CS3 2021- Cloud Storage Synchronization and Sharing

Monday, 25 January 2021 - Thursday, 28 January 2021

Book of Abstracts

Contents

A Proposal for OCM Governance 1

Brainstorming - OCM in EFSS and beyond (CANCELLED) 1

Building solutions for Sensitive Data Projects 1

CERNBox: beyond 2020 1

CS3 Site Reports Summary 2

CS3MES4EOSC - creating a mesh for Open Science 2

Campfire discussion 2

CodiMD in CERNBox: leveraging the WOPI protocol to provide collaborative markdown editing 3

Collabora Online: Secure, on-premise collaborative editing 3

Conclusion 3

Cubbit Hive: the private distributed cloud 4

Data access and integration challenges in a distributed computational environment for medical research 4

Data encryption and crypto-flexibility in collaboration 5

Describo and RO-Crate - the FAIR data research helpers 5

Digital Sovereignty with email & calendars - Is the client the bottleneck? 6

Distribution of container images: From tiny deployments to massive analysis on the grid 6

EOS-wnc: EOS Client for Windows 7

European Open Science Cloud: A single European market for research data and services 8

Evolution of the CS3APIS and the Interoperability Platform 9

FAIR and Open Data: challenges and opportunities 9

Forget 2020: Building an integrated AARNet Cloud Ecosystem in 2021 9

GPFS – protocols and watchfolder function 10

HIFIS: Sync&Share Federation for Helmholtz . . . 10

How to join the Science Mesh? . . . 11

Introducing an open source Sync and Share solution in Sweden . . . 11

JupyterLab for Earth Observation applications with HTCondor scaling and Voilà dashboarding . . . 11

Making Reva talk to EOS: ultimate scalability and performance for CERNBox . . . 13

National Data Storage's synergies with sync & share systems . . . 13

Nextcloud - State of the nation . . . 14

OCM - from the initial design to a core component of Nextcloud Hub . . . 14

On-premise or in the cloud? . . . 14

OnlyOffice and Collabora Online Experience at CERN . . . 15

Open Heart Surgery with 60 Terabytes of Data: Migrating 40.000 users from PowerFolder to Nextcloud . . . 15

Oracle Data Science Cloud . . . 16

Progress of Sciebo Research Data Services . . . 17

Q/A . . . 17

Q/A and Feedback . . . 18

Q/A discussion . . . 18

REVA on CephFS . . . 18

Running Parallel, Distributed ROOT Analysis with PyRDF on Public Cloud - AWS Lambda Case Study . . . 18

SWAN, Rucio, and Jupyter . . . 19

Samba and CERNBox: experience in providing HA online access to Windows-based users 20

Say hello to "SeaTable", the next generation spreadsheet . . . 20

Scalable Metadata Management Using Onedata and OpenFaaS . . . 21

Science Mesh beyond science – perspectives for adoption in a wider business context. . 22

Science Mesh for site administrators: operation, security, trust . . . 22

Science Mesh in a nutshell . . . 23

ScienceMesh for developers: how to contribute to CS3APIs and IOP . . . 23

ScienceMesh for users: applications, use cases & workflow demo . . . 23

SeaTable: easy like a spreadsheet, powerful like a database . . . 23

Seafile and OCM 24
Seafile: Review and Next Steps 24
Site Report sciebo 24
Sync and Share for Research Data Management 24
The OCM test suite 25
The State of OCM 25
The benefits of using Nextcloud Talk 25
The ocis storage driver - a deconstructed filesystem 26
The triangle of digitalization – sustainability and democracy within digital collaboration 26
To Quality, and Beyond! 26
Towards push notifications in OCIS 27
Welcome 27
Welcome & Warm-up 27
Welcome and objectives 28
iRODS Policy Composition: Configuration, Not Code 28
ownCloud - News and Roadmap 28

OCM Workshop / **193**

A Proposal for OCM Governance

Corresponding Author: peter.szegedi@oracle.com

OCM Workshop / **191**

Brainstorming - OCM in EFSS and beyond (CANCELLED)

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

FAIR and Open Research Data Services / **147**

Building solutions for Sensitive Data Projects

Author: Robert Pocklington[1]

[1] *AARNet*

Corresponding Author: robert.pocklington@aarnet.edu.au

As the world shifts evermore to online, real-time and collaborative environments which span countries, continents and cultures, the challenge of security in online projects and research is becoming more important each year. Projects which contain sensitive data, whether it be ecological, medical, financially or commercially sensitive, introduce specific challenges for private companies and research institutions in order to manage and maintain control of their data.

These projects may also include specific requirements such as legal compliance, data sovereignty, institutional process standards or ethical requirements about managing, tracking and disposing of the projects which leverage sensitive data.

In this presentation we will uncover how AARNet is engaging, developing and validating our proof of concept for a Sensitive Data platform which universities and research institutions can leverage to collaborate, perform data analysis and manage their sensitive project data with customised security features and workflows to support the needs of these projects, while delivering an improved user experience with the upcoming version of OwnCloud.

Site reports / **164**

CERNBox: beyond 2020

Author: Samuel Alfageme Sainz[1]

Co-author: Hugo Gonzalez Labrador [1]

[1] *CERN*

Corresponding Author: samuel.alfageme.sainz@cern.ch

2020 was a very challenging year for everyone.

Many of us have had to adapt to a completely new reality, as working from home made its way into our lives. Since the very beginning, our services have been in the spotlight to make this new way of working possible for everyone.

This fact has translated to increased responsibility of being resilient and trustworthy for all our users, now distributed across the globe and heavily relying on the data stored in CERNBox to keep up with their work.

In these rough times, we do not only need to keep the lights on. But doing it while focusing on improving the existing user experience, the agility with which we can deploy changes in our services as well as coming up with new applications and ways of bringing us together thanks to technology.

Site reports / **174**

CS3 Site Reports Summary

Corresponding Author: steiger@id.ethz.ch

EOSC & Federations: Future research infrastructures / **159**

CS3MES4EOSC - creating a mesh for Open Science

Authors: Pedro Ferreira[1]; Jakub Moscicki[1]

[1] *CERN*

Corresponding Author: pedro.ferreira@cern.ch

Over the last decade, several CS3 services at universities, research institutions and NRENs have reliably provided data storage and collaborative services to the Scientific Community, normally through Open-Source on-premise solutions. Examples of such CS3 sites include CERNBox, SWITCHdrive, PSNCBox, SURF's ResearchDrive, DeiC's ScienceData, AARNet's CloudSTOR, WWU's Sciebo, JRC's Earth Observation Data Processing Platform, and CESNET's DataCare. But these services are in most cases isolated islands and do not necessarily fit with the collaborative nature of many research activities. It is very often difficult to give access to users located in other institutions, share data across platforms and run applications on top of this fragmented landscape.

These concerns have led the institutions mentioned above to come together and, along with four other partners (Ailleron, Cubbit, ESADE and Trust-IT), conceive a plan to interlink these services; and they have received the support of the European Commission to do this a Project: CS3MESH4EOSC (cs3mesh4eosc.eu). The idea is to boost open science by presenting a joint, coordinated service — Science Mesh (sciencemesh.io) — to the users in the research and education space, at a pan-European level; and to answer a major question: can we manage and store research data in Europe?

In this presentation, we will introduce the CS3MESH4EOSC project and summarize the main achievements of this first year of work, with a focus on the conception and design of the Science Mesh platform. We will also explain how the CS3 ecosystem can participate in what we hope will be a community-wide effort towards a true federation of CS3 services for Science.

On-premise, hybrid or cloud? / **175**

Campfire discussion

Tech Short Talks / **152**

CodiMD in CERNBox: leveraging the WOPI protocol to provide collaborative markdown editing

Authors: Giuseppe Lo Presti[1]; Michal Kolodziejski[1]

[1] *CERN*

Corresponding Author: giuseppe.lopresti@cern.ch

This contribution describes the integration of CodiMD, a popular markdown nodejs-based editor, in CERNBox.

CodiMD is the open-sourced version of a cloud service run by HackMD.io. Designed to store users' content in a relational database and blobs in a filesystem-based or cloud (e.g. S3) storage, there is no provision for interfacing external storages. But thanks to its open REST API, it is straightforward to programmatically pull and push content to it: add a few tweaks and you get a fully fledged collaborative editor integrated in CERNBox.

