Evolution of the CERNBox platform to support the Malt project

Hugo González Labrador[1,*], *Vincent Nicolas* Bippus[1], *Sebastian* Bukowiec[1], *Diogo* Castro[1], *Sebastien* Dellabella[1], *Michal* Kwiatek[1], *Giuseppe* Lo Presti[1], *Luca* Mascetti[1], *Jakub T.* Mościcki[1], *Esteban* Puentes[1], *Piotr Jan* Seweryn[1], *Apostolos* Smyrnakis[1]

[1]CERN, 1 Esplanade des Particules, Meyrin, Switzerland

Abstract.
CERNBox is the CERN cloud storage hub for more than 25,000 users at CERN. It allows synchronising and sharing files on all major desktop and mobile platforms (Linux, Windows, MacOSX, Android, iOS) providing universal, ubiquitous, online- and offline access to any data stored in the CERN EOS infrastructure. CERNBox also provides integration with other CERN services for big science: visualisation tools, interactive data analysis and real-time collaborative editing.

Over the last two years, CERNBox has evolved from a pure cloud sync and share platform into a collaborative service, to support new applications such as DrawIO for diagrams and organigrams sketching, OnlyOffice and Collabora Online for documents editing, and DXHTML Gantt for project management, as alternatives to traditional desktop applications. Moving to open source applications has the advantage to reduce licensing costs and enables easier integration within the CERN infrastructure. This move from commercial software to open source solutions is part of the MALT project, led by the IT department at CERN to reduce the dependencies on commercial solutions.

As part of the MALT project, CERNBox is the chosen solution to replace Home directories of the Windows DFS file system. Access to storage from Windows managed devices for end-users is largely covered by synchronization clients. However, online access using standard CIFS/SMB protocol is required for shared use-cases, such as central login services (Terminal Services) and visitor desktop computers. We present recent work to introduce a set of Samba gateways running in High Availability cluster mode to enable direct access to the CERNBox backend storage (EOS).

1 Introduction

The background of the MALT project was provided in CERN Computing Newsletter [1]:

> Over the years, CERN's activities and services have increasingly relied on commercial software and solutions to deliver core functionalities, often leveraged by advantageous financial conditions based on the recognition of CERN's status as an academic, non-profit or research institute.

[*] e-mail: hugo.gonzalez.labrador@cern.ch

> Given the collaborative nature of CERN and its wide community, a high number of licenses are required to deliver services to everyone, and when traditional business models on a per-user basis are applied, the costs per product can be huge and become unaffordable in the long term.

The initial objective of the MALT project is to investigate the migration from commercial software products to open-source solutions. The project also helps to minimise the exposure of CERN to the risks of unsustainable commercial conditions. By doing so, the laboratory is playing the pioneering role among public research institutions, especially in Europe, most of which have recently been faced with the same dilemma as well as questions about Digital Sovereignty.

The IT Storage group participates actively in this important project exposing alternatives to commercial software to the CERNBox[3–5] user community (more than 25,000 user accounts).

2 CERNBox

The CERN IT Storage group is responsible for ensuring a coherent development and operation of the storage services at CERN for all aspects of physics data. The group supports the data requirements for the experiments at CERN and as well as supporting storage services for the whole CERN users community (CERNBox and AFS).

CERNBox is the CERN cloud storage hub and has become one of de facto platforms for collaboration at CERN. CERNBox was launched 2014 to address the necessity of offering an easy and convenient way to access and share the physics data. It was also introduced to better protect the data of the organization by ensuring clear data protection policies (all CERNBox data is stored on CERN premises) and clear data management policies.

The goal with CERNBox is to consolidate different home directory use-cases (DFS for Windows users and AFS for Linux users) into only one service that can be accessed from any platform satisfying most of the requirements of the systems currently running.

