CS3 2025 - Cloud Storage Synchronization and Sharing

Report of Contributions

Contribution ID: **203** Type: **Presentation**

KubePie: Leveraging Kubernetes for Scalable End-User Data Access

Wednesday 19 March 2025 12:30 (15 minutes)

Does the idea of running and managing "own"web server for every user with the corresponding user/group IDs and securing access to it sound crazy? Definitely 10 years ago it was, but not in the current Cloud Native world. KubePie facilitates streamlined access to end-user data by harnessing Kubernetes' scalability and deployment capabilities to actually running, managing and securing web-servers on a large scale.

Inspired by the success of JupyterHub on Kubernetes, KubePie implements *Permissions Impersonated Environments ("Pies")*, constructed from open-source components to provide a personalized data consumption experience for end-users through any web browser or HTTP client.

The microservices framework within KubePie is structured around various "Pies":

- **PieData:** The personal web-server Pie, stuffed with data services. Wrapped in a nice Helm chart.
- **PieTrack:** Transfer activities keeping for PieData. Tracks your calories consumption over time.
- **PiePass:** Ensure secure passage to the KubePie bakery services for the customers.
- **PieCut:** The "Control Unit" of KubePie bakery, taking orders and delivering PieData.
- **PieDeck:** The "Data Endpoints Controller for Kubernetes" operates the baking process of the PieDatas for end-users.
- **PieSec:** Admission webhook for the security seasoning of the PieData. Adds the UID/GID and SGID spices. Make sure no rotten eggs are added.
- **PieEat:** "Erases at Timeout" the stale PieDatas.

PiePass serves as the authentication and authorization entry point for end-users, relying strictly on OIDC/OAuth2 protocols and leveraging Apache web servers, equipped with mod_oauth2 and mod_auth_openidc modules from OpenIDC.

PieCut is a main action point for end-users. Technically it is Kubernetes Controller that creates PieData Custom Resource Definitions (CRDs), based on authenticated user information from HTTP headers set by PiePass.

PieDeck is a Helm-based Operator using OperatorSDK to manage PieData deployments via PieData CRDs. It is a central configuration point of common PieData settings (e.g. Images, Resources).

PieSec is the Admission Controller for Kubernetes providing a securityContext based on the JWT claims information and a pluggable mapping backend config (static mapfile and LDAP are implemented). Keeping PieSec separate (compared to defining securityContext by PieCut) improves security, UID/GID comes as an enforcement not as a request and is separated from the main Pod spawning logic.

PieData is taking care of servicing the main data traffic. End-user is redirected to a PieData endpoint via the proper Ingress configuration and uses Apache with OpenIDC/OAuth2 modules as well to authenticate the end-user. Optionally, ephemeral stateless credentials for WebDAV can be generated as well.

PieData is capable of carrying multiple containers inside the same Pod to enhance the functionality. In particular, data analysis services, such as HDF5 viewer or even JupyterLab, can be provided. Thanks to PieSec, all PieData services are running with restricted permissions and storage access.

PieTrack is essentially a custom Prometheus exporter for Apache. This way KubePie offers both, generic observability and an interface for PieEat. With Prometheus metrics it is easy to define a dashboard for data transfers monitoring, including the per-user traffic statistics.

PieEat is another rather simplistic Kubernetes Controller that acts on data-transfer inactivity and removes unused PieDatas. This contributes to both efficient resource utilization and security. KubePie framework is designed to run services on demand, avoiding unnecessary infrastructure usage by idle services.

The framework deployment at MAX IV, illustrates the utilization of KubePie for scientific data access with integrated HDF5 viewer. It runs on a bare-metal Kubernetes cluster with native SpectrumScale data access, Keycloak SSO for authentication and LDAP for user mapping, delivering high-speed data transfer rates via standard HTTPS protocol.

Another use-case KubePie helps to solve at MAX IV is on-demand OpenVSCode Server instances, mounting user own “home”directory, providing easy IDE and web-terminal experience.

Authors: SALNIKOV, Andrii (Lund University (MAX IV)); ERMAKOV, Dmitrii (Lund University)

Presenter: SALNIKOV, Andrii (Lund University (MAX IV))

Session Classification: CS3 Jupyter SIG & Data Science and Visualisation Platforms

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **204** Type: **Presentation**

A community has to do what a community has to do.

Thursday 20 March 2025 09:00 (45 minutes)

When proprietary products are discontinued or bought up by competitors, your own (decision-making) freedom can quickly become precarious. A discontinued product forces you to migrate, incurs costs, and imposes unwelcome decisions.

Not so with open source software: it offers me the freedom to develop the code further at any time, either on my own or with new partners, and to set up new support chains. The freedom not to have to migrate and the freedom not to have to work with an unpleasant service provider. Even drastic changes in product and business strategy can be avoided by taking matters into your own hands.

At least that's the theory. But does it work in practice?

This report is about a very recent case: about software that wasn't supposed to die, and a team that absolutely wanted to continue. It also shows where theory meets practice and the headwinds you have to face when you just go for it. But: a community has to do what a community has to do. And that can also mean: let's fork!

Author: Mr HEINLEIN, Peer (CEO and Founder OpenCloud GmbH)

Presenter: Mr HEINLEIN, Peer (CEO and Founder OpenCloud GmbH)

Session Classification: Keynote

Track Classification: Keynotes

Contribution ID: **205** Type: **Presentation**

Conquering new horizons: The latest from ONLYOFFICE

Thursday 20 March 2025 15:45 (20 minutes)

In this session, we'll provide an in-depth look into the latest achievements and features of ONLYOFFICE, a leading open-source office software project with focus on secure document processing.

We will cover the following:

• What are the novelties of ONLYOFFICE over the year, including a full-featured collaborative PDF Editor, integration news, etc.
• How to organize effective teamwork in any branch, be it research, education, public sector, etc.
• Provide valuable insights into ONLYOFFICE's evolving capabilities.

Author: GODUHINA, Galina

Presenter: GODUHINA, Galina

Session Classification: Collaboration Products

Track Classification: Main sessions: Collaborative Applications, Data Privacy and Data Classification

Contribution ID: **206** Type: **Presentation**

The combo of AI & office software: Unlimited possibilities to work with documents

Friday 21 March 2025 14:30 (15 minutes)

Innovations in Artificial Intelligence have led to it becoming an integral part of the society and finding applications in a variety of fields. We analyzed recent requests and cases which we have faced ourselves and came to the conclusion that the use of AI in this or that field is somehow related to documents.

In this session, we will:

- find out what is the connection between document editors and AI;
- highlight what benefits and issues AI can bring to users when working with documents;
- cover AI implementation into office software using the experience of ONLYOFFICE.

Author: GODUHINA, Galina

Presenter: GODUHINA, Galina

Session Classification: AI-based Innovations

Track Classification: Main sessions: AI and storage

Contribution ID: **207** Type: **Presentation**

Security in office software: an ever-important trend

Wednesday 19 March 2025 17:20 (15 minutes)

Data security is a crucial question for everyone who works with docs online, especially now with all-around implementing the AI technology.

The data security aspect involves protecting sensitive information, such as research data or customer details from unauthorized access, corruption, or theft. Therefore, robust security measures are necessary to safeguard this information. These measures may include the use of strong, unique passwords, two-factor authentication, signatures, regular software updates to patch any security vulnerabilities, etc.

In this session, we will cover how ONLYOFFICE provides a comprehensive level of security for online document editing and collaboration, including:

- various security tools and services;
- flexible access rights;
- 3 levels of encryption.

Author: GODUHINA, Galina

Presenter: GODUHINA, Galina

Session Classification: Features & Principles

Track Classification: Main sessions: Technology & Research

Contribution ID: **208** Type: **Presentation**

Status Update of the no-code platform SeaTable

Thursday 20 March 2025 16:25 (20 minutes)

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.