The integration work has been prototyped in the context of the ScienceMesh project, and it is now proposed to CERNBox users. A key aspect lies in the use of the WOPI protocol, which enables storing some arbitrary state within a WOPI lock. Despite locking can only be considered advisory in a sync-and-share context, in this case WOPI locks help designing a stateless integration, where no state is added to the CodiMD application, and a *WOPI Bridge* microservice fills the gap between CodiMD and WOPI.

Collaborative applications / **132**

Collabora Online: Secure, on-premise collaborative editing

Author: Michael Meeks[1]

[1] *Collabora*

Corresponding Author: michael.meeks@collabora.com

Come and hear how Collabora Online delivers scalable, secure, on-premise control of your data with a simple, easy to deploy and manage approach.

Hear about our significant improvements in functionality in the last year everywhere from a host of interoperability improvements, to a polished mobile and tablet interface, with native mobile and ChromeOS versions. See our re-worked user-experience, and the options that gives as well as the ability to easily theme & style to your taste.

Hear how the one-click install app-image approach for easy setup (that was prototyped at the last CS3) works, and why you don't want that in a large deployment.

Hear some thoughts on how our simple architecture allows easy deployment, simple scaling, high availability and more for your EFSS.

Science Mesh workshop / **184**

Conclusion

Corresponding Author: jakub.moscicki@cern.ch

On-premise, hybrid or cloud? / 150

Cubbit Hive: the private distributed cloud

Authors: Gianluca Granero[1]; Lorenzo Posani[1]

[1] *Cubbit*

Corresponding Authors: gianluca.granero@cubbit.io, lorenzo.posani@cubbit.io

In the scientific community, we have - at the same time - a strong need to **seamlessly store and share files** with colleagues all around the world, and a major constraint of **data sovereignty**.

As many publicly-funded institutions are not allowed to use commercial cloud storage products - as they are run and hosted by foreign companies obeying their own local laws - a common solution is to host a **private cloud infrastructure** within their own premises.

However, not every institution can afford the heavy hardware and connectivity infrastructure, along with the required IT workforce for maintenance, that is needed to run efficiently a private cloud solution.

In this talk, we present Cubbit Hive: an innovative approach that **decouples the storage service from the need for dedicated infrastructure**, virtualizing a private cloud on a pre-existing network of connected devices.

Cubbit Hive intelligently **collects spare storage and computing resources** within the premises of the institution (workstations, small servers, etc.) to enable a distributed storage service that is **encrypted**, **fast**, and **compliant-by-design** with the strongest needs of security and sovereignty.

Novel Data Science Environments / 139

Data access and integration challenges in a distributed computational environment for medical research

Authors: Piotr Nowakowski[1]; Maciej Malawski[2]; Marian Bubak[3]; Tadeusz Satława[4]; Jan Meizner[5]; Tomasz Gubała[5]; Marek Kasztelnik[1]

[1] *ACC Cyfronet AGH*

[2] *AGH University of Science and Technology*

[3] *AGH Krakow*

[4] *u*

[5] *Sano Centre for Computational Medicine*

Corresponding Author: ymnowako@cyf-kr.edu.pl

Carrying out research at the forefront of medical science with the use of distributed computing resources poses a number of unique challenges related to securely managing, integrating, storing and accessing data. The authors present their experience in developing a distributed computational environment for a range of studies in the area of computational medicine, with particular focus on how the environment - and its users - treat the underlying data, both with respect to inputs and outputs of medical simulations. These experiences are presented against the backdrop of work performed in a number of collaborative research projects, and currently ongoing at the Sano Centre for Computational Medicine.

This project has received funding from the EU H2020 research and innovation programme under grant agreement No 857533 and from the International Research Agendas Programme of the Foundation for Polish Science No. MAB PLUS/2019/13.

Collaborative applications / 135

Data encryption and crypto-flexibility in collaboration

Author: Mikhail Korotaev[1]

[1] *ONLYOFFICE*

Corresponding Author: mikhail.korotaev@onlyoffice.com

As electronic collaboration patterns evolve, we must think of adequate new calculus in document protection that neither draws new security gaps nor limits the casual editing and collaboration tools in the name of data protection.

Most of the existing document encryption schemes do not provide all deliverables important in keeping the process agile yet backed up on the security side. Electronic documents can be protected by the standard secure AES encryption, but sharing of the data necessary for decryption via insecure channels or preliminary decryption in plain sight in a public cloud seems to make the efforts futile.

With a brand new Private Rooms technology, ONLYOFFICE attempts at creating a safe space for documents where every bit of information is encrypted end-to-end, securely shared at any scale with no complication in user experience, and is ultimately collaboration-ready.

The talk will cover the following topics:

· Evolution of the document encryption approach in ONLYOFFICE,
· Chosen principles and methods of document encryption in the present scheme,
· Mechanics of file sharing and collaboration,
· Allocation of Private Rooms within the user environment,
· Application of the technology in existing integration scenarios,
· Future steps.

FAIR and Open Research Data Services / 154

Describo and RO-Crate - the FAIR data research helpers

Authors: Marco La Rosa[1]; Peter Sefton[2]

[1] *The University of Melbourne*

[2] *UTS*

Corresponding Authors: m@lr.id.au, pt@ptsefton.com

Too often, research projects don't involve work-practices that describe the data generated over the lifetime of the project so that it meets the FAIR principles. Making data Findable so that it is Accessible, Reusable by others and Interoperable with archives, preservation systems and analytical systems. To address this we will introduce Describo; a desktop and online tool. Describo enables researchers to turn their folders of content (on their Desktop or in cloud-based share/sync services; initially implemented with Microsoft OneDrive but with an architecture that allows it to be used with many cloud/file services) into Research Object Crates (RO-Crate) suitable for sharing, reuse and long term preservation in archival systems.

Describo allows research teams from ANY discipline to describe their data using Linked Data methods; associating it with its creators using ORCID; unambiguously identifying funding sources, equipment and other provenance to maximise the Reuse and Interoperability potential of the data.

Collaborative applications / **141**

Digital Sovereignty with email & calendars - Is the client the bottleneck?

Author: Andreas Rösler[1]

[1] *Kopano BV*

Corresponding Author: a.roesler@kopano.com

Even super-humans have every hour of the day only once. To collaborate with teams a requirement is to have one calendar per person and to share its free/busy-information with team-members. (How) Is it possible to collaborate beyond the borders of the own organizations in a digital sovereign world with a colourful bouquet of different collaboration solutions? Hint: I do not trust in the 'Ring that rules them all". And the standard must not be a protocol.

My talk is about the thing we call ' groupware' and we do mean e-mail, calendar, contacts and tasks by this. I am going to focus on calendars and will reflect on a history starting at MS Exchange 2000, following a path via MAPI, IMAP, CalDAV, ActiveSync-over-the-air, z-push, and completely bypassing MS Outlook. The view I am going to present is from a 3rd party integrations perspective as well as from an end-users perspective. Hereby I will touch the role of APIs to connect to standards to realize interoperability with other tools, even the ones outside your own organization. An example I will touch is the exchange of calendar availability.

Tech Short Talks / **153**

Distribution of container images: From tiny deployments to massive analysis on the grid

Author: Enrico Bocchi[1]

[1] *CERN*

Corresponding Author: enrico.bocchi@cern.ch

- Distribution of container images: From tiny deployments to massive analysis on the grid.

In recent years the use of containers proliferated to the point that they are now the de-facto standard to package and run modern applications. A crucial role in the successful distribution of container images is played by container registries, specialized repositories meant to store container images. Due to the popularity of containers, public registries had to constantly increase their storage and network capacity to withstand the huge demand from users. In August 2020, the Docker Registry has announced changes to the retention policy such that images hosted in free accounts would be deleted after 6 months of inactivity. While the enforcement of the new retention policy was postponed to mid-2021, many users of containers at CERN started to investigate alternative solutions to store their container images.

CERN IT offers a centralized GitLab Container Registry based on S3 storage. This registry is tightly integrated with code repositories hosted on CERN GitLab and allows for building and publishing images via CI/CD pipelines. Plans are to complement the GitLab Registry with Harbor (https://goharbor.io/),

the Open Cloud Initiative container registry, which provides advanced capabilities including security scans of uploaded images, authentication and authorization (via LDAP, AD, OIDC, RBAC), non-blocking garbage collection of unreferenced blobs, and proxying/replication across other Harbor instances or Docker Hub.