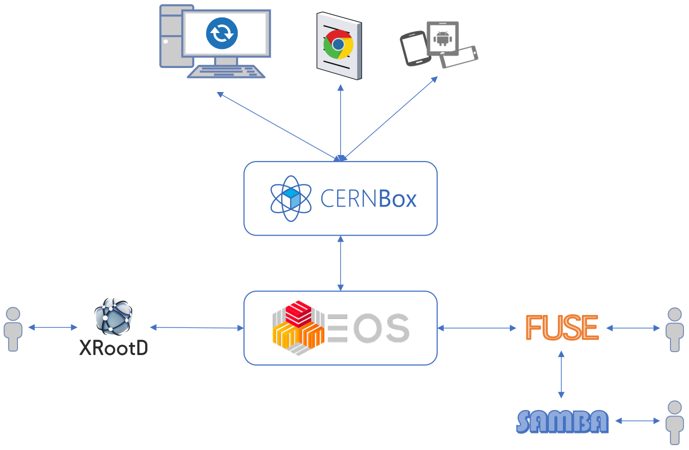

Figure 1: How to access CERNBox

CERNBox is based on two open source components: ownCloud[2] and EOS[16, 17]. ownCloud is an open source cloud collaboration software and EOS is CERN open source elastic, adaptable and scalable storage system for storing the data generated by the LHC.

Users can connect to CERNBox using various methods as depicted in Figure 1:

- FUSE access: EOS storage can be mounted as a local file-system. This is important for computers inside the computer centre that do not have graphical interfaces and need to perform batch job processing. Also, many analysis tools are not adapted to work with remote storage systems and they expect a local filesystem.
- WebDAV: the storage is accessible through WebDAV enabled clients like Finder for OSX, Windows Network Drives for Windows or third-party tools like CyberDuck, which brings a user-friendly access methods to end-users.
- XROOTD: the storage can also be accessed by using the xrootd high performance, scalable and fault tolerant access protocol.
- Synchronization client: this is a component, integrated in the user desktop environment, which keeps one (or more) local folders in synchronization with the central storage server. Users can work on their devices without connectivity and the synchronization client reconciles changes when the network connectivity is restored. This component enables users to work offline.
- Web access: via any web browser the user can manipulate files and share them with other users. Users can also take advantage of the available applications in CERNBox to work with certain file types. See section 3 for more information on the applications.
- Mobile devices: CERNBox provides mobile and tablet applications for Android and iOS to access the data.
- Samba (SMB) access: CERNBox provides online access via the SMB/CIFS networking protocol, used mainly by Windows clients. SMB support was added due to the increasing demand for online access from Windows client, following the ongoing migration from DFS to CERNBox. Refer to sections 4 and 5 for more details.

3 MALT Applications

CERNBox offers a convenient and universal (browser-based) access to a collection of useful applications, designed to boost the productivity of our user community. Users can just use these applications without having to install any software on their computers. A web browser is the only tool needed to access a catalog of diverse applications, ranging from basic text editors to complex and high performance web based software analysis platforms such as SWAN[18]. In this section we explore a variety of applications in close relationship with the MALT objectives: replacing commercial applications with open source alternatives.

Editing of Microsoft Office documents has been available in CERNBox for a number of years using integration with SharePoint. This enabled our user community to view and edit documents online, both standalone and to a certain extent in a collaborative fashion.

In an attempt to improve collaborative edition, as well as to identify suitable alternatives in the market, we are proposing to our user community other similar products. OnlyOffice is a cloud-oriented solution with a modern and intuitive user interface allowing multiple people to collaborate on documents, spreadsheets and presentations [11]. Currently, we are finishing our testing phase using Canary Testing Mode, a feature flag mechanism of CERNBox which allows to expose volunteer users to new applications.

At the moment, OnlyOffice is deployed as a single container application running on a bare metal server. This has been largely sufficient for the testing phase, where the number of concurrent users has never exceeded some tens. However, for the long term operations and scalability of the new service, we are planning to use our Puppet-managed infrastructure on top of OpenStack. Following vendor's deployment recommendations, the initial plan includes installation of two front-end OnlyOffice Document Servers in a high-availability configuration, as well as other required back-end services on separate VMs.

In addition to OnlyOffice, Collabora Online [12] also offers a cloud-based service specifically optimized for the Open Data file formats. After an initial successful proof-of-concept integration in CERNBox, we plan to prototype its integration in the near future.