Author: DYLLICK-BRENZINGER, Christoph

Presenter: DYLLICK-BRENZINGER, Christoph

Session Classification: Collaboration Products

Track Classification: Main sessions: Collaborative Applications, Data Privacy and Data Classification

Contribution ID: **209** Type: **Presentation**

The power of automation and how to control it

Wednesday 19 March 2025 17:50 (15 minutes)

In today's fast-paced academic and research environments, efficiency is key. This presentation introduces n8n, an open-source workflow automation platform that can transform how educational institutions and researchers manage their daily tasks and data processes.
Key Points:
- Seamless Integration: n8n connects with over 350 applications, allowing for easy automation of tasks across various platforms used in education and research.
- Time and Resource Optimization: By eliminating repetitive tasks, n8n saves valuable time and reduces the risk of human error, allowing educators and researchers to focus on high-value activities.
- Customizable Workflows: With its intuitive graphical interface, users can create complex workflows tailored to specific research or administrative needs.
- Data Privacy and Security: n8n offers self-hosting options, ensuring that sensitive educational and research data remains within your control.
- Community-Driven Innovation: A strong user community contributes to an ever-growing list of integrations and pre-defined workflows, perfect for sharing best practices in academia.
- Cost-Effective Solution: n8n provides a comprehensive free version, making it accessible to educational institutions with limited budgets.
- Advanced Capabilities: For tech-savvy users, n8n supports custom node development and integration with AI functionalities, opening up possibilities for cutting-edge research automation.

This presentation will demonstrate how the open source workflow automation platform n8n can streamline various processes in education and research. By adopting n8n, educational institutions and researchers can significantly enhance their productivity, allowing them to dedicate more time and resources to their core mission of education and discovery.

Author: DYLLICK-BRENZINGER, Christoph

Presenter: DYLLICK-BRENZINGER, Christoph

Session Classification: Features & Principles

Track Classification: Main sessions: Technology & Research

Contribution ID: **210** Type: **Presentation**

Data transfer, synchronization and sharing solution for PSDI

Thursday 20 March 2025 17:45 (15 minutes)

PSDI is the UK nationally funded programme that analyses physical sciences needs in a common data infrastructure and develops guidance, training and technology to address these needs. The PSDI main objective is to serve research use cases originating in experimental "bench science" and simulations with applications in physics, chemistry, materials research or engineering. The main challenge to address by PSDI in the well-known "three Vs"of Big Data: Volume, Variety and Velocity is not Volume but primarily Variety aspect with some considerations given to Velocity, too.

Data transfer, synchronization and sharing solution is a part of broader technology works in PSDI and relies on Open Source components with a strong inclination to containerised and cloud deployments. The IT stack for developing the solution includes OCIS (Own Cloud Infinite Scale) with Ceph object store as a backend, combined with additional tools and an orchestration component based on Apache Airflow. We are reporting on integration of components, on performance measurements and on implementation of data policies that are essential to have in a common data infrastructure. We are discussing the potential of combining the data transfer, synchronisation and sharing solution with data pipelines and a data indexing solution that are also in scope of PSDI technology works.

[1] PSDI –Physical Sciences Data Infrastructure. www.psdi.ac.uk

[2] Towards data sharing service for Physical Sciences Data Infrastructure. CS3 2024. https://indico.cern.ch/event/1332413/c

[3] Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. Volume 35, Issue 2, April 2015, Pages 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Author: BUNAKOV, Vasily (STFC UKRI)

Co-authors: BELOZEROV, Alexander (STFC UKRI); PAWULA HEWAGE, Amali (STFC UKRI); WRIGHT, Paul (STFC UKRI); MEACHAM, Stuart (STFC UKRI); UNDERWOOD, Tom (STFC UKRI)

Presenter: BUNAKOV, Vasily (STFC UKRI)

Session Classification: Data sharing infrastrcutures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **211** Type: **Presentation**

AI-Powered File Management: Transforming Digital Chaos into Intelligent Order

Friday 21 March 2025 14:45 (15 minutes)

Imagine a world where your documents organize themselves, where finding the right file is as simple as asking a question, and where AI does the heavy lifting of managing your digital information.

From legal firms to healthcare, from small startups to global enterprises, AI is reshaping how we interact with our digital information.

In this presentation, I will unveil how artificial intelligence is revolutionizing file management:
- Discover how AI can automatically categorize, tag, and analyze your documents
- Learn how intelligent search can turn hours of manual searching into seconds of precise retrieval
- Explore a real-world application that are saving businesses time, money, and frustration

Author: DYLLICK-BRENZINGER, Christoph

Presenter: DYLLICK-BRENZINGER, Christoph

Session Classification: AI-based Innovations

Track Classification: Main sessions: AI and storage

Contribution ID: **212** Type: **Presentation**

Sciebo site report: Migrating 230k users to Nextcloud

Thursday 20 March 2025 10:40 (15 minutes)

ownCloud 10 is EOL and for sciebo we needed to make a difficult choice on how to continue for the upcoming years.
As a result, we are now in the process of migrating 230k users across approximately 45 instances from ownCloud 10 to Nextcloud.

We will report on:
- why we chose to continue our own helm chart
- fixes and features we needed before making the move
- first roadblocks and how we've overcome them
- current progress
- hopefully no horror stories
- streamlining database operations of 45 galera clusters
- evolution of our application landscape

Author: WUNDERLICH, Marcel

Co-author: ANGENENT, Holger (University of Münster)

Presenter: WUNDERLICH, Marcel

Session Classification: Operations and development of CS3 services

Track Classification: Main sessions: CS3 Community Site Reports

Contribution ID: **213** Type: **Presentation**

Nextcloud. State of the nation

Thursday 20 March 2025 14:45 (30 minutes)

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.

Author: KARLITSCHEK, Frank

Presenter: KARLITSCHEK, Frank

Session Classification: EFSS Products

Track Classification: Main sessions: File Sync & Share Solutions and Requirement from the Community

Contribution ID: **214** Type: **Presentation**

Scalable, Secure, Seamlessly Integrated Document Editing

Thursday 20 March 2025 16:05 (20 minutes)

Join us to hear about the latest work from the world of Collabora Online (COOL). Hear about the new integration points and APIs that let us create a richer integration between storage and our 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 how EFFS can easily deploy and configure rich document functionality such as AutoText giving an even better experience for users.

For administrators seeking cutting-edge solutions, we will look at our multi-tenant remote configuration capabilities, and touch on auto-scaling in Kubernetes. From seamless deployment and integration of external tools, to easy configuration 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, @user notification, improved style handling, Impress editing improvements, 3D transitions, speaker notes and presenter console, as well as new APIs to make automatic document generation easy.

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.

Author: MEEKS, Michael

Presenter: MEEKS, Michael

Session Classification: Collaboration Products

Track Classification: Main sessions: Collaborative Applications, Data Privacy and Data Classification

Contribution ID: **215** Type: **Presentation**

Sunet Drive - Sweden - Community Site Report

Thursday 20 March 2025 10:25 (15 minutes)

Sunet Drive is Sweden's national data storage solution, and part of the ScienceMesh and active member of the Open Cloud Mesh Community Group. 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 and rolling out Secure Zones and Step-up-Authentication, including the transition from a Nextcloud Global Scale to a "purely federated"deployment, as well as the integration of RDS-NG and the development of a new connector for the Swedish National Dataservice system DORIS. In addition, we will briefly talk about the status of "Scalable JupyterHub", funded through GN5-1 - GÉANT Project Incubator.

Authors: NORDIN, Micke (SUNET); FREITAG, Richard

Co-authors: Mr DELHAGE, Lars (Sunet); Mr DANIELSSON, Rikard (Sunet)

Presenter: FREITAG, Richard

Session Classification: Operations and development of CS3 services

Track Classification: Main sessions: CS3 Community Site Reports

Contribution ID: **216** Type: **Presentation**

Tales of FEII - Function, Efficiency, Innovation & Integration

Wednesday 19 March 2025 17:00 (20 minutes)

Effective storage solutions have grown into complex systems, providing services with requirements from a variety of stakeholders. Tales of FEII is a story-driven presentation from the perspective of an NREN, namely Sunet in Sweden, on how such systems can be implemented and improved over time by applying four core values to everything they do: Function, Efficiency, Innovation & Integration.
Function represents the fundamental deliverables of services, provided by implementing off-the-shelf solutions, covering the majority of the stakeholder requirements. Efficiency is not only required to scale a solution to a certain size but also to remain performant or to use as few (human) interactions as possible to achieve a certain task. Some like to describe those through non-functional requirements or other KPIs. Innovation is needed when the stakeholders have requirements that have not yet been implemented by the vendor, often requiring careful assessment of the known unknowns. Integration is the often underestimated effort of combining standalone functionality into a seamlessly integrated solution.
During the presentation, we will walk you through the four core values of FEII by providing representative examples, elaborating further on how one can apply them to their own projects, and how software vendors can use them to prioritize some of their development efforts.