Containers are also becoming more and more popular in the High Energy Physics community, where scientists encapsulate their analysis code and workflow inside a container image. The analysis is firstly validated on a small dataset to then run on the massive computing capacity provided by the Worldwide LHC Computing Grid. In this context, the typical approach of pulling a container image from a registry and extract it on the worker node show its limitations and results very inefficient. For this specific use-case, the CERN VM FileSystem (https://cernvm.cern.ch/fs/), a well-established service for the distribution of software at a global scale, comes to help. It features a dedicated ingestion engine for container images (based on per-file deduplication instead of per-layer) and an optimized distribution and caching mechanism that allows to greatly save on network bandwidth and local storage. The integration between CVMFS and the GitLab Registry (and Harbor) is being investigated to provide the end-users with a unified management portal for their container images (GitLab or Harbor) while supporting the large-scale analysis scenarios typical of the HEP world.

Scalable Storage / **160**

EOS-wnc: EOS Client for Windows

Author: Gregor Molan[1]

[1] *Comtade*

Corresponding Author: gregor.molan@cern.ch

Most of CERN experiments are using Linux systems for data collection so EOS was designed to work primarily with Linux. But there are some high demanding users that are using Windows systems for data collection. Possible solution for them is to use Samba to mount Linux file system as Windows drive. Most of Windows users are used to get instant applications and instant file systems, and consequently, they try to avoid the use of a combination of different programs for wanted functionality. And that it is not just a users' caprice – using a mixture of different programs is not a Windows way to work. Such a compilation also degrades the performance of EOS file system.

The solution for Windows high-performance users is *EOS-wnc* – an EOS Windows Native Client and a direct interface between Windows and EOS cluster. EOS Windows Native Client (*EOS-wnc*) provides EOS administration and file-related operations between local Windows file system and remote EOS file system on Linux. It is designed and implemented to directly access EOS cluster from Windows platform. Development of such a client is based on CERN-Comtrade collaboration on EOS Productisation project that has been living and growing since 2015.

The *EOS-wnc* project started with investigation of Windows file systems as an upgrade of Comtrade's extensive knowledge about EOS. Studied were many possible implementations for Windows file system:
* XtreemFS, github.com/xtreemfs
* Chef infra, chef.io/products/chef-infra
* UrBackup, urbackup.org
* Syncthing, syncthing.net
* Duplicati, duplicati.readthedocs.io
* Bacula, bacula.org

The natural decision was to provide EOS client on Windows environment. There were two options to do it:
* Use Windows Subsystem for Linux (Microsoft WSL)
* Write a new Windows EOS client

As the first option did not solve high-performance demands, there following options to write a new EOS client for Windows were studied:

* Win32 API (Windows API)
* Windows Presentation Foundation (WPF)
* Windows Forms (.NET)
* Universal Windows Platform (UWP)
* Extension to the Windows Runtime (WinRT)

Furthermore, there were two possible directions of development:
* Port the existing XrootD and existing EOS Linux client to Window
* Write a completely new EOS client for Windows with a new architecture

Naturally, we decided to develop a completely new EOS Windows native client. The architecture of this new EOS client is different that the architecture of EOS client on Linux. The result is an implementation that leverages the advantages of Windows system and implements a completely new EOS client: *EOS-wnc*. This will be the first public live presentation of *EOS-wnc* after some CERN internal live demo presentations.

EOSC & Federations: Future research infrastructures / 166

European Open Science Cloud: A single European market for research data and services

Author: Bob Jones[1]

[1] *CERN*

Corresponding Author: robert.jones@cern.ch

Building an open and trusted environment for accessing and managing research data and related services, the European Open Science Cloud (EOSC) will transform how researchers access and share data and give Europe a global lead in research data management.

The EOSC will federate existing scientific data infrastructures across disciplines and the EU Member States.
Supporting the EU's Open Science policy and European Data Strategy, EOSC will help European scientists reap the full benefits of data-driven science and give Europe a global lead in research data management.

Key features of EOSC will be its smooth access to data and interoperable services and its trusted digital platform that addresses the whole research data cycle, from discovery, mining, storage, management, analysis and re-use.
With open and seamless services for storage, management, analysis and re-use of research data across borders and scientific disciplines, EOSC aims to create a virtual environment for 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences.

Seeking to make digital assets findable, accessible, interoperable and reusable (FAIR), EOSC will tackle societal challenges including early diagnosis of major diseases and climate change.

Speaker's Bio

Bob Jones is a director of the recently formed EOSC Association: https://www.eosc.eu/
He is a senior member of the scientific staff at CERN (http://www.cern.ch) and was the coordinator for the award winning Helix Nebula Science Cloud Pre-Commercial Procurement project (http://www.hnscicloud.eu/) procuring innovative cloud services for the European research community and contributing to the EOSC. His experience in the distributed computing arena includes mandates as the technical director and then project director of the EGEE projects (2004-2010) which led to the creation of EGI (http://www.egi.eu/).

Publications: http://orcid.org/0000-0001-9092-4589
Twitter: @BobJonesAtCERN

Tech Short Talks / **165**

Evolution of the CS3APIS and the Interoperability Platform

Author: Ishank Arora[1]

Co-author: Hugo Gonzalez Labrador [1]

[1] *CERN*

Corresponding Author: ishank.arora@cern.ch

CS3APIs started out as a means to offer seamless integration of applications and storage providers, to solve the issue of fragmentation of services. Its reference implementation - Reva, has evolved over time to effortlessly allow plugging in numerous authentication mechanisms, storage mounts and application handlers through dynamic rule-based registries. As an Interoperability Platform REVA offers opportunities for application discovery and scalability to meet the growing storage needs of the next decade.
These functionalities allow the concept of 'Bring Your Own Application' to be introduced, allowing admins to offer a catalogue of applications and protocols and affording the end-users the freedom to choose among them. The inter-operability has also allowed Reva to serve as the benchmark for the EU project, CS3MESH4EOSC, which aims to tackle the challenges of fragmentation across CS3 platforms and promote the application of FAIR principles. Efforts into developing a federated service mesh providing frictionless collaboration for hundreds of thousands of users are already underway. These developments would pave the way for a data-driven application ecosystem in the European Open Science Cloud.

FAIR and Open Research Data Services / **167**

FAIR and Open Data: challenges and opportunities

Author: TBD TBD[1]

[1] *TBD*

This talk will summarise the requirements on the FAIR and Open Data in the future EOSC and the practical impact on service providers and implementors.

Novel Data Science Environments / **148**

Forget 2020: Building an integrated AARNet Cloud Ecosystem in 2021

Authors: Gavin Charles Kennedy[None]; Carina Kemp[1]; Brad Marshall[2]

[1] *AARNet*

[2] *AARNet*

Corresponding Author: gavin.kennedy@aarnet.edu.au

2020 has been a terrible year for everyone and has delivered significant challenges for AARNet (Australia's Academic Research Network provider) in the delivery of online researcher facing services. A surge in demand across multiple services from March onwards has kept the operational teams busy, while online-only work practices initially slowed our development activities. However AARNet have embraced these challenges by starting the build on a more comprehensive cloud ecosystem

that extends the research value of AARNet's CloudStor, and has laid the foundations for a productive 2021.

The AARNet Cloud Ecosystem (working title) brings together compute, storage and data movement across a range of services. At the foundation layer are three separate file storage platforms: CloudStor which provides EFSS services, as well as SWAN Jupyter Notebooks compute capability; S3 Storage providing direct high performance disk storage; and Sensitive Data providing isolated high security sensitive data storage and management services in an ownCloud/EOS stack. Integrated with these storage platforms are multiple OpenStack based compute platforms providing PAAS+ hosting services for high value research community applications such as the Galaxy bioinformatics suite. Then, to provide long term low cost storage services we are implementing a tape based Cold Storage service, based on the CERN Tape Archive (CTA) platform. And finally, to enhance our data movement capabilities and make the most out of our network, we are licensing Globus Data Transfer for our research customers and plan to integrate it into our Cloud Ecosystem.

This talk will discuss the business drivers, the high level architecture and the logistical challenges of delivering this scaled up service ecosystem. It will also discuss the value of the multiple technologies used to build and deploy the services (OpenStack, Terraform, Kubernetes, etc). It will finish with highlighting the importance of the REVA storage driver and the CS3APIs as they are being implemented by our software partners, and the opportunity this program of work represents for AARNet's integration with the EU Science Mesh.

