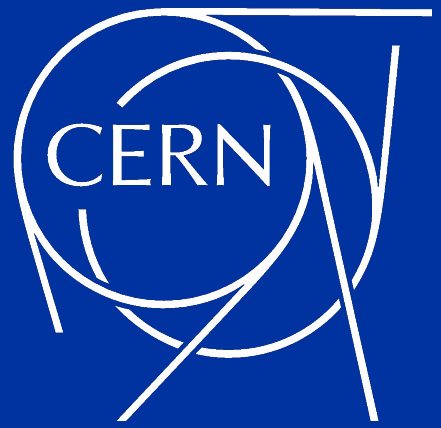


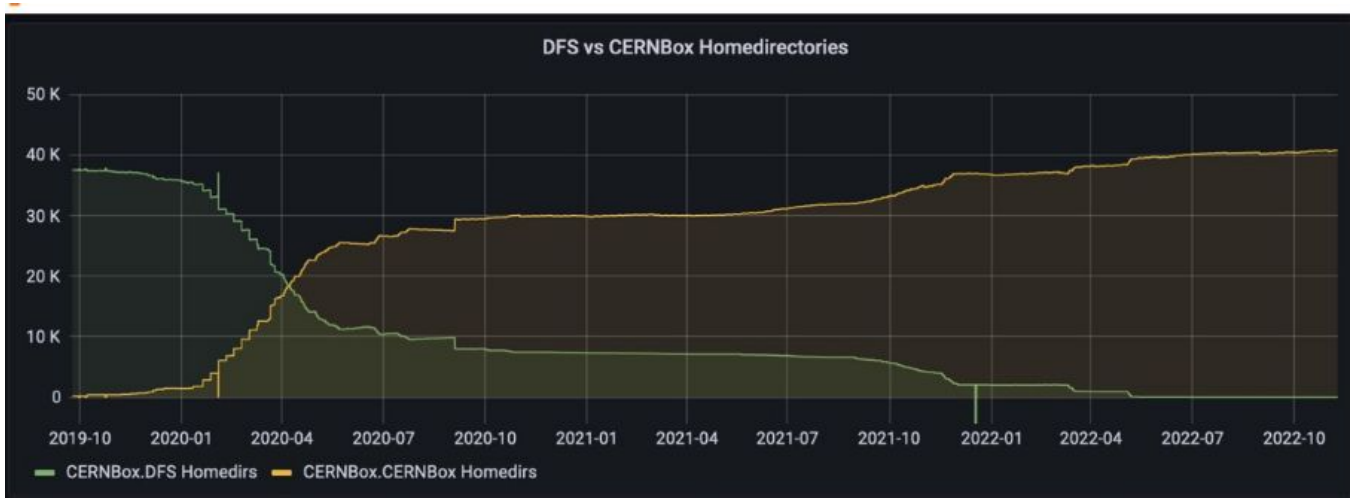
CERN Storage Services Status and Outlook for 2023

Julien Leduc for CERN IT-SD

30/03/2023

Future of storage for Windows at CERN (DFS)

- **In 2019 it was decided to discontinue the DFS service at CERN in order to consolidate storage solutions and rationalise resources.**
  - 40k user home directories migrated to CERNBox [DONE]
  - 3.5k DFS project spaces [In progress]



Target storage backend for migration is currently under investigation: possible combination of CephFS and EOS/CERNBox depending on the project requirements

CEPH BC/DR: Consolidation of critical-power storage

- Due to risk of electricity supply interruptions over winter,
  the Ceph team improved storage **Business Continuity** through the consolidation of:
  - **RBD** block volumes,
    by migrating critical ones to a cluster fully contained in the diesel-backed region of the CERN DC
  - **CephFS** through the physical move of an entire cluster

- **RBD consolidation** required migration of volumes:
  - Achieved through OpenStack Cinder, with downtime (mirroring features not mature enough)
  - 150+ bootable and storage volumes (~120 TB), each one in coordination with owner

- **CephFS consolidation** achieved with the **physical relocation of an entire cluster**:
  - Full-flash storage, 18x2U nodes, 16x1.92 TB SSDs each, ~550 TB raw capacity
  - Cluster hosts data for GitLab, TWiki, Linuxsoft, K8s/OKD clusters, …
  - **Physical move completed in 6 days** (3 batches of 6 nodes each), with **zero downtime**
    - Metadata operations blocked for ~40 seconds when moving metadata server

CEPH BC/DR: Ongoing efforts for improved resiliency

- The Ceph infrastructure is being rationalized and expanded
  to provide independent storage zones and harden upstream features for backups.