Author: FREITAG, Richard

Co-authors: NILSSON, Anders; NORDIN, Micke (SUNET)

Presenter: FREITAG, Richard

Session Classification: Features & Principles

Track Classification: Main sessions: Technology & Research

Contribution ID: **217** Type: **Presentation**

What's new in Seafile for 2024

Thursday 20 March 2025 14:15 (30 minutes)

Seafile is a popular open-source file sync and share solution, used by many organizations (edu and for-profit). Its features include robust and efficient file syncing, cross-platform virtual drive clients, efficient usage of server resources, and encrypted libraries.

In this talk we'll present updates of Seafile in the year 2024. Notable updates:
- Redesigned user interface
- A new wiki module
- SeaDoc, a light-weight collaborative document editor, is production ready

Author: XU, Jonathan

Presenter: XU, Jonathan

Session Classification: EFSS Products

Track Classification: Main sessions: File Sync & Share Solutions and Requirement from the Community

Contribution ID: **218** Type: **Presentation**

NORTRE - national collaboration for federated access to secure HPC

Friday 21 March 2025 12:00 (15 minutes)

Working on highly sensitive research data is challenging and is often hindered by both legal and technical obstacles. It requires not just secure data storage, but also secure data processing, and a secure collaborative platform. The purpose of this talk is to show how the University of Oslo (UiO), the University of Bergen (UiB) and the Norwegian University of Science and Technology (NTNU) in Trondheim are working together to provide federated access to centralised secure HPC. Each university has its own trusted research environment (TRE). UiO has services for sensitive data (TSD), UiB has Safe, and NTNU has HuntCloud. Each TRE is developed in-house over several years, and offer scalable, secure storage and data processing software in a secure environment. However, only TSD has its own compute cluster, for running jobs on large collections of sensitive data. By establishing trust and using APIs for data migration between the TREs, users will be able to send compute jobs directly to the TSD cluster from within one of the other two TREs. Through this, the users get the best of all three worlds, and an easier way of collaborating across platforms, without building a grand new solution, but instead linking existing secure research platforms.

Author: Dr BERGSAKER, Anne (University of Oslo)

Presenter: Dr BERGSAKER, Anne (University of Oslo)

Session Classification: HPC data access and integration

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **219** Type: **Presentation**

Computational Intelligence Architecture for Continuous Learning in Medical Centres

Friday 21 March 2025 14:10 (20 minutes)

In the early days of artificial intelligence (AI) during the 1950s, two primary approaches emerged. One was engineering-oriented, while the other focused on computational modeling of human decision-making processes, later termed "computational intelligence", and is strongly determined by three fundamental time-constrained limitations: data, computation, and communication. Modern AI development emphasizes scaling data and computational resources, operating on the premise that machines are not bound by the constraints of limited data and computational capacity.
This work presents a Computational Intelligence Architecture used to support continuous learning processes and deployment of classification models within Medical and Research Centers on health data, and presents mechanism of communicating findings between these centers.
The architecture implements a distributed network of Agents that run containerized classification models on local medical data stored on Medical Center's premises, display the obtained results locally for doctors' decision making process support and share results via a knowledge data bank available to all participating centers without sharing the data itself. Additionally, the system allows for models' meta-analysis for further improvement with a growing number of medical cases.

Authors: Mr NOWAK, Adam (Sano Centre for Computational Personalised Medicine); Mr KATULSKI, Filip (Sano Centre for Computational Personalised Medicine); Dr SOUSA, Jose (Sano Centre for Computational Personalised Medicine)

Presenter: Mr KATULSKI, Filip (Sano Centre for Computational Personalised Medicine)

Session Classification: AI-based Innovations

Track Classification: Main sessions: AI and storage

Contribution ID: **220** Type: **Presentation**

State of file sync and share at University of Oslo

Thursday 20 March 2025 11:15 (15 minutes)

The Educloud service at University of Oslo has provided our researchers and collaborators access to a suite of tools for collecting, storing and sharing data, and contains a suite of services including a Nextcloud service.

After a survey of the different on premises and cloud storage services and sync tools we currently offer, we have seen the need for a more strategic approach to end user storage. One of the things we consider is to replace the old home directories and file shares with a larger nextcloud installation.

Our new nextcloud installation is currently in a PoC phase. In the talk we will give you an overview of the service, what goals we try to achieve, and how we work to form our roadmap for end user storage.

We will talk a bit about how we are implementing federated authentication for university users and collaborators, and we will give a very brief overview of the architecture of the service.

At the date of the conference, we will hopefully also have some experience from the PoC to share, and we will talk a bit about our future plans for the service.

Authors: BRUVIK, Anders; MEDEIROS-LOGEAY, Francis Augusto (University of Oslo)

Presenters: BRUVIK, Anders; MEDEIROS-LOGEAY, Francis Augusto (University of Oslo)

Session Classification: Operations and development of CS3 services

Track Classification: Main sessions: CS3 Community Site Reports

Contribution ID: **221** Type: **Presentation**

Cloud storage with Seafile at Elettra

Thursday 20 March 2025 11:30 (15 minutes)

Elettra is an multidisciplinary research center running two particle accelerators producing synchrotron light. In 2017 we looked into software that will replace our aging Windows file share server, used for hosting non scientific data. This research led us to Seafile. My presentation will introduce you to our initial reasoning, it will also show you the evolution of our Seafile infrastructure and will guide you through the problems our users have been encountering when using the service.

Author: GREGORI, Iztok (SINCROTRONE-ELETTRA)

Presenter: GREGORI, Iztok (SINCROTRONE-ELETTRA)

Session Classification: Operations and development of CS3 services

Track Classification: Main sessions: CS3 Community Site Reports

Contribution ID: **222** Type: **Presentation**

INFN Cloud service integration: software distribution by leveraging CernVM File System over a CEPH RGW multisite

Thursday 20 March 2025 11:45 (15 minutes)

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 four years INFN Cloud, a production-level, integrated and comprehensive cloud-based set of solutions, delivered through distributed and federated infrastructures.

INFN Cloud offers to its users and collaborations an S3-based Object Storage service for data archiving, on top of a multisite CEPH RGW infrastructure, accessible via a web ui or programmatically.

Taking advantage by the S3-based Object Storage service, the CernVM-File System services have been deployed and integrated with other technologies (such as Vault identity-based secrets and encryption management system and RabbitMQ open-source message broker) to define a user-friendly solution aimed at sharing software and related configuration files, among heterogeneous and distributed resources.

The solution we provide implements an abstraction layer that hides the underlying complexity and allows the final user to easily interact with an S3 object storage interface for distributing software, libraries and related dependencies among different sites, under a common path and with a POSIX access, via the CernVM-File System.

We will describe the main features of our setup, focusing on the integration process of the different services on the INFN Cloud distributed infrastructure.

Authors: ALKHANSA, Ahmad (INFN - CNAF); Dr COSTANTINI, Alessandro (INFN-CNAF); MICHELOTTO, DIEGO (INFN - National Institute for Nuclear Physics); SPIGA, Daniele; DEL CORSO, Francesca; MALATESTA, Giada (INFN); GASPARETTO, Jacopo (CNAF); VERLATO, Marco (Universita e INFN, Padova (IT)); SGARAVATTO, Massimo (Universita e INFN, Padova (IT)); TRALDI, Sergio; STALIO, Stefano

Presenter: Dr COSTANTINI, Alessandro (INFN-CNAF)

Session Classification: Operations and development of CS3 services

Track Classification: Main sessions: CS3 Community Site Reports

Contribution ID: **223** Type: **Presentation**

Seeking Cost-Optimal Infrastructure Size for Distributed File Systems: A Ceph Case Study

Wednesday 19 March 2025 16:15 (15 minutes)

Here, we present a preliminary study to evaluate how hardware configuration choices can affect the performance of a distributed file system.

While it is straightforward to size hardware for intensive computational tasks to achieve a given performance target, the complexity of the I/O hardware and software stack makes it challenging to predict - and even assess [1] - file system performance based solely on hardware specifications.