Scalable Storage / **169**

GPFS – protocols and watchfolder function

Author: Weiser Olaf [1]

[1] *IBM*

IBM's high performance file system solution scales out over multiple nodes and storage technologies. Common data access over multiple paths and protocols like direct-POSIX, NFS, SMB, HDFS, OBJ and a full integration into Kubernetes container can help to manage your data more efficiently. SpectrumScale's so called watchfolder function can be used to take advantage from central file system notifications of a distributed environment to post process data automatically.

EOSC & Federations: Future research infrastructures / **134**

HIFIS: Sync&Share Federation for Helmholtz

Authors: Sander Apweiler[None]; Andreas Klotz[1]; Matthias Leander-Knoll[2]

[1] *Helmholtz Berlin*

[2] *KIT*

Corresponding Authors: sa.apweiler@fz-juelich.de, andreas.klotz@helmholtz-berlin.de

While HIFIS1 has released its initial Helmholtz Cloud Service Portfolio2 in October 2020, there has also been specific development in terms of Sync&Share within the Helmholtz Association. The Helmholtz center's Sync&Share administrators are meeting under the umbrella of HIFIS and are planning possible implementations of a Helmholtz-wide Sync&Share federation, which will be provided as a service within the Helmholtz Cloud.

Out of 19 Helmholtz centers with over 42.000 employees, there are currently nine centers operating their own Sync&Share instance. Four of those are currently planning to open their instances for other Helmholtz users via the Helmholtz Cloud. Our goal will be to equally distribute the majority

of the remaining center's users on the providing instances. A federation of all Helmholtz based Sync&Share instances is also planned. Both of these goals bring multiple challenges, which lie ahead of us. Each has its own set of possible solutions.

We already observe a big trend to groups of members from multiple centers working in Sync&Share group folders and calendars. This brings up the question: How can we federate in such a way that users reside on only one instance, while collaborating with their group members on other instances, using group folders, calendars and additional extensions in the future?

Another big issue is GDPR compliance, especially since all 19 Helmholtz centers are independent legal entities. Each having their own Data Protection Officer as well as employee representatives. Our approach is based on a Helmholtz-wide cooperation contract, including a specific set of Helmholtz AAI Policies[3], based on international guidelines and templates (e.g., WISE, AARC).

1 HIFIS Website: https://hifis.net
2 Helmholtz Cloud - Initial Service Portfolio: https://hifis.net/news/2020/10/13/initial-service-portfolio
[3] Helmholtz AAI Policies: https://hifis.net/policies

Science Mesh workshop / **180**

How to join the Science Mesh?

Corresponding Author: daniel.mueller@uni-muenster.de

Context, limitations and practical steps to be part of the Science Mesh

Site reports / **158**

Introducing an open source Sync and Share solution in Sweden

Authors: Anders Bruvik[1]; Gabriel Paues[None]

[1] *Safespring*

Corresponding Authors: anders.bruvik@safespring.com, gabriel.paues@safespring.com

Time to introduce a new service!
Safespring is together with Sunet building a sync and share solution based on Nextcloud for Sweden. Storage is hosted on S3/Ceph based backends based on Safesprings Private Cloud offering. We started working on the implementation in 2020, and the service will be released to the public in the early 2021. The goal is to offer this as a service to all Higher educational institutions in Sweden.

In this talk, we will talk about the the current status of the project, arcitechtural choices - which is all based on open source software - and we will talk about the timeline going forward.

Novel Data Science Environments / **156**

JupyterLab for Earth Observation applications with HTCondor scaling and Voilà dashboarding

Authors: Davide De Marchi[1]; Armin Burger[1]; Pierre Soille[1]

[1] *European Commission*

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

The Joint Research Centre (JRC) of the European Commission has set up the JRC Big Data Analytics Platform (BDAP) as a multi-petabyte scale infrastructure to enable EC researchers to process and analyse big geospatial data in support to EU policy needs 1. One of the service layer of the platform is the JEO-lab environment 2 that is based on Jupyter notebooks and the Python programming language to enable exploratory visualization and interactive analysis of big geospatial datasets. JEO-lab is set-up with deferred processing, using multiple service nodes to execute the Jupyter client processing workflow starting from data stored in the CERN EOS distributed file system deployed on the BDAP. In this context, recent developments were done in these areas:

• Scaling to HTCondor: batch processing submission from the notebook
BDAP uses HTCondor [3] as the scheduler for the batch processing activities. Users can submit complex tasks to the HTCondor master that will then allocate jobs to the HPC nodes and control their progressing. The submit activity is generally done from a remote desktop environment which consists of a linux instance accessible from the browser. For some specific tasks, in particular those involving geospatial datasets, the direct submit of simple processing tasks from the interactive notebook environment has been developed, calling the python bindings of HTCondor. This allows for easy check of the results, thanks to a new GUI created in Jupyter to control the status of the processing jobs and to instantly visualize the produced datasets on an interactive map. This new development complements the usual deferred mode of the JEO-lab environment, which manages big geospatial datasets by processing them at the screen zoom level and at the visible viewport, with a tool that can scale the prototyped processing chains to operate at the full resolution of the input datasets, in order to create and permanently save any type of derived product. The processing is based on python: the user can create a custom python function that is applied to all the input image files to generate the output product by using any Numpy, Scipy, gdal, pyjeo [4] function. Examples will be demonstrated for the numerical evaluation of the effect of the lockdown measures on air quality at European level using the data coming from the Sentinel-5P satellite of the EU Copernicus programme.

• Web dashboards derived from Jupyter notebooks using Voilà
Voilà [5] turns Jupyter notebooks into standalone web-dashboard applications; it supports Jupyter interactive widgets like ipywidgets [6], charting libraries like plotly [7], etc., while not permitting arbitrary code execution, thus posing less security threats. Many applications developed inside the JEO-lab environment have been brought into the Voilà world, where they are accessible without the need for user authentication, and thus greatly expanding the impact of the BDAP platform and providing an easy way to publish complex interactive visualization environments. Among the most important, the CollectionExplorer dashboard that enables users to browse all the geospatial datasets available in the platform with an easy-to-use interface, dedicated dashboards like S2explorer and DEMexplorer, to perform typical GIS operations on Sentinel2 products and Digital Elevation Models, and many dashboards for the monitoring of the Covid-19 spread in various regions and continents.

Both developments were partially financed by the H2020 project CS3MESH4EOSC, led by CERN and to which JRC participates providing support in the Earth Observation use case. In this context, the CERN SWAN Jupyter environment was also deployed in view of future full adoption of the IOP and the connection to the ScienceMesh federated infrastructure.

The JRC Big Data Analytics Platform is a living demonstration of a complex ecosystem of cloud applications and services that allows data scientists' navigation inside a multi-petabyte scale world. In particular, the exploratory visualization and interactive analysis tools in the JEO-lab component can run custom code to prototype the generation of scientific evidence as well as create GUI applications that can be used by end-users ranging from policy makers to citizens.

1 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.

2 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.

[3] https://research.cs.wisc.edu/htcondor/

[4] P. Kempeneers , O. Pesek, D. De Marchi, P. Soille. “pyjeo: A Python Package for the Analysis of Geospatial Data”, ISPRS International Journal of Geo-Information, Volume 8, Issue 10, October 2019. Special Issue “Open Science in the Geospatial Domain”.
https://doi.org/10.3390/ijgi8100461

[5] https://blog.jupyter.org/and-voil%C3%A0-f6a2c08a4a93
https://github.com/voila-dashboards/voila

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

[7] https://plotly.com/

Scalable Storage / **157**

Making Reva talk to EOS: ultimate scalability and performance for CERNBox

Author: Fabrizio Furano[1]

Co-authors: Hugo Gonzalez Labrador [1]; Ishank Arora [1]; Samuel Alfageme Sainz [1]

[1] *CERN*

Corresponding Author: fabrizio.furano@cern.ch

The Reva component, at the heart of the CERNBox project at CERN will soon get new plugins that build on the experience accumulated with the current production deployment, where its data is stored centrally in a system called EOS. EOS represents since 10 years the ultimate development effort into providing an extremely scalable data storage system that supports the demanding requirements of the massive physics analysis, together with the more regular requirements of a wider community (scientists, engineers, administration): synchronisation and sharing, online and universal access and real time collaborative workflows.

Making Reva natively interfaced to EOS through high performance gRPC and standard HTTPS interfaces will open a new scenario in terms of scalability and manageability of the CERNBox service, whose requirements in terms of data will continue to grow in the next decade. In this contribution we will technically introduce this near-future scenario.