As far as other office productivity applications are concerned, Gantt Chart Viewer is an application designed for users who just need to view Gantt charts [10]. Draw.io is an application that allows opening and creating different types of diagrams [13]. Both applications were integrated in CERNBox using dedicated integration applications based on the work available at [15]. The diagram presented in Figure 2 describes the data flow between these tools and CERNBox. Once the user logs in and opens a file, the integration application creates an *iframe* and loads a particular tool by setting its URL. The file data is transferred using the *window.postMessage* API, part of the HTML5 standard. Integration uses the *client.js* library for obtaining the files' content either as a text or an array buffer. Finally, both applications depend on micro services hosted on-site for importing and exporting files in different formats, which allows us to keep our users' data on premise.

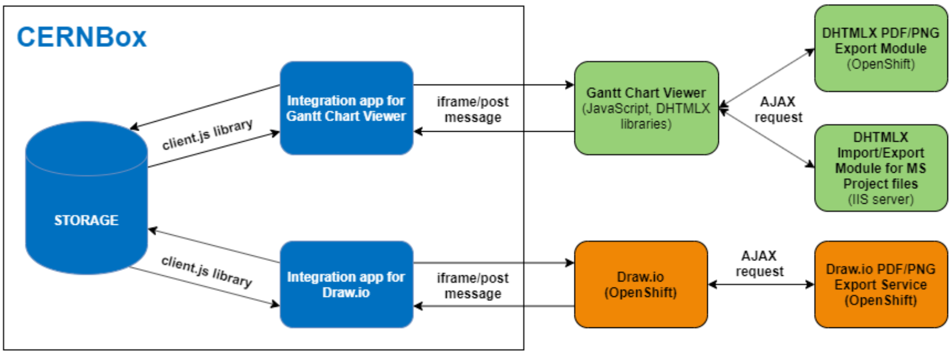
CERNBox
STORAGE
client.js library
client.js library
Integration app for Gantt Chart Viewer
Integration app for Draw.io
iframe/post message
iframe/post message
Gantt Chart Viewer (JavaScript, DHTMLX libraries)
Draw.io (OpenShift)
AJAX request
AJAX request
DHTMLX PDF/PNG Export Module (OpenShift)
DHTMLX Import/Export Module for MS Project files (IIS server)
Draw.io PDF/PNG Export Service (OpenShift)

Figure 2: Integration of Gantt Chart Viewer and Draw.io with CERNBox.

4 SMB Access to CERNBox

With the increased demand of online access to the CERNBox backend storage (EOS) from Windows, a cluster of SMB/CIFS gateways has become critical. To this end, in order to provide a High Available (HA) setup, the Samba and CTDB software suites [14] have been deployed on a cluster of four nodes with an IP-based load balancing over floating IPs, as required by the software. The cluster is in production since September 2019, and it permanently serves several hundreds Windows clients in the CERN network.

This cluster is instrumental to perform the ongoing mass migration of CERN Windows-based users from DFS to CERNBox, as detailed in the following section.

5 DFS Migration to CERNBox

In the context of the MALT project, the CERN users' Windows Home directories are being migrated from the Distributed File System (DFS) into CERNBox. The devised procedure involves the following three steps:

- Communication with the user
- Provisioning of a CERNBox home and data/metadata copy
- Configuration of end-user computer(s) to switch to CERNBox

5.1 Communication with the user

Communicating with users is of essential importance: the DFS team contacts the users one week before the migration to provide them with all the necessary information and instructions for the procedure. The users are given the option to postpone the intervention in case they have important work constraints on the suggested date. The most important action points in this communication are two:

1. Users need to save their work and do not access their files during the migration time-window
2. The DFS Home folder will be read-only after the migration

5.2 CERNBox home provisioning and data copy

An overnight automated procedure is provisioning a new Home directory in CERNBox, creating also the necessary folder layout structure to accommodate the typical folders found on a Windows system:

- Documents (from DFS "Documents" folder, excluding the sub-directories My Music, My Pictures and My Video)
- Music (from the DFS "Documents/My Music" folder)
- Pictures (from the DFS "Documents/My Pictures" folder)
- Videos (from the DFS "Documents/My Video" folder)
- WINDOWS/Desktop (from the DFS "Desktop" folder)
- WINDOWS/Favorites (from the DFS "Favorites" folder)

Under no circumstances this automated procedure overwrites or deletes existing data. Files and folders are already safely copied in CERNBox. In the case the user had a CERNBox space already, the script checks for name clashes; in case they exists, the migration procedure will not start. To copy the data from DFS to CERNBox, a server side copy procedure is implemented. Before the selected migration date, data is copied transparently from DFS to CERNBox progressively, using the *rsync* tool.