However, being able to predict overall performance and associated hardware cost is of utmost importance in many cases, like, for instance, large scale sharing platforms (i.e., NextCloud and others) based on distributed storage solutions as their backbone and HPC facilities with I/O intensive workload (i.e., large scale ML training).

Our experiments are conducted using the Ceph file system; the test infrastructure comprises 10 storage nodes, each of them is equipped with 192 GB of RAM, 2 processors with 16 cores each, while the storage comprises 12x22TB HDD for data and 2x14TB NVMe for metadata.

We explore three different HW parameters: number of CPU cores, amount of memory, and disk speed to see how the performances are affected in terms of bandwidth and IOPS of sequential and random read/write operations.

As a benchmarking tool, we use FIO [2], which is run multiple times, adjusting the number of available CPU cores and the memory capacity by means of the Linux hotplug interface while controlling disk I/O speeds through Linux cgroups.

Our preliminary results reveal that for some workloads, increasing hardware resources does not yield proportional performance gains. Surprisingly, sequential I/O operations are not significantly influenced by additional CPU cores or memory, indicating that the lower tiers of the hierarchical storage model do not benefit enough to justify noteworthy resource increases.

In contrast, IOPS-intensive tasks benefit from increased resources (CPUs and RAM).
Finally, performance consistently improves as disk speeds increase, but only up to a certain point; this suggests that the maximum potential performance of the disks is never fully realized in practice.

We are currently performing additional tests and examining different network configurations, including speed, link layer, and LAG settings, to achieve a comprehensive hardware analysis.

[1] Vasily Tarasov, Saumitra Bhanage, Erez Zadok, and Margo Seltzer.
"Benchmarking File System Benchmarking: It IS Rocket Science." Proceedings of the 13th Workshop on Hot Topics in Operating Systems
(HotOS XIII), May 2011. Napa, CA: USENIX Association.

[2] Jens Axboe. Flexible I/O Tester (fio). Version 1.2.0, 2022.

Authors: PASIANOTTO, ISAC; TOSATO, NICCOLÒ; Dr LOT, Ruggero (Area Science Park)

Co-author: Prof. COZZINI, Stefano (Area Science Park)

Presenter: TOSATO, NICCOLÒ

Session Classification: Storage Technology

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **224** Type: **Presentation**

Infrastructure and Practices for Sharing and Disseminating In-Silico Medicine Research Data

Thursday 20 March 2025 18:00 (15 minutes)

A significant amount of research data remains underutilized due to being unpublished or poorly described, leading to a loss of funding and scientific potential. To recover lost data and prevent further waste, researchers must be encouraged to use FAIR data sharing repositories and adopt good publishing practices, such as providing descriptive metadata and utilizing datasets from the community. Since this involves additional complications, the data retrieval and publication process should be simplified, and extra motivation should be provided.
Our solution comprises integration between the Model Execution Environment platform for in-silico model simulations on HPC resources, and open-source data sharing repositories, along with the required infrastructure, including an instance of the Dataverse repository, and a set of practices for its effective use, to enable convenient collaboration within the community. It aims to facilitate data management for execution of medical simulations, provide scientists with tools for cooperation, and engage the scientific community in further advancing research through data contributions. Additionally, the Model Execution Environment leverages HPC resources for the scientists, providing a straightforward interface and structuring complex computational workflows.
The aforementioned practices include rule-based data sharing based on an incentive-driven mechanism, fueling research even after the data becomes public. This approach is embodied by the Sano Dataverse instance, part of the RODBUK Krakow Open Research Data Repository, through the publication of a dataset from the DPValid case study, conducted by UNIBO within the InSilicoWorld project. The publication preparation process, which includes data processing on HPC using the MEE platform, uploading to Sano Dataverse or Zenodo, configuring rule-based data sharing, and ensuring ongoing curation and support, is presented on the basis of the requirements of the ISW scientific community which the authors are part of.
This publication is partly supported by the EU H2020 grants Sano (857533), ISW (101016503) and by the Minister of Science and Higher Education “Support for the activity of Centers of Excellence established in Poland under Horizon 2020” number MEiN/2023/DIR/3796.
We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017022

Authors: Mr ZAJĄC, Karol (Sano Centre for Computational Medicine); ZHYHULIN, Taras

Co-authors: Ms BOTTIN, Francesca (Alma Mater Studiorum - University of Bologna); Dr DAVICO, Giorgio (Alma Mater Studiorum - University of Bologna); Mr STANIC, Goran (Alma Mater Studiorum - University of Bologna); MEIZNER, Jan (Sano Centre for Computational Medicine); MALAWSKI, Maciej (AGH University of Science and Technology); Mr KASZTELNIK, Marek (ACC Cyfronet AGH); BUBAK, Marian (AGH Krakow); NOWAKOWSKI, Piotr (ACC Cyfronet AGH); Mr POŁEĆ, Piotr (ACC Cyfronet AGH)

Presenter: ZHYHULIN, Taras

Session Classification: Data sharing infrastrcutures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **225** Type: **Presentation**

Onedata Distributed Storage for HPC Galaxy Workflows

Friday 21 March 2025 11:45 (15 minutes)

Onedata [1] is a high-performance, distributed data management system designed for global infrastructures. It provides seamless access to heterogeneous storage resources and supports diverse use cases ranging from personal data management to large-scale scientific computations. Leveraging a fully distributed architecture, Onedata facilitates the creation of hybrid cloud environments that integrate private and public cloud resources. The system enables users to collaborate, share, and publish data while supporting high-performance computations on distributed datasets via various interfaces, including POSIX-compliant native mounts, pyfs (Python filesystem) plugins, REST/CDMI APIs, and an S3 protocol (currently in beta).

Recent advancements in Onedata include the development of the *fs.onedatarestfs* Python library, a lightweight *pyfilesystem* client built upon the *OnedataFileRESTClient* library. Within the scope of the EuroScienceGateway project [2], these libraries have been instrumental in integrating Onedata with the Galaxy Project [3] , an open-source platform for data analysis workflows predominantly used in the life sciences. This integration has resulted in a new File Source Plugin and an Object Store for Galaxy. The File Source Plugin enables users to import and export datasets between Onedata and Galaxy, while the Object Store integration allows Onedata to function as a backend storage system for Galaxy datasets. This implementation takes advantage of Onedata's distributed architecture, creating a synergy with Galaxy's distributed network of Pulsar endpoints (workflow execution services). By tracking data distribution, it opens the door to locality-aware, smart workflow scheduling, which can reduce data transfer costs, processing delays, and energy usage.

Onedata is currently deployed in several European projects, including EUreka3D [4], EuroScienceGateway [2], DOME [5], and InterTwin [6]. In these projects, Onedata provides a data transparency layer for managing large, distributed datasets in dynamic, hybrid cloud environments with containerized deployments.

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. https://onedata.org.
2. EuroScienceGateway Project: Open Infrastructure for Data-Driven Research. https://galaxyproject.org/projects/esg/.
3. The Galaxy Project. https://galaxyproject.org/.
4. EUreka3D: European Union's REKonstructed in 3D. https://eureka3d.eu.
5. DOME: A Distributed Open Marketplace for Europe Cloud and Edge Services. https://dome-marketplace.eu.
6. InterTwin: Interdisciplinary Digital Twin Engine for Science. https://intertwin.eu.

Authors: Dr ORZECHOWSKI, Michał (ACK Cyfronet AGH); Mr OPIOŁA, Łukasz (ACK Cyfronet AGH); Dr DUTKA, Łukasz (ACK Cyfronet AGH)

Presenter: Dr ORZECHOWSKI, Michał (ACK Cyfronet AGH)

Session Classification: HPC data access and integration

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **226** Type: **Presentation**

SWAN: Status, Custom Environments, and Cloud Integration

Wednesday 19 March 2025 12:00 (15 minutes)

SWAN (Service for Web-based Analysis) is CERN's cloud-based platform that streamlines scientific data analysis and collaboration by offering users an integrated Jupyter-based environment with seamless access to resources such as EOS/CERNBox and CVMFS. In this talk, we will present the current state of the project and highlight the advancements made in 2024 to address new use cases. These include enabling user-defined custom software environments without requiring pre-built container images, as well as exploring and evaluating diverse approaches to integrate with external clouds and services, enhancing SWAN's flexibility and scalability for modern scientific workflows.