EOSC & Federations: Future research infrastructures / **163**

National Data Storage’s synergies with sync & share systems

Authors: Maciej Brzezniak[1]; Norbert Meyer[2]; Pawel Wolniewicz[2]; Radosław Januszewski[3]; Mirosław Kupczyk[3]; Rafał Mikołajczak[3]; Krzysztof Wadówka[None]; Pokora Eugeniusz[3]

[1] *PSNC Poznan Poland*

[2] *Unknown*

[3] *PSNC*

Corresponding Author: maciekb@man.poznan.pl

Sync & share services are nowadays a natural component of the cloud computing and storage services. This is also the case of the services and applications provided to academic and research environment in Poland by Poznan Supercomputing and Networking Centre (PSNC), a Polish NREN and HPC center. Since 2015 PSNC provides a country-wide sync & share service based on Seafile software and GPFS storage back-end. In The solution serves long tail users and scientists addressing typical sync & share scenarios and use-cases as well as a provides high-performance, scalable data storage platform for PSNC cloud computing and HPC systems.

In 2021 PSNC will start implementing the Polish government funded R&D National Data Storage (NDS) project that focuses on designing, building and deploying the large scale data infrastructure (200+PBs of tapes, 200+PB of disk storage and 10+PB of SSD/NVMe storage) across 4 HPC centres in Poznan, Kraków, Gdańsk and Wrocław and 4 MAN centers in Łódź, Białystok, Częstochowa and Kielce. Based on this infrastructure and the PIONIER broadband network PSNC and partners will develop and integrate set of data management services for data storage and access, long-term preservation, data exploration, analysis, and processing. Natural part of the project is development and integration of scalable and feature-full sync & share services.

In our presentation we will discuss the next stage of the sync & share development at PSNC that will be achieved in synergy with the National Data Storage project. First of all, NDS project will both provide infrastructure and services that will be used as the back-end of the sync & share service, including large-scale Ceph-based disk servers clusters along with the SSD/NVMe pools for data storage and access acceleration. In additional NDS project has dedicated tasks on integrating the sync & share services into the data management hardware and software stack of the national cloud computing, HPC and storage infrastructure. NDS project includes also activities related to implementing the extended scenarios for sync & share such as using sync & share systems in the data collection and processing workflows, agile deployment of dedicated and secure sync & share environments - per user group or project as well as combining sync & share systems with long-term storage systems and data repositories.

File Sync&Share Products for Home, Lab and Enterprise / **137**

Nextcloud - State of the nation

Author: Frank Karlitschek[1]

[1] *Nextcloud*

Corresponding Author: frank@nextcloud.com

This talk will give an overview of the big improvements that happened in Nextcloud in the last year. In the last 12 month Nextcloud Hub 18, 19 and 20 were made available. Nextcloud Hub 21 will be released in a few days. During this time a lot of significant improvements in functionality, performance, scalability and security were released. This talk will give an overview together with some real world example how the new capabilities can be used.

OCM Workshop / **192**

OCM - from the initial design to a core component of Nextcloud Hub

Corresponding Author: bjoern@nextcloud.com

On-premise, hybrid or cloud? / **168**

On-premise or in the cloud?

Author: Alberto Pace[1]

[1] *CERN*

Corresponding Author: alberto.pace@cern.ch

Discussion on different service provisioning models.

Collaborative applications / **144**

OnlyOffice and Collabora Online Experience at CERN

Authors: Maria Alandes Pradillo[1]; Piotr Jan Seweryn[1]; Mario Rey Regulez[1]; Giuseppe Lo Presti[1]

[1] *CERN*

Corresponding Author: maria.alandes.pradillo@cern.ch

Collaboration features are nowadays a key aspect for efficient team work with productivity tools. During 2020, CERN has deployed and monitored OnlyOffice and Collabora Online solutions in CERNBox. CERN users working online can now choose OnlyOffice, which is the default application to work with OOXML files in CERNBox; and Collabora Online, which is the default application for ODF files in CERNBox.

This presentation will focus on the technical aspects of deploying and maintaining OnlyOffice and Collabora Online within CERN, and the integration with our CERNBox infrastructure. It will also give an overview of the main advantages and disadvantages our user community has faced when interacting with these applications. The presentation will also describe how these new applications have been rolled out to the user community in terms of communication, training and user support.

Site reports / **162**

Open Heart Surgery with 60 Terabytes of Data: Migrating 40.000 users from PowerFolder to Nextcloud

Author: Sascha Wiswedel[1]

[1] *Nextcloud GmbH*

Corresponding Author: sascha.wiswedel@nextcloud.com

tl;dr: It works!

System migrations are rarely easy. Neither do you want to lose any data (worst case) or metadata (close to the above) along the way, nor - and that might be the worst case in the long run - do you want to lose the trust of the end users after having decided to go for a new solution.

When migrating 40.000 users from 55 universities and higher education organizations with 60 TB of data, from PowerFolder to Nextcloud, you want to make perfectly sure that on Monday morning everybody finds everything to be as complete as it was on Friday afternoon.

By the following Thursday you also hope to be able to stop thinking about the word “rollback” coming from the customer to your private Telegram in the middle of the night.

Let’s look at some learnings from this major migration project of 2020 and what had to be done in terms of logic analysis and custom programming (which ended up in Nextcloud’s core, open for everyone) so students, teachers, staff and guests can work effectively and efficiently with a top of class content collaboration platform.

On-premise, hybrid or cloud? / 149

Oracle Data Science Cloud

Author: Peter Szegedi[1]

[1] *Oracle*

Corresponding Author: peter.szegedi@oracle.com

Oracle Data Science Cloud is a collaborative platform for data scientists to build and manage ML models. Leveraging open source technology, it provides a scalable cloud-based platform for data scientists to explore data and train, save, and deploy models, while utilizing the rich Python ecosystem as well as Oracle’s proprietary Python libraries.
• Data Analysts - Oracle Analytics Cloud delivers cutting-edge visualization, augmented analysis, and natural language processing through an easy-to-use interface. Powered by AI and machine learning, Oracle Analytics Cloud makes it possible for any level of user to generate deep insights and create forward-thinking reports.
• Data Engineers - Oracle Database’s Machine Learning capabilities bring the latest automation and self-learning tools to the database space. The result is an experience that’s both powerful and user-friendly. Using Oracle’s tools, it’s easier than ever to manage data and support application development on a secure and scalable infrastructure.

The platform combines three key components:
• Infrastructure - systems tasks (like spawning servers) are abstracted and handled automatically so data scientists can focus on the substance of their work.
• Tools - open source tools (like Jupyter, R Shiny, or modeling libraries) that data scientists need are integrated into a centralized place.
• Workflow - automation for tasks, collaboration, and communication that let data science teams effectively deliver on their mission.

The presentation is going to include a live showcase of the Oracle Data Science Cloud platform using a simple user scenario!

Oracle Cloud Infrastructure offers exceptional performance, security, and control for today’s most demanding high-performance computing (HPC) research workloads. The Oracle HPC cloud consists of:
• Bare metal instances support applications requiring high core counts, large amounts of memory, and high memory bandwidth. Users can build cloud environments with significant performance improvements over other public clouds and onsite data centers. Bare metal compute instances provide researchers with exceptional isolation, visibility, and control.
• NVIDIA-based offerings provide a range of options for graphics-intensive workloads, along with the high performance demanded by AI and machine learning algorithms.
• Virtual cloud networks enable researchers to easily move their existing network topology to the cloud. Standard bare metal servers support dual 25 Gbps Ethernet for fast front-end access to your compute clusters. Oracle’s groundbreaking, back-end network fabric lets them use Mellanox’s ConnectX-5, 100 Gbps network interface cards with RDMA over converged Ethernet (RoCE) v2 to create clusters with the same low-latency networking and application scalability that one would expect on premise.

Oracle provides several high-performance storage options suitable for HPC workloads.
• Local NVMe SSD: High-speed local flash storage ideal for large databases, high performance com-

puting (HPC), and big data workloads such as Apache Spark and Hadoop.
• Block volumes: Standard block storage services offering 60 IOPS per Gb, up to a maximum of 32,000 IOPS per volume, backed by Oracle's highest performance SLA.
• Parallel file systems: HPC requires larger data sets and higher performance than standard enterprise file servers can provide. Customers often build their own parallel file systems either on premise or in the cloud using open source software such as Lustre.