5.3 Final configuration on end-user computer(s)

Once users log into their computers, a script is automatically run to configure their machines to switch from DFS to CERNBox. The script identifies the main user of the computer and applies one of the following configurations depending on the device:

1. Configuration of the CERNBox sync client to store data locally: directories listed above are saved locally in the Hard Drive. The CERNBox client will use the local `cernbox` path to synchronize its content with the server. The Windows Registry is then configured accordingly for the typical Windows folders (refer to Section 5.2).

2. Configuration of a network mapped drive using the SMB protocol: in this case, data is not stored locally. A network drive is created pointing to the SMB gateway for CERNBox and the Windows Registry is updated to point to the corresponding remote folders. This configuration is suggested for shared computers (e.g. meeting room PCs) where having the client synchronising the profile of many users would not be optimal and could result in filling up the Hard Drive quickly.

6 Outlook

We plan to introduce more application integrations with alternative open source solutions, as discussed in section 3, to reduce our dependencies on existing commercial software. The usage and easy of adoption of these alternative products will be monitored during the coming years. At the same time, we aim at increasing the productivity of our users in their daily workflows, in particular when it comes to uploading and downloading files from different clouds: for example, many users use the Overleaf platform to edit LaTeX documents to prepare scientific papers, and the data to prepare them typically is already in CERNBox. We are also exploring a concept we name *Bring Your Own Application*, to efficiently allow users and developers to connect their preferred applications to CERNBox (refer to [9] for a more detailed explanation) using the micro-services deployment pattern [6, 7].

References

[1] E. Ormancey, *Migrating to open-source technologies*, CERN Computing Newsletter 12 June 2019, https://home.cern/news/news/computing/migrating-open-source-technologies

[2] ownCloud, https://owncloud.com (access time: 10/03/2018)

[3] Mascetti, Luca and Labrador, H Gonzalez and Lamanna, M and Mościcki, JT and Peters, AJ, Journal of Physics: Conference Series **664**, 062037, *CERNBox + EOS: end-user storage for science* (2015)

[4] H. G. Labrador, *CERNBox: Petabyte-Scale Cloud Synchronisation and Sharing Platform* (University of Vigo, Ourense, 2015) EI15/16-02

[5] H. G. Labrador et all, EPJ Web of Conferences **214**, 04038, *CERNBox: the CERN storage hub* (2019)

[6] Thönes, Johannes, IEEE software 32.1 *Microservices*, (2015)

[7] Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L, Springer, Cham, *Microservices: yesterday, today, and tomorrow. In Present and ulterior software engineering* (2017)

[8] WOPI, https://wopi.readthedocs.io/en/latest/ (access time: 10/03/2020)

[9] Hugo G. Labrador al., *Increasing interoperability: CS3APIS*, Proceedings of 24th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2019)

[10] Maria Alandes Pradillo et al., *Experience Finding MS Project Alternatives at CERN*, Proceedings of 24th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2019)

[11] Ascensio System SIA, https://www.onlyoffice.com/ (access time: 10/03/2020)

[12] Collabora Ltd, https://www.collabora.com/ (access time: 10/03/2020)

[13] Draw.io (Seibert Media GmbH), https://about.draw.io/ (access time: 10/03/2020)

[14] Samba, https://www.samba.org (access time: 10/03/2020)

[15] Pawel Rojek et al., https://github.com/pawelrojek/nextcloud-drawio (access time: 10/03/2020)

[16] Peters, AJ and Sindrilaru, EA and Adde, G, Journal of Physics: Conference Series **664**, 062037 (2015)

[17] Peters, Andreas J and Janyst, Lukasz Journal of Physics: Conference Series **331**, 052015 (2011).

[18] Piparo, Danilo, et al. "SWAN: A service for interactive analysis in the cloud.", **1071-1078**, Future Generation Computer Systems 78 (2018).