are already mature projects available for running LLM:s on commodity hardware with the same API endpoints as OpenAI presents. This opens up the possibility to integrate with the recent tools that Nextcloud have developed for summarizing text, writing headlines and more (also integrated in text and mail apps for example).

1. https://localai.io
2. https://github.com/nextcloud/integration_openai
3. https://github.com/nextcloud/assistant
4. https://github.com/nextcloud/mail

Technology Bricks: advanced integration / **163**

Closed Domain QA System for LBL ScienceIT: Fine-Tuned and Retrieval Augmented Generation Models

Author: Fengchen Liu[1]

Co-author: Jordan Jung [1]

[1] *Lawrence Berkeley National Laboratory*

Corresponding Author: fengchenliu@lbl.gov

This paper proposes the development of a closed-domain Question-Answering (QA) system for LBL ScienceIT, using the ScienceIT website as the data source. The focus is on evaluating different models, specifically two fine-tuned pre-trained language models and three retrieval-augmented generation (RAG) models. Through this comparison, insights into the performance of these models, based on several evaluation metrics, are derived, ultimately highlighting the potential of a certain approach for the specific task. Through this comparative study, we aspire not only to present a robust QA framework for LBL ScienceIT but also to shed light on the dynamics of model selection and optimization for domain-specific tasks, setting the stage for future advancements in the realm of specialized QA systems.

Technology Bricks: advanced integration / **167**

Leaf.cloud

Authors: David Kohnstamm[1]; Dennis Pennings[1]

[1] *The Good Cloud*

Corresponding Author: david@leaf.cloud

The presentation focuses on the environmental impact of the technology industry, challenging the assumption that it is inherently eco-friendly. It highlights the significant carbon emissions from the tech sector, projected to triple by 2040 without intervention by the growth of AI. The content then shifts to the positive impacts of technology in various sectors like healthcare, education, and business. The main question posed is how to reduce the carbon footprint of digital technology without hindering its progressive capabilities. Solutions proposed include sustainable design, green engineering, and sustainable operations. Additionally, the presentation introduces an alternative, sustainable cloud solution that utilizes residual heat from servers for building heating locally, putting servers where the heat is used instead, this cuts costs per watt of installed IT by 5 to 10 times and saves the environment at the same. Offering an innovative approach to greenify the digital industry by offsetting natural gas use with server waste heat.

191

Summary and Conclusions

Corresponding Authors: guido@sunet.se, jakub.moscicki@cern.ch, massimo.lamanna@cern.ch, ron.trompert@surf.nl, tilo.steiger@id.ethz.ch

Co-located GEANT SIG-CISS Meeting / 193

SIG-CISS Opening and plans for 2024

Corresponding Author: mario.reale@geant.org

Co-located GEANT SIG-CISS Meeting / 194

The GÉANT Community Programme: overview and updated strategy

Corresponding Author: dawn.ng@geant.org

In this presentation an overview of the GEANT community programme will be provided

Co-located GEANT SIG-CISS Meeting / 195

Activities and future plans for the Cloud workpackage WP4 of the GN5-1 project

Corresponding Author: maria.ristkok@eenet.ee

Co-located GEANT SIG-CISS Meeting / 196

Updates from the CS3 community

Corresponding Author: guido@sunet.se

Co-located GEANT SIG-CISS Meeting / **197**

Cloud update from CSC (FInland)

Corresponding Author: kalle.happonen@csc.fi

Co-located GEANT SIG-CISS Meeting / **198**

Cloud update from KIFU (Hungary)

Corresponding Author: molnar.peter@kifu.gov.hu

Co-located GEANT SIG-CISS Meeting / **199**

Cloud update from SWITCH (Switzerland)

Corresponding Author: bernard.landon@switch.ch

Co-located GEANT SIG-CISS Meeting / **200**

Development of a national shared storage service for active research data in Ireland

Corresponding Author: roberto.sabatino@heanet.ie

Co-located GEANT SIG-CISS Meeting / **201**

eduMEET : new release 4.0

Corresponding Author: idzik@man.poznan.pl

Co-located GEANT SIG-CISS Meeting / **202**

Meeting closure

Corresponding Author: mario.reale@geant.org

Wrap-up and closure

FAIR Data Management / 170

Open Data and Open Science at CERN