FAIR and Open Research Data Services / 131

Progress of Sciebo Research Data Services

Authors: Peter Heiss[1]; Holger Angenent[2]; Lennart Hofeditz[3]

[1] *University of Muenster*

[2] *University of Münster*

[3] *University of Duisburg-Essen*

Corresponding Authors: peter.heiss@uni-muenster.de, lennart.hofeditz@uni-due.de

In order to follow the Open Science idea, accurate research data management (RDM) becomes increasingly important. As one consequence, research institutions and third-party organizations began to develop e-science technologies such as data storages and digital research environments 1–[3]. However, on operational level, there hardly is an appropriate infrastructure. Existing services are poorly linked to the RDM steps which are demanded by public funders and institutions [4].
Last year's CS3, we provided the first results of the project sciebo RDS (research data services) which is a highly modular RDM infrastructure in order to support open science aspirations and connect already existing services. A key aspect of the project is to develop and improve low-threshold services that will result in an increasing use of RDM guidelines among potential users. This year, we want to present our progress and the next steps.
So far, we implemented various connectors (e.g. to Zenodo and Open Science Framework) and a functional user interface within ownCloud. In a cooperation with the Science Mesh project, we further plan to integrate the tool "Describo" to enable the collection of metadata and various metadata schemes to a research project without leaving ownCloud as a RDM platform. In the future, it should be possible to use the RDS interface in other cloud storage implementations such as Nextcloud or Seafile without major code changes. Usability optimizations and improvements towards a user-centered GUI will be developed in a dedicated research project. For this, we will use various principles of digital nudging to increase the awareness for the implemented workflows, based on the DINU-model by Meske & Potthof [5].

1 S. Stieglitz et al., "When are researchers willing to share their data? – Impacts of values and uncertainty on open data in academia," PLoS One, vol. 15, no. 7 July, 2020.
2 K. Wilms et al., "Digital Transformation in Higher Education – New Cohorts, New Requirements?," in Proceedings of the 23rd Americas Conference on Information Systems (AMCIS), 2017, pp. 1–10.
[3] R. Vogl, D. Rudolph, and A. Thoring, "Bringing Structure to Research Data Management Through a Pervasive, Scalable and Sustainable Research Data Infrastructure," in The Art of Structuring, K. Bergener, M. Räckers, and A. Stein, Eds. SpringerLink, 2019, pp. 501–512.
[4] L. Hofeditz et al., How to design a research data management platform? technical, organizational and individual perspectives and their relations, vol. 12185 LNCS, no. July. Springer International Publishing, 2020.
[5] C. Meske and T. Potthoff, "The DINU-Model - A Process Model For The Design Of Nudges," in Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal, 2017, vol. 2017, pp. 2587–2597.

Welcome and Keynote / 173

Q/A

Corresponding Author: steiger@id.ethz.ch

Science Mesh workshop / **183**

Q/A and Feedback

Follow up with our attendees and early adopters to discuss their expectations and requirements in terms of services to be deployed by Science Mesh.

Site reports / **186**

Q/A discussion

Scalable Storage / **170**

REVA on CephFS

Author: Theofilos Mouratidis[1]

[1] *CERN*

Corresponding Author: theofilos.mouratidis@cern.ch

CephFS is a distributed file system based on the popular storage system Ceph. This filesystem is a scalable system with POSIX features that makes it a compelling candidate for a Sync&Share backend. Sync&Share applications have a lot of users interacting with them, doing constant I/O tasks. Therefore, the need for a filesystem that can handle this I/O load is necessary. Even if the performance part is important, Sync & Share applications have a lot of features that improve the QoL of the system, for example, ACLs, file versions, fast file discovery for synchronisation and so on. We present a proof of concept module of CephFS on REVA, to evaluate whether this filesystem can fully support a Sync & Share application or not. This module is currently implemented using a local filesystem module, where the operations are performed on a file path using the tools that the Linux OSes provide. The CephFS module of REVA currently supports recursive mtime propagation for the fast detection of the files that changed. It also provides file versions based on snapshots, with the current limitation that the file versions are fixed based on a time difference rather than I/O based. Other features such as ACLs, or using a CephFS client instead of an OS interface are not present, but can easily be implemented. This proof of concept shows promising results and we firmly believe that with some minor changes it can support a Sync & Share application.

Novel Data Science Environments / **138**

Running Parallel, Distributed ROOT Analysis with PyRDF on Public Cloud - AWS Lambda Case Study

Authors: Jacek Kusnierz[1]; Piotr Pasternak[1]; Maciej Malawski[1]

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

Corresponding Author: jacek.andrzej.kusnierz@cern.ch

ROOT framework is a standard tool for HEP data storage and analysis. Current approaches for parallel and distributed processing in ROOT require either batch systems such as HTCondor, or big data frameworks like Apache Spark supported by the recently added PyRDF interface. However, these approaches are heavily reliant on existing scientific infrastructure. On the other hand, cloud computing allows provisioning resources on demand in increasingly elastic and fine-grained ways, as in example of serverless computing and function-as-a-service model.

To make it possible to run parallelized ROOT without any dependency on existing resources in research computing centers, we explore the possibility of using public clouds with serverless paradigm, creating our infrastructure on-demand. The goal is to allow deploying and running the ROOT analysis on public cloud resources. As a proof of concept, we developed a prototype of a new distributed processing backend to PyRDF interface, to support running ROOT analysis workflows on AWS Lambda in the same way as on the existing Apache Spark backend.

We report on our experience with serverless infrastructure, starting with compiling and deployment of the entire ROOT environment with its dependencies, through usage of CERN resources without any CERN-specific software, up to the point of connecting multiple Serverless Functions using PyRDF to mimic the existing PySpark environment. The technologies we employed include Terraform for deployment of our application, boto for client-AWS integration, Docker for simpler installation, and AWS S3, Lambda and EFS services for underlying infrastructure. We outline the issues, the possibilities, technical limitations and current roadblocks waiting to be solved for the tool to be used easily by anyone.

Acknowledgments:

We would like to thank the ROOT team for their support and discussions, in particular to Vincenzo Padulano, Enric Tejedor and Vassil Vassilev.

This work was supported by the Polish Ministry of Science and Higher Education, grant DIR/WK/2018/13.

Novel Data Science Environments / **136**

SWAN, Rucio, and Jupyter

Authors: Muhammad Aditya Hilmy[None]; Mario Lassnig[1]; Martin Barisits[1]; Riccardo Di Maria[1]; Diogo Castro[1]; Enric Tejedor Saavedra[1]; Enrico Bocchi[1]

[1] *CERN*

Corresponding Author: mario.lassnig@cern.ch

The LHC experiments at CERN produce an enormous amount of scientific data. One of the main computing challenges is to make such data easily accessible by scientists and researchers. Technologies and services are being developed at CERN and at partner institutes to face this challenge, ultimately allowing to turn scientific data into knowledge.

SWAN (Service for Web-based ANalysis) is a platform allowing CERN users to perform interactive data analysis directly using a web browser. This service builds on top of the widely-adopted Jupyter Notebooks. It integrates storage, synchronisation, and sharing capabilities of CERNBox and the computational power of Spark/Hadoop clusters. Both scientists at CERN and at partner institutes are using SWAN on a daily basis to develop algorithms required to perform their data analysis. Full analyses can be performed using Notebooks as long as all the required data are available locally.

The Rucio data management system was principally developed by the ATLAS experiment to deal with Exabytes of data in a scalable, modular, and reliable way. Nowadays, Rucio has become the

de-facto data management system in High Energy Physics and many other scientific communities such as astronomy, astrophysics, or environmental sciences are evaluating and adopting it.

In the Exabytes-scale era, the challenge to move large amounts of data in the local file system of a Notebook is faced on a daily basis by each individual scientist, causing duplication of effort and delaying the analysis results. The integration of Rucio in the Jupyter Notebook environment is a challenging but necessary R&D activity from which the worldwide scientific community would greatly benefit.

Starting from an idea at the previous CS3 conference, in less than a year a JupyterLab extension was developed and tested in the context of Google Summer of Code and the EU-funded project ESCAPE. This extension integrates Rucio functionalities inside the JupyterLab UI, to link experiment data into notebooks that require them, and to transparently make the data present in the ESCAPE DataLake available using Rucio.

Scalable Storage / **151**

Samba and CERNBox: experience in providing HA online access to Windows-based users

Authors: Giuseppe Lo Presti[1]; Aritz Brosa Iartza[1]

[1] *CERN*

Corresponding Author: aritz.brosa.iartza@cern.ch

This contribution presents the experience in providing CERN users with direct online access to their CERNBox storage from Windows. In production for about 15 months, a High-Available Samba cluster is regularly used by a significant fraction of the CERNBox user base, following the migration of their central home folders from Microsoft DFS in the context of CERN's strategy to move to open source solutions.