Author: CASTRO, Diogo (CERN)

Presenter: CASTRO, Diogo (CERN)

Session Classification: CS3 Jupyter SIG & Data Science and Visualisation Platforms

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **227** Type: **Presentation**

Introduction to OpenCloud: Team, Key Features and Plans

Thursday 20 March 2025 13:45 (30 minutes)

This talk will provide an overview of OpenCloud, including its team, core features, and future plans. We will describe the platform's architecture, its capabilities for data storage, processing, and collaboration, and present the roadmap for upcoming updates and enhancements. The session will offer insights into how OpenCloud aims to support scientific research and data-driven projects.

Author: BAADER, Tobias (OpenCloud)

Presenter: BAADER, Tobias (OpenCloud)

Session Classification: EFSS Products

Track Classification: Main sessions: File Sync & Share Solutions and Requirement from the Community

Contribution ID: **228** Type: **Presentation**

Reva and CERNBox: developments and prospected evolution

Thursday 20 March 2025 10:10 (15 minutes)

In this contribution we will touch on the recent developments and the operational experience running Reva as part of CERNBox, the CERN cloud storage, and on our plans to support Reva and the CS3APIs for the community.

Having the CS3APIs reached a good level of maturity, in the past year we consolidated the Reva implementation and improved its dependability, in particular with respect to sharing, on one hand, and to supporting the two main storages, EOS and CephFS, on the other.

On the operations side, we will share our experience in integrating external applications, and we show how we intend to provision storage with different QoS fulfilling the multiple requirements of our diverse user community.

We will conclude with an outlook on the CS3APIs governance and on our plans to evolve CERNBox as a sync & share ecosystem for the community.

Authors: CASTRO, Diogo (CERN); Dr LO PRESTI, Giuseppe (CERN); GEENS, Jesse

Presenters: Dr LO PRESTI, Giuseppe (CERN); GEENS, Jesse

Session Classification: Operations and development of CS3 services

Track Classification: Main sessions: CS3 Community Site Reports

Contribution ID: **229** Type: **Presentation**

Future of Active Archive Data Storage: Striking the Right Balance Between Performance and Energy Efficiency

Wednesday 19 March 2025 16:00 (15 minutes)

Learn how Host-Managed Shingled Magnetic Recording (HM-SMR) drives and selective write-grouping can transform software-defined storage (SDS) environments. Selective write-grouping and Popular Data Concentration (PDC) both work with Shingled Magnetic Recording (SMR) and Conventional Magnetic Recording (CMR) disks using erasure coding. By restricting write operations to fewer drives, selective write-grouping lets the remaining drives power down, reducing energy consumption by up to 43% in SMR configurations. This method also eliminates the PDC "staging"phase, enabling immediate data partitioning and placement within designated groups. Join us to explore how to effectively manage power usage, scale capacity, and maintain high performance in modern Active Archive Data Storage.

Author: Mr MODRZYK, Piotr (Leil Storage)

Presenter: Mr MODRZYK, Piotr (Leil Storage)

Session Classification: Storage Technology

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **230** Type: **Presentation**

Current state of collaborative storage in OpenCloud

Wednesday 19 March 2025 17:35 (15 minutes)

OpenCloud comes with a never before seen level of integration with the underlying storage system. It allows transparently accessing files on a posix filesystem. The new driver allows users to either work with OpenCloud or use external tools that directly access files on the storage. Local filesystems as well as enterprise network filesystems have already been integrated. OpenCloud picks up changes in real time and notifies all clients that have access. In this presentation we will show the available options and explain some of the technical details.

Authors: Dr DREYER, Jörn (OpenCloud GmbH); FREITAG, Klaas

Presenter: Dr DREYER, Jörn (OpenCloud GmbH)

Session Classification: Features & Principles

Track Classification: Main sessions: Technology & Research

Contribution ID: **231** Type: **Presentation**

SCION ScienceDMZ: now with FTS integration!

Thursday 20 March 2025 17:30 (15 minutes)

In today's research landscape, managing and processing a high volume of data has become crucial in many fields. Many researchers make use of remote computing resources to process large data volumes. High-volume data transfers between research institutions and High-Performance Computing Clusters (HPCC) have thus increased in importance, as large data sets can require hours or days to transmit in non-optimized settings. Moreover, adhering to security compliance requires the use of firewalls, which are often costly and / or slow down data transfers.

To resolve these issues, we have developed Hercules and LightningFilter, which make use of the SCION next-generation Internet to achieve security and efficiency. The Hercules data transfer application provides a high-speed implementation, offering sustained transmission and reception speeds of around 100 Gbps including reliable delivery as well congestion control. LightningFilter is an open-source firewall implementation, which can process minimum-sized packets in excess of 100 Gbps on a standard mid-range server. LightningFilter can satisfy firewall compliance rules, and enables ASes to cryptographically verify, restrict, and police the incoming connections, whether from other ASes or specific hosts, allowing the HPCC to implement distinct rate limits for different universities while ensuring a guaranteed throughput for particular hosts. The open-source implementation of these tools facilitates a low-cost yet high-performance file transfer service.

A key component of this architecture is the deployment of data transmission nodes, which play a crucial role in optimizing data flow. These nodes, strategically positioned within the network, facilitate high-speed data transfers and ensure reliable connectivity between the HPCC and researches.

A new development in this infrastructure is the integration of Hercules with the File Transfer Service (FTS) through the gfal2 library. This integration streamlines the data transfer process, enhancing efficiency and reliability. By leveraging the gfal2 library, data transfers can be integrated into existing data processing pipelines, bridging the gap between diverse systems and technologies, and allowing for more flexible and robust data handling capabilities.

We are excited to present the latest advancements in SCION-based Science DMZs and share insights from further deployments and proofs of concept, highlighting the tangible benefits this infrastructure offers to the research community.

Author: WIRZ, Francois (ETHZ)

Presenter: WIRZ, Francois (ETHZ)

Session Classification: Data sharing infrastrcutures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **232** Type: **Presentation**

Using Federated Data Infrastructure for a European Open Web Index

Thursday 20 March 2025 17:15 (15 minutes)

In an era where web search serves as a cornerstone driving the global digital economy, an open, impartial and transparently produced web index is a key opportunity for Europe and beyond. Currently, the landscape is dominated by a select few gatekeepers who provide their web search services with minimal scrutiny from the general public. Moreover, web data has emerged as a pivotal element in the development of AI systems, particularly Large Language Models. The efficacy of these models depends upon both the quantity and quality of the data available. Consequently, restricted access to web data and search capabilities severely curtails the innovation potential, particularly for smaller innovators and researchers who lack the resources to manage petabyte platforms.
In this talk, we present the OpenWebSearch.eu project which is currently developing the core of a European Open Web Index (OWI) as a basis for a new Internet Search in Europe. We mainly focus on the setup of a Federated Data Infrastructure leveraging geographically distributed data and computing resources at top-tier supercomputing centres across Europe. This data infrastructure leverages MINIO/S3, iRODS, EUDAT (B2SAFE, B2HANDLE) and our previous work on the LEXIS Platform for distributed computing and data management. The system developed facilitates efficient execution of complex processing and indexing workflows.

Author: HAYEK, Mohamad

Co-authors: WAGNER, Andreas (CERN); MARTINOVIČ, Jan (VSB - Technical University of Ostrava); MANKINEN, Katja (CSC –IT Center for Science); GOLASOWSKI, Martin; SHARIKADZE, Megi; GRANITZER, Michael; Ms FATHIMA, Noor Afshan (CERN); ZERHOUDI, Saber (University of Passau); HEINEKING, Sebastian (Webis-Group); HACHINGER, Stephan (Leibniz Supercomputing Centre (LRZ) of the BAdW)

Presenter: HAYEK, Mohamad

Session Classification: Data sharing infrastrcutures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **233** Type: **Presentation**

Sharing across sync-and-share services made easy

Wednesday 19 March 2025 14:45 (15 minutes)

In the current implementation of federated sharing between OCM-compliant sync-and-share services require the exchange of Federated Cloud IDs by end users. Now although this approach works, it is not very user-friendly. During the course of the EU-funded CS3MESH4EOSC project a so-called invitation workflow has been implemented in golang based on REVA-based sync-and-share services like CERNbox. What we have done is that we have implemented this invitation workflow in php so that it will work for both Owncloud10 as well as Nextcloud. Our goal here is to make sharing data with users of remote sync-and-share services just as easy for end-users as sharing data with local users.