We will describe the configuration of the cluster, which is based on standard components: the EOS-backed CERNBox storage is mounted via FUSE, and an additional mount provided by CephFS is used to share the cluster's state. Further, we will describe some typical shortcomings of such a setup and how they were tackled.

Finally, we will show how such an additional access method fits in the bigger picture, where the EOS-backed storage can seamlessly be accessed via sync clients, FUSE/Samba mounts as well as the web UI, whilst aiming at a consistent view and experience.

Collaborative applications / **127**

Say hello to "SeaTable", the next generation spreadsheet

Author: Christoph Dyllick-Brenzinger[1]

[1] *datamate*

Corresponding Author: cdb@datamate.org

SeaTable is the new spreadsheet solution from the creative minds behind Seafile.

SeaTable will change the way people handle information and data. It helps to bring order and structure to any kind of information. In contrast to other spreadsheet solutions, SeaTable offers more than just text and numbers. The magic word here is rich text fields for all kind of information: files, photos, selection lists, checkboxes, e-mail addresses, links, URLs or people.

Views, Grouping and Sorting give every team member the perfect view of the data. As a web application, SeaTable allows any number of people to work together at the same time. Every line, every entry can be commented or expanded.

SeaTable is a tool for companies and teams of all sizes and does not require any prior IT knowledge. It is a flexible toolbox to organize data and build your processes.

The presentation will cover:

- Features
- Team
- Deployment options
- Development outlook

Scalable Storage / **185**

Scalable Metadata Management Using Onedata and OpenFaaS

Authors: Lukasz Dutka[None]; Michał Orzechowski[1]; Bartosz Kryza[2]

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

[2] *ACC Cyfronet-AGH*

Corresponding Author: lukasz.dutka@cyfronet.pl

Onedata 1 is a global high-performance, transparent data management system, that unifies data access across globally distributed infrastructures and multiple types of underlying storages, such as NFS, Amazon S3, Ceph, OpenStack Swift, WebDAV, XRootD and HTTP and HTTPS servers, as well as other POSIX-compliant file systems.
Onedata allows users to collaborate, share, and perform computations on data using applications relying on POSIX compliant data access. Thanks to a fully distributed architecture, Onedata allows for the creation of complex hybrid-cloud infrastructure deployments, including private and commercial cloud resources.

Onedata comprises the following services: Onezone - authorisation and distributed metadata management component that provides access to Onedata ecosystem; and Oneprovider - provides actual data to the users and exposes storage systems to Onedata and Oneclient - which allows transparent POSIX-compatible data access on user nodes. Oneprovider instances can be deployed, as a single node or an HPC cluster, on top of high-performance parallel storage solutions with the ability to serve petabytes of data with GB/s throughput.

Onedata introduces the concept of Space, a virtual volume, owned by one or more users, where they can organize their data under a global namespace. The Spaces are accessible to users via a web interface, which allows for Dropbox-like file management, a Fuse-based client that can be mounted as a virtual POSIX file system, a Python library (OnedataFS 2), or REST and CDMI standardized APIs. As a distributed system Onedata can take advantage of modern scalable solutions like Kubernetes and thanks to a rich set of REST APIs and OnedataFS library it can process at scale data and metadata alike using FaaS systems like OpenFass.

Currently Onedata is used in European Open Science Cloud Hub 2, PRACE-5IP [3], EOSC Synergy [4], and Archiver [5] project, where it provides data transparency layer for computation deployed on hybrid clouds.

Acknowledgements: This work was supported in part by 2018-2020's research funds in the scope of the co-financed international projects framework (project no. 3905/H2020/2018/2, and project no. 5145/H2020/2020/2).

1 Onedata project website. http://onedata.org.
2 OnedataFS - PyFilesystem Interface to Onedata Virtual File System. https://github.com/onedata/fs-onedatafs.
[3] European Open Science Cloud Hub (Bringing together multiple service providers to create a single contact point for European researchers and innovators.). https://www.eosc-hub.eu.
[4] Partnership for Advanced Computing in Europe - Fifth Implementation Phase. http://www.prace-ri.eu.
[5] European Open Science Cloud - Expanding Capacities by building Capabilities. https://www.eosc-synergy.eu.
[6] Archiver - Archiving and Preservation for Research Environments). https://www.archiver-project.eu.

On-premise, hybrid or cloud? / **161**

Science Mesh beyond science – perspectives for adoption in a wider business context.

Author: Marcin Sieprawski[None]

Corresponding Author: marcin.sieprawski@softwaremind.com

Sync & Share services have been in use in science and research communities for many years. We all see the potential of further integration of these platforms with research-oriented services – and a possible adoption of these technologies in a wider commercial context – but the question is when this happens, or even if this happens. Software Mind, a medium-sized software house, joined the CS3MESH4EOSC project with the belief, that the answer to the second question is "yes", and that we can use this to grow our cloud software development business.

But let us introduce ourselves. Software Mind, part of Ailleron group, is a global IT service provider based in Poland, delivering skilful managed teams for even most demanding projects. With Software Mind you can ramp-up an innovative, effective, agile development team in just a few weeks – to expand your teams, grow your startup or accelerate your IT. The company develops solutions based on cutting-edge technologies, including Big Data Integration, Internet of Things, semantic technologies, machine learning, cloud computing and Smart Cities.

In case you skipped the previous paragraph I'll give a condensed version, because you probably have not heard about us: we are an Agile-software-development company, we were one of the first commercial users of Hadoop back in early 2004 when we provided the technology for the first web-scale Semantic Web startup (with Tim Berners-Lee in the team), now we mostly develop microservice solutions in the cloud.

As you probably guess, currently we develop solutions based on American hyperscalers' walled gardens: AWS and Azure. But we have some ideas how this may change.

In CS3MESH4EOSC project we provide the expertise on microservices architecture, integration, DevOps, agile software development process and Data Science. We lead tasks on Reference interoperability platform and distributed Data Science environments. In this talk I'll show how we see this as a part of the strategy of growing our business of application services in the cloud, microservice-based architectures, Data Science, Big Data integration and analytics. Maybe I'll even share some of our secrets, if you promise not to tell anyone.

Science Mesh workshop / **179**

Science Mesh for site administrators: operation, security, trust

Corresponding Author: ron.trompert@surf.nl

Science Mesh workshop / **178**

Science Mesh in a nutshell

Corresponding Author: pedro.ferreira@cern.ch

Overview of objectives, activities and assets of the CS3MESH4EOSC project

Science Mesh workshop / **182**

ScienceMesh for developers: how to contribute to CS3APIs and IOP

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

Information on processes for contributing to the design and development of the Science Mesh

Science Mesh workshop / **181**

ScienceMesh for users: applications, use cases & workflow demo

Corresponding Author: maciekb@man.poznan.pl

Technical demo presenting the initial Science Mesh services

Tech Short Talks / **128**

SeaTable: easy like a spreadsheet, powerful like a database

Author: Christoph Dyllick-Brenzinger[1]

[1] *datamate*

Corresponding Author: cdb@datamate.org

At first glance, SeaTable looks like an Excel spreadsheet, but it is as powerful as a database. In this presentation I will show you with concrete examples how seatable positions itself in comparison to Excel and a traditional SQL database.

After this presentation you will understand what kind of tasks you can perform with SeaTable and the limitations that can prevent you from using SeaTable.

The presentation will cover:

- Comparison of Excel and SeaTable with regards to content types, collaboration, data crunching, automation
- Strengths and Weaknesses of SeaTable in comparison to a SQL-database
- Concrete Use Cases

OCM Workshop / **190**

Seafile and OCM

Corresponding Author: jonathan.xu@seafile.com

File Sync&Share Products for Home, Lab and Enterprise / **129**

Seafile: Review and Next Steps

Author: JiaQiang Xu[None]

Corresponding Author: jonathan.xu@seafile.com

Seafile is one of the most popular open sourced and self-hosted cloud storage solutions. The main advantage is its focus files, providing high performance and reliability. It's widely used by many educational and research institutions, such as HU Berlin, University of Mainz, Max Planck Society, PSNC.