Authors: Mr PRINS, Antoon (SURF); TROMPERT, Ron

Presenter: TROMPERT, Ron

Session Classification: OCM CS3 SIG & Federated Infrastructures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **234** Type: **Presentation**

Storage, Data, AI in the fast changing world

Friday 21 March 2025 09:00 (1 hour)

Presenter: Dr KOESTER, Axel (IBM Storage Chief Technologist)

Session Classification: Keynote

Track Classification: Keynotes

Contribution ID: **235** Type: **Presentation**

FAIR Data Management –Current Requirements, Best Practices and Opportunities

Friday 21 March 2025 10:00 (15 minutes)

From the perspectives of different data (re-)use cases (from University Library, Research Funding Support, Super Computing Centre and LMU Physics Department), this talk will focus on the many aspects of FAIR data. In practice, data should be handled in accordance to the FAIR (Findable, Accessible, Interoperable, Reusable) principles –but what does this mean in scientific day-to-day work? Starting with the framework provided by research funders and scientific journals, we will explore what makes your data FAIR and how you can also benefit from having your data FAIRified. In this context, the role of data management plans, metadata, data publication and Open Data will show, how inseparable data driven research and FAIR are linked. To wrap it up, helpful tools and infrastructures are shown as well as best practice examples from how large datasets are handled by LMU Munich and LRZ.

Authors: Dr SCHRECK, Florian (Ludwig-Maximilians-Universität München, Unit for Research Funding); MEIER, Laura (Ludwig-Maximilians-Universität München, University Library); Dr SPENGER, Martin (Ludwig-Maximilians-Universität München, University Library); ALDENHÖVEL, Pauline (Ludwig-Maximilians-Universität München, University Library); GEBHARDT, Stefan (Ludwig-Maximilians-Universität München, University Library)

Co-authors: WELLMANN, Alexander (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities); MUNKE, Johannes (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities); REDL, Robert (Ludwig-Maximilians-Universität in Munich, Faculty of Physics, Munich, Germany); HACHINGER, Stephan (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities)

Presenters: Dr SCHRECK, Florian (Ludwig-Maximilians-Universität München, Unit for Research Funding); Dr SPENGER, Martin (Ludwig-Maximilians-Universität München, University Library)

Session Classification: FAIR Data Management

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **236** Type: **Presentation**

To Open up Data in Sync-and-Share

Friday 21 March 2025 10:15 (15 minutes)

We have been running sync-and-share services for about 11 years now and there has been a recurring question from our users is the ability to "park" data from finished projects to somewhere else. Somewhere else often also means a place where others can find it. For this reason we have developed SURF Research Data Connector (SRDC). This is a service sitting between the sync-and-share service and a data repository or data publication service. SRDC is based on plugins for the various repository and publication solutions. We have implemented but exporting data and metadata to a repository as well as importing data from a repository into the sync-and-share service. In addition we have also implemented importing open data where users only have to supply a DOI url.

Currently, we have implemented plugins for figshare, iRODs, Dataverse, OSF, Zenodo and the two local Dutch repo's SURF ShareKit and 4TU.ResearchData. Currently, work is ongoing using SRDC to offload data to object-locked buckets in object storages to ensure immutability of the data.

SRDC has been implemented for Owncloud10 as well as Nextcloud.

Authors: Mr TROMP, Dave (SURF); TROMPERT, Ron

Presenter: Mr TROMP, Dave (SURF)

Session Classification: FAIR Data Management

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **237** Type: **Presentation**

Research Data Management for Huge Datasets: ”Cloudifying” Data not on the Cloud

Friday 21 March 2025 10:30 (15 minutes)

Data repositories play an essential role in Research Data Management according to the “FAIR principles”(Wilkinson et al. 2016) and leave less and less to be desired. However, they usually cannot accommodate huge datasets towards the PB range, as they e.g. come from supercomputing - for technical, financial or organisational reasons. In fact, for such datasets even movement to an external (paid) cloud storage is often not possible due to the size or costs.
At Munich, the University Libraries and the Leibniz Supercomputing Centre are collaboratively aiming at providing basic FAIR RDM also for such huge datasets. As these datasets are usually produced by user groups with technical excellence, the FAIR solution for them can arguably depend on IT skills. However it should not impose a strong data-lifecycle management or control,as these users often have their own ideas and project (or institutional) policies that shape the data management concept.
We therefore aim at a minimally-invasive approach, where users add metadata as YAML “sidecar files” (with DataCite-compliant contents) to their huge datasets and have those published manually via one of several portals.
At the LRZ, a python-scripted workflow has been prototyped to then push these metadata into the InvenioRDM repository framework, generating a DOI and a landing page. In this usage scenario, InvenioRDM is used purely as a metadata-publication frontend, while the data remain on back-end “Big Data” storage and are linked from the metadata.
We have implemented this concept with two InvenioRDM instances / demonstrators, one in collaboration with LMU University Library and Physics Department (”Open Data LMU - Physics”) and one for LRZ in general (”LRZ FAIR Data Portal”). These two use different approaches for making the actual data available.
Open Data LMU - Physics uses the iRODS ‘’data-grid middleware” and a web-based frontend to make the data openly available, while the LRZ FAIR Data Portal relies on GLOBUS and its data-transfer middleware. These approaches show how data-cloud-like functionality can be implemented on top of classical data storage, which is a central point in making huge datasets FAIR.

Authors: WELLMANN, Alexander (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities); MUNKE, Johannes (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities); Dr HACHINGER, Stephan (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities)

Co-authors: Dr SPENGER, Martin (Ludwig-Maximilians-Universität München, University Library); Dr REDL, Robert (Ludwig-Maximilians-Universität München, Faculty of Physics)

Presenter: WELLMANN, Alexander (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities)

Session Classification: FAIR Data Management

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **238** Type: **Presentation**

Bridging HPC and Cloud: Deploying an S3 Frontend for POSIX-compliant Storage

Friday 21 March 2025 11:30 (15 minutes)

As research communities increasingly rely on cloud-native tools and workflows, integrating High-Performance Computing (HPC) environments with cloud storage has become increasingly important. From the perspective of an IT support team, this presentation outlines our initial approach towards a lightweight, easily deployable S3-layer on existing POSIX-compliant storage, enabling scientists to leverage cloud-native tools while collaborating with backend storage users. A key driver for this work is also the need to share very big data with external collaborators who lack direct access to the internal storage systems. By providing an S3 interface to existing storage, we can facilitate secure and controlled data sharing across institutional boundaries, enabling researchers to work together more effectively.

We discuss the challenges and first solutions for mapping S3 objects to POSIX files, ensuring seamless data access and changes via both file system and S3. We look into existing open-source implementations like the Versity S3 Gateway with the aim to provide read/write functionality on supported backend storage systems while addressing security considerations, user mapping, and compliance to the S3 standard. We will share our first experiences with running workflows of the reproducible analysis platform (REANA) on our initial setup as well as future directions for this project, which aims to bridge the gap between HPC and cloud storage, ultimately enhancing the productivity and collaboration of researchers.

Author: REDL, Robert (Ludwig-Maximilians-Universität in Munich, Faculty of Physics, Munich, Germany)

Co-authors: SACCHI, Elena (Leibniz-Institut für Astrophysik Potsdam (AIP), Potsdam, Germany); KHALATYAN, Arman (Leibniz-Institut für Astrophysik Potsdam (AIP), Potsdam, Germany); RAUSCHER, Felix (Ludwig-Maximilians-Universität in Munich, Faculty of Physics, Munich, Germany)

Presenter: REDL, Robert (Ludwig-Maximilians-Universität in Munich, Faculty of Physics, Munich, Germany)

Session Classification: HPC data access and integration

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **239** Type: **Presentation**

A web portal for hydrodynamical, cosmological simulations

Wednesday 19 March 2025 12:15 (15 minutes)

Since 2017, the cosmo sim web portal allows accessing and sharing the output of large, cosmological, hydro-dynamical simulations with a broad scientific community and contentiously grows in services and data which are made available. It is based on a multi-layer structure: a web portal, a job control layer, a computing cluster and a HPC storage system. The outer layer enables users to choose an object from the simulations. Objects can be selected by visually inspecting 2D maps of the simulation data, by performing highly compounded and elaborated queries or graphically by plotting arbitrary combinations of properties. It also allows users to receive related scientific data products by directly processing the raw simulation data on a remote computing cluster.

Author: DOLAG, Klaus

Presenter: DOLAG, Klaus

Session Classification: CS3 Jupyter SIG & Data Science and Visualisation Platforms

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **240** Type: **Presentation**

Sharing Field Campaign Data using IPFS

Thursday 20 March 2025 18:15 (15 minutes)

In the recent atmospheric and oceanic measurement campaigns (EUREC4A and ORCESTRA), we use the InterPlanetaryFileSystem (IPFS) to store, use, synchronize and share measurement data. IPFS uses content addressing (instead of location-based addressing) and provides an easy-to-set-up peer-to-peer network for sharing data on servers and portable devices.

Because of these features, we were able to immediately start collaborating on newly collected data using local network connections while in the field. At the same time, the data was continuously synchronized with external data centers so that it was also available to outsiders. Since the data is addressed globally by content (CID), data access remains unchanged regardless of where the data is stored or analyzed. Scripts using the data could therefore be exchanged immediately between participants on site and remotely, and of course still work after we have dismantled our infrastructure in the field.

We plan to continue to build on our practical experience and develop more tools and user interfaces around the current setup. This should eventually lead to a replacement of the currently used data management platform, which was previously used to store the data recorded by the research aircraft.

Author: KÖLLING, Tobias (MPI für Meteorologie Hamburg)

Co-author: Mr BRÖTZ, Björn (Deutsches Zentrum für Luft- und Raumfahrt)

Presenters: Mr BRÖTZ, Björn (Deutsches Zentrum für Luft- und Raumfahrt); KÖLLING, Tobias (MPI für Meteorologie Hamburg)

Session Classification: Data sharing infrastrcutures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **241** Type: **Presentation**

Report on the CS3 Jupyter Workshop

Wednesday 19 March 2025 11:30 (30 minutes)

On May 15 2024, about 30 participants from our community joined an online workshop about JupyterHub. Among them was also the project lead for the JupyterHub project. Seven presentations were given and after each one the participants were able to discuss and ask questions. The workshop was concluded with a longer community discussion.

This presentation will highlight some key take-aways from the workshop and give some examples of what was discussed.

Author: NORDIN, Micke (SUNET)

Presenter: NORDIN, Micke (SUNET)

Session Classification: CS3 Jupyter SIG & Data Science and Visualisation Platforms

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **242** Type: **not specified**

Report on the CS3 OCM Workshop

Wednesday 19 March 2025 14:00 (30 minutes)

On Nov 20 2024 about 12 participants and developers from our community joined an online workshop about Open Cloud Mesh. On top of the three presentations, a lively technical discussion took place to address issues such as security and discovery in a federated cloud environment.

This presentation will highlight the main take-aways and serves as an introduction for the Campfire session about OCM.

Author: Dr LO PRESTI, Giuseppe (CERN)

Presenter: Dr LO PRESTI, Giuseppe (CERN)

Session Classification: OCM CS3 SIG & Federated Infrastructures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **243** Type: **Presentation**

EOSC EU Node contributionto sync & share services in EU

Wednesday 19 March 2025 14:30 (15 minutes)

In our talk at CS3 2025 we will discuss the EOSC EU Node-provided sync & share service and its contribution to the horizon of the sync & share services for research and academia in Europe.

EOSC EU Node is a set of data-centric cloud services supporting Open Science, Open Data and FAIR. The node is owned by European Commission and implemented by a group of contractors. PSNC is the main contractor, and collaborates with Safespring (Sweden), Owncloud/Kiteworks (Germany), Nordunet (Denmark), Sikt (Norway), SUNET (Sweden), CESNET (Czechia) and EGI (Netherlands).

EOSC EU Node provides application-level, user-facing and data-centric services including: File Synchronisation and Sharing Service, Interactive Notebook Service and Large File Transfer Service. I also provides infrastructure and platform-level services including orchestrated Virtual Compute Infrastructure, Container Platform and massive data transfer, targeted to a more advanced users.

Among the services provided in EU Node the Manage File Synchronisation and Sharing Service is of special interest of the CS3 community. We believe it provides an interesting contribution to the spectrum of the sync & share services for research and academia in Europe provide by European R&D and academic community.

EU Node's sync & share service is implemented based on ownCloud/Kiteworks OCIS platform integrated with Geant-provided federated AAI service. It offers basic ownCloud functionality along with extra integrations including collaborative document editing platform based on Collabora and storage integration with Interactive Notebook Service. Users of Jupyter Notebooks provided within EOSC EU Node in parallel to ownCloud service may use their data storage space in ownCloud, in order to keep, store and access the Jupyter kernel codes as well as compute and analysis input/output data. ownCloud data sharing capability can also be used for collaborative computing.

During our talk at CS3 2025 we will present the details of the ownCloud/Kiteworks OCIS platform deployment on EOSC EU Node, its key feature, provided in response to EC requirements.

We will also project on possible future developments and extensions. Among these, OCM usage will be discussed. ownCloud OCIS platform supports OCM and thus enables cross-site data-based collaboration among users of EOSC EU Node and other sync & share platforms. In our talk we will discuss current and future possibilities related to the integration of EU Node with other sync & share services provided withing CS3 community and elsewhere in Europe that will enable and facilitate data-centric, distributed and collaborative data collection, improvement, analysis and research.

Author: BRZEZNIAK, Maciej

Co-authors: MANZI, Andrea; Mr PAUES, Gabriel (Safespring AB); ABEN, Guido (SUNET); DYROFF, Holger; BLONIARZ, Krzysztof; WADÓWKA, Krzysztof (PSNC); FLORIO, Licia; ZIMNIEWICZ, Michał (Poznan Supercomputing and Networking Center); MEYER, Norbert (Unknown); SUSTR, Zdenek (Czech Technical University (CZ)); BŁAŻEWICZ, marqs

Presenter: BRZEZNIAK, Maciej

Session Classification: OCM CS3 SIG & Federated Infrastructures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **244** Type: **Presentation**

R(evolution) in storage back-ends for sync & share services

Wednesday 19 March 2025 15:45 (15 minutes)

In our talk at CS3 2025 we will discuss the (r)evolution in the storage and network back-ends for cloud-based data centric services with a special focus on the sync & share service as a storage killer application.

We will review various storage back-ends for cloud platforms along with their performance, scalability and maintenance complexity and economical aspects. We will also touch on the new trends in the cloud platforms that convert into hyper-convergent, fully software-defined and open source-based systems, where compute, storage and network components act together to provide flexible, functional, scalable, reliable and high-performance platform that is cost-effective, easy to implement, maintain, automate, monitor and optimize.

In face of the AI revolution and overall shift of economy, science, research and education towards data-centric applications and services we observe increasing requirements vs cloud platforms that help dealing with large data sets. At the very infrastructure level this means a growing pressure on performance (IOPS, GB/s) and economical scalability. We can also observe that 'classical'storage technologies: disk arrays, Fibre Channel-based SAN networks etc. become obsolete and legacy as they lack flexibility, do not support programmatic orchestration widely, and are relatively costly to purchase, implement and maintain. Their usage requires specialised knowledge and can lead to vendor lock-in situations. On the other end of technology spectrum, commodity technologies are rapidly developed, breaking the next barriers of reliability and performance. Growing popularity and improving economic affordability of flash storage systems (SSD, NVMe) enables advancing the cloud platforms and applications performance so that they address today's IO requirements. Also recent development of Ethernet enables using this protocol as a carrier for I/O traffic. RoCE is in particular useful to implement RDMA in Ethernet networks that are widely spread in today' s IT infrastructure.

These two advancements in storage and network technology support for reliability, scalability and economical efficiency, re-use of existing knowledge as well as compute, storage and network platform integration and simplification. Among the new technologies NVMeoF gains the momentum, with increasing number of vendors offering NVMeoF products. Also the increasing adoption of software defined networks (SDN) and cloud software stacks supporting SDN facilitates compute, storage and network resources provisioning, orchestration and automation.