In this talk we'll review new features added in Seafile 7.1 and 8.0 version and look into the roadmap of Seafile. Main topics include:
- What's new in Seafile server 7.1 and 8.0
- Update about SeaDrive client 2.0
- Roadmap update for Seafile development

Site reports / **133**

Site Report sciebo

Authors: Marcel Wunderlich[1]; Holger Angenent[2]

[1] *University of Muenster*

[2] *University of Münster*

Corresponding Author: wunderlich@uni-muenster.de

Sciebo is an owncloud based sync and share solution for academic institutions in North Rhine-Westphalia, Germany. It was started in 2015 and is operated by the University of Münster. With close to 200k users and over two petabyte of used storage, the hardware and administrative requirements are quite demanding.
In its current form sciebo runs an ownCloud instance for each participating institution, distributed across three sites at the universities of Bonn, Duisburg-Essen and Münster. All instances share access to a single OnlyOffice installation in Münster.
Currently we are migrating the three sites to a redundant two-site setup in Münster. Towards a more reliable deployment with version controlled configurations and software versions, this includes moving operations from manual ssh-ing and semi-automated updates with Ansible to a declarative deployment with Kubernetes. Each site runs an on-premise Kubernetes cluster, with a GPFS containing the user data cluster spanned across both sites. Version control, workflow management and documentation are unified in the university's own GitLab instance.

FAIR and Open Research Data Services / **145**

Sync and Share for Research Data Management

Author: Tom Wezepoel[1]

[1] *SURF*

Corresponding Author: tom.wezepoel@surf.nl

Since a number of years SURF has been running a sync-and-share service called SURFdrive as a personal cloud storage system. Later on it appeared that this service could not fulfill all the requirements coming from the research community. This had to do with flexible quota, project-based storage rather than personal storage and multiple means of authentication. The latter was an absolute necessity in order to allow people accessing the service outside of the Dutch SURFconext identity federation. For this reason SURF has started Research Drive that is delivered mostly to higher educational and research institutes in the Netherlands. This has become a great success with 20+ different sync-and-share instances operated by SURF.

The institutes use this service to manage their research data. Apart from the regular users like students, teachers and researchers there are also the local IT, primary investigators and data stewards. Each having their role in the data management process. Since Research Drive is completely self-service we have developed a dashboard where these different roles have been implemented, each with the different capabilities suiting their role.

In this presentation we will give an overview of how a sync-and-share service like Research Drive can be used for Research Data Management at an institutional level.

OCM Workshop / **189**

The OCM test suite

Corresponding Author: michiel@unhosted.org

OCM Workshop / **188**

The State of OCM

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

Tech Short Talks / **155**

The benefits of using Nextcloud Talk

Author: Olivier Paroz[1]

[1] *Nextcloud GmbH*

Corresponding Author: olivier.paroz@nextcloud.com

Today's organisations are increasingly looking for tools which allow their members to communicate with one another while collaborating online.
And while some are focused on trying to get various solutions to talk to one another, there are several advantages in using an integrated solution such as Nextcloud Talk on top of an existing Nextcloud instance.

Scalable Storage / **143**

The ocis storage driver - a deconstructed filesystem

Author: Jörn Dreyer[1]

[1] *ownCloud GmbH*

Corresponding Author: jfd@owncloud.com

Looking up files by a stable id is an inefficient operation in most filesystems.
While an efficient lookup by file id can be cached inside an OCIS storage provider this cache needs to be kept up to date. By deconstructing a filesystem and storing every node by its uuid we can evade the cache invalidation problem at the cost of more stat requests. The ocis storage driver allows an efficient lookup of file metadata by path and by file id while relying solely on the filesystem as a persistence layer. Furthermore, the layout on disk can be used to implement trash, versions and in the future deduplication. It serves as a blueprint for seperating metadata and content in an S3 or librados storage driver.

Welcome and Keynote / **172**

The triangle of digitalization – sustainability and democracy within digital collaboration

The keynote presentation discusses three theses concerning the triangle of digitalization, sustainability and democracy.

First: How we design this triangle will determine to a large extent the quality of our livelihoods and of our democracy. The question is not whether digitalization plays a crucial role, but how it will play out regarding sustainability and democracy.

Second: Digitalization has the potential to be a driving force for the needed transformation of the energy, transportation, building, industry and housing sectors on the path towards greenhouse gas neutrality and circular economy (e.g. European Green Deal). But so far it is more a driver for increased emissions and the use of resources than a limiting factor.

Third: The legitimization of democracy is based on two pillars: representation and deliberation. Digitalization has a huge potential to improve both representation (without people even being physically present) and deliberation. However, currently it rarely improves representation and it rapidly undermines deliberation in society. Digitalization can be used to increase transparency of governments and businesses and to limit their power, but it can also be used as an instrument to control people and cement power of businesses and governments.

At the end, the question will be raised whether time has come to stop using naively all kinds of digitalization as if different kinds of digitalization did not create very different but decisive path dependencies determining our future.

Christoph Bals is a policy director at Germanwatch NGO. Germanwatch is dedicated to sustainable development and topics such as World Trade and Food Security, Climate Protection and Adaptation, Corporate Accountability, the Financial Sector and Sustainability as well as the Financing of Development Cooperation.

Site reports / **146**

To Quality, and Beyond!

Author: Bradley William Marshall[None]

Corresponding Author: brad.marshall@aarnet.edu.au

AARNet provide CloudStor, Australia's largest EFSS suite for research collaboration, with end points distributed across the nation. As the number of users scale so has our related products, with a corresponding increase in complexity.

2020 has provided many challenges for the Cloud Services team at AARNet, among those keeping the services stable while scaling to meet demand. As part of our response to these issues we have focused on quality, with the deployment of a Openstack cluster to improve our ability to test changes and new features.

Our first aim with this platform was to create a deployment of the components of CloudStor, using Terraform to spin up the VMs and ansible (using the same scripts we do in production) for the services. The next step was to deploy the CS3Mesh IOP platform, as this involved Kubernetes and helm charts. We've also taken our first steps into continuous integeration with automated Docker containers being built post git commits, with some linting and testing as part of it.

The ultimate aim is to have a fully tested automated deployment of the full environment, including monitoring and performance testing, but the journey has just started.

Tech Short Talks / **142**

Towards push notifications in OCIS

Author: Jörn Dreyer[1]

[1] *ownCloud GmbH*

Corresponding Author: jfd@owncloud.com

To make clients pick up changes to a shared file the etag of all recipients root folders needs to be updated. The current implementation in OCIS jails shares into a /Shares folder to calculate a dynamic etag, based on all accepted shares. This multiplies the stat requests made to the underlying storage system by the number of shares on every propfind. By letting the storage registry cache the root etag of every storage id we can reduce this number to one stat request per storage. This cache would not only allow the gateway to calculate the etag for any path based on the mount point of storages: it would allow sharing the stat cache for all clients, the ocis instance could dynamically adjust the cache timeout to react to system load, and storages providers can push etag changes to the storage registry to prepare for push notifications to clients.

OCM Workshop / **187**

Welcome

Corresponding Author: pedro.ferreira@cern.ch

Welcome and Keynote / **171**

Welcome & Warm-up

Corresponding Author: jakub.moscicki@cern.ch

Science Mesh workshop / **177**

Welcome and objectives

Corresponding Author: jakub.moscicki@cern.ch

Scalable Storage / **130**

iRODS Policy Composition: Configuration, Not Code

Author: Terrell Russell[1]

[1] *iRODS Consortium / RENCI*

Corresponding Author: tgr@renci.org

With a twenty-five year history, iRODS open source technology has been used to automate data management across many scientific and business domains. The scale and value of data across these domains drives the necessity for automation. This variety also demands a flexibility in data management policies over time. Organizations have satisfied their own needs by investing in the development of their own policy, but this has led to monolithic, specific policy sets tailored to a particular organization.

As a community, we have observed common themes and duplication of effort across these organizations and now worked to provide generalized implementations that can be deployed across multiple domains. This new approach allows multiple policies to be configured together, or composed, without the need for custom development. For example, our Storage Tiering capability is a composition of several basic policies: Replication, Verification, Retention, and the Violating Object Discovery.

File Sync&Share Products for Home, Lab and Enterprise / **140**

ownCloud - News and Roadmap

Author: Patrick Maier[None]

Corresponding Author: pmaier@owncloud.com

On mission for digital sovereignty, ownCloud is not only the center for an organization's unstructured data but will also provide the integrated, digital workplace of the future. Equipping these pillars with a powerful set of workflows, ownCloud is the platform for secure digital collaboration that keeps data under control.
In this talk we give an update on the recent achievements around the ownCloud platform and shed light on the milestones that lie ahead of us.