In our presentation we will discuss the PSNC cloud computing, storage and network platform with a special focus on usage of SSD, NVM, RoCE technologies and usage of SDN. PSNC platform is based of Openstack, OKD compute components and both specialised (storage arrays, NAS appliances, HDDs, SSD/NVMe) and software-defined storage components (Ceph on HDD and SSD/NVMe), software defined network (SDN) as well as legacy and modern storage networks (Fibre Channel, iSCSI, Ethernet, NVMeoF). The platform provides storage, compute and network resources for our sync & share systems including country-wide box.pionier.net.pl service offered to academic and research community in Poland since 2015, based on Seafile software as well as our EOSC EU Node ownCloud/Kiteworks OCIS-based sync & share service devoted to EOSC users.
Seafile and ownCloud/Kiteworks OCIS are interesting and challenging benchmark applications to storage systems, back-end networks and compute infrastructure components due to their extreme IO requirements. In our presentation we will share information on current setup of these applications implemented at PSNC as well as make projections on future improvements of our sync & share services compute, storage and network back-ends.

Authors: PRYCKI, Adam; POKORA, Eugeniusz; BRÓŹDZIAK, Jan (PSNC); BLONIARZ, Krzysztof; WADÓWKA, Krzysztof (PSNC); BRZEZNIAK, Maciej

Presenter: BRZEZNIAK, Maciej

Session Classification: Storage Technology

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **245** Type: **Presentation**

Open-source Windows storage with CERNBox

Wednesday 19 March 2025 16:30 (15 minutes)

We report on the new, CERNBox-based stack, incorporating both Samba and CephFS, which is now providing Windows storage at CERN.
For nearly 30 years, the Distributed File System (DFS) ecosystem served as the main data storage platform for the Windows operating system at CERN. With a new demand for collaboration and access from anywhere worldwide, CERNBox has become the natural candidate to replace DFS.
The replacement process was planned in two phases. The first included Windows home directories for each user account at CERN. The second is presently underway and involves large shared spaces called projects.
This presentation will discuss the differences between native DFS and that offered by the CERNBox service, and the architectural choices that have made CERNBox a viable option for the Windows operating system community at CERN.

Author: BUKOWIEC, Sebastian (CERN)

Presenter: BUKOWIEC, Sebastian (CERN)

Session Classification: Storage Technology

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing

Contribution ID: **247** Type: **Presentation**

Collective infrastructure makes you stronger

Wednesday 19 March 2025 15:00 (15 minutes)

In the GN5-2 project The GÉANT Association will develop investment proposals for three service concepts, with a large, long-term expected impact that may require significant future investment and commitment from the NREN community. Each service concept will only succeed if a concerted collective effort is made, over time. The presentation will present each concept at high level, explain the reasoning and hypothesis behind each and explain the process that will be followed to produce the investment proposals which would -hopefully- lead to a significant increase of community capability over time.

The work has just begun, so this is a key opportunity for the wider community to influence its direction.

These concepts are:

Common PaaS Cloud Middleware for Integrated Trusted Research - A common baseline suite for Platform as a Service cloud research data environments, integrating NREN services such as T&I and security by design, deployable to any Infrastructure Cloud (commercial or community), offering Virtual/Trusted Research Environment functionality to end -users, anchored in the R&E trust ecosystem and with federating capabilities.

Data movement infrastructure for research datasets - A research data transfer infrastructure aims to make large transfers of significant size (i.e. sub-CERN, 0.5 TB < X < 1PB) trivial for any researcher, by integrating available tooling and protocols into a coherent, widely deployed, easy-to-use and secure infrastructure that goes where the network goes. It simplifies many of the challenges around point-to-point data transfers, allowing users to move on from bits and files to managing datasets. This solution will fully leverage the available network capabilities to all researchers.

Pan-European, Sovereign Object Storage for Research Data - This topic will investigate a pan-European infrastructure for research data storage. Baseline requirements include a collectively procured, built and managed technology platform that offers simple object-storage interfaces, deployed in a distributed way that satisfies digital sovereignty requirements. This targets research data repositories with the promise of providing better and cheaper low-level storage by leveraging European-size economies of scale. Such storage will manage technical complexity with great potential savings by unlocking the benefits of building a collective, European infrastructure while recognising the national character of most sovereign data storage infrastructure efforts.

Author: MEIJER, Jan

Presenter: MEIJER, Jan

Session Classification: OCM CS3 SIG & Federated Infrastructures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.

Contribution ID: **248** Type: **not specified**

CS3 Site Reports Summary

Thursday 20 March 2025 09:50 (20 minutes)

Session Classification: Operations and development of CS3 services

Contribution ID: **249** Type: **not specified**

Modern Storage Technologies for the AI era

Wednesday 19 March 2025 10:30 (1 hour)

In the era of data-driven innovation, modern storage technologies are pivotal in addressing the exponential growth of information and the demand for higher performance, efficiency, and scalability. This presentation explores key advancements shaping the storage landscape, including on-premise/cloud storage tiering, computational storage, archival storage, and ultra-fast NVMe SSDs. In this presentation I will provide a comprehensive overview of these technologies, highlighting their roles in modern IT ecosystems and their synergistic impact on optimizing data workflows.

Presenter: Prof. PARADIES, Marcus (LMU)

Session Classification: Keynote

Contribution ID: **250** Type: **Presentation**

Providing a frontend for file sync and share solutions integrated into end-to-end researcher workflows - RSpace & owncloud

Friday 21 March 2025 10:45 (15 minutes)

In the era of data-intensive research and data science, the challenge of managing research data effectively while ensuring FAIR principles isn't just a technical problem—it's a collaborative one. This presentation explores how the needs of researchers, research software providers, and research IT can be addressed together by a commitment to vertical interoperability between research tools and infrastructure to provide end-to-end workflows across tools and research phases. As an example, we will show how a file sync and share solution commonly provided by research infrastructure, Owncloud, can be integrated with a generalist research tool, RSpace, to seamlessly facilitate FAIR data management as part of everyday workflows.

RSpace is a generalist solution for the active phases of research that provides researchers with an electronic laboratory notebook and sample management solution interoperable with a variety of research tools and services. The extended RSpace ecosystem thereby covers most phases of the research data lifecycle from planning to publishing data and enables end-to-end solutions integrated into everyday workflows. Additionally, RSpace integrates with IT infrastructure for e.g. storing, sharing and managing digital assets, which often pose usability challenges for researchers and frustration for the IT organisation when services are not adopted or not properly used. RSpace addresses these problems by providing a frontend for managing data in such filestore solutions and integrating these into typical researcher workflows.

In 2024, RSpace extended their integration with the iRODS file management solution, so that users can directly store the data they collect in everyday workflows in iRODS. The ongoing development phase focusses on metadata exchange between iRODS and RSpace to further improve the robustness of links to remote locations in RSpace documents as well as increase the discoverability and manageability of files in an institutional file store through metadata contextual to the research. This general idea is now being transferred to other file management solutions.

Many European research infrastructures use Owncloud for collaborative file storage, which has recently been prominently adopted by the first EOSC node. SUNET and RSpace are currently collaborating on extending the basic Owncloud functionality already integrated in RSpace in analogy to the iRODS approach. The goals are to increase the usage of managed file store solutions, improve the persistence of links to Owncloud files to prevent link-rot, and to allow efficient metadata exchange between Owncloud, RSpace and the RSpace ecosystem of tool integrations.

Besides reporting on the progress for the development of these integrations, we'll explore how a collaboration between infrastructure and tool providers addresses adoption challenges in research data management (RDM) by creating a unified frontend that seamlessly connects researchers' daily workflows with institutional storage solutions such as iRODS and owncloud.

Authors: ABEN, Guido (SUNET); Mr MACNEIL, Rory (ResearchSpace); MATHES, Tilo

Presenters: ABEN, Guido (SUNET); Mr MACNEIL, Rory (ResearchSpace); MATHES, Tilo

Session Classification: FAIR Data Management

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data

Contribution ID: **251** Type: **not specified**

SIG-CISS session

Friday 21 March 2025 16:10 (1h 50m)

More information: https://wiki.geant.org/display/CISS/15th+SIG-CISS+meeting

Presenter: REALE, Mario (GEANT)

Session Classification: Co-located GEANT SIG-CISS Meeting