- Inherent complexity from:
  - **Diverse storage offer** – Blocks, Files, Objects
  - **Different BC/DR needs**, from high(er) availability (**active-active**) to Offline copy (**backup-and-restore**)
  - Integration with compute environment (OpenStack) and user-facing self-service portals

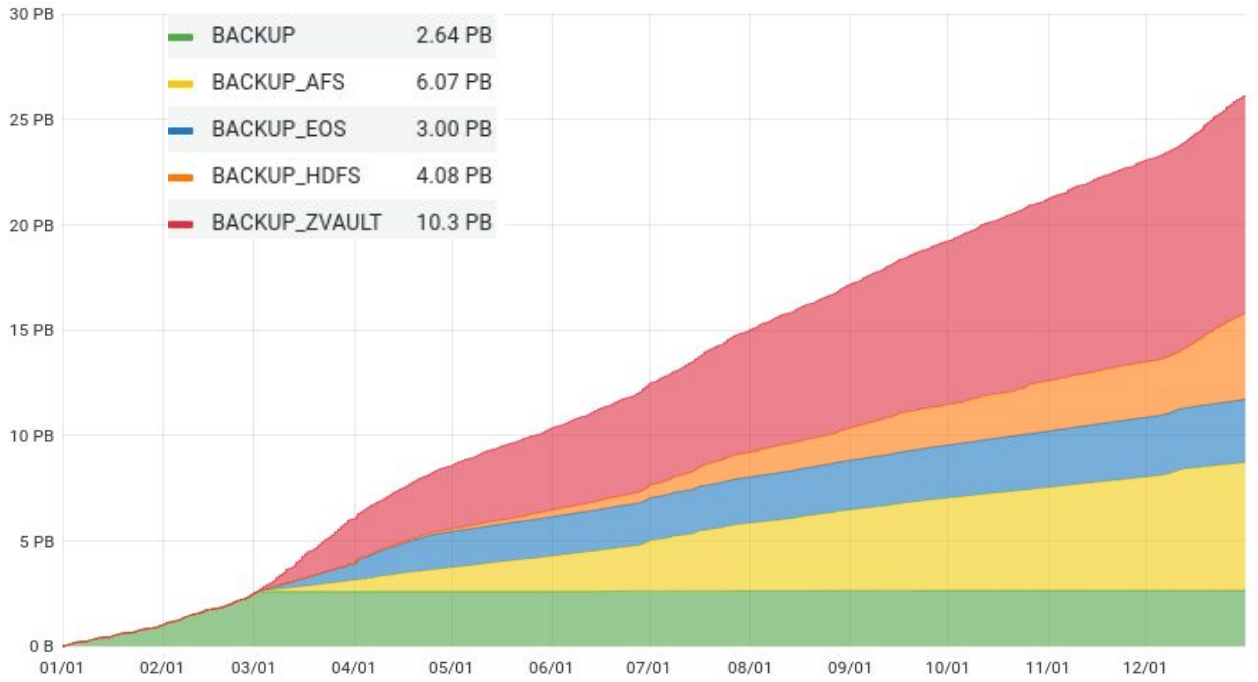
- Several options under evaluation and testing:
  - Multi-DataCentre **stretch** clusters and storage **availability zones**
  - RBD block volumes mirroring and backups
  - CephFS snapshots and **backups to S3 (and tape)** via restic
  - **Multi-site object storage**, backed by archival zones or external cloud resources
  - S3-based **immutable backups** through object locks and versioning

BC/DR: CTA BACKUP use cases

Demanding internal IT Storage backups already performed on eosctabackup instance

- **Over 2PB of monthly write traffic**
- **automatic repack in place for pruned backup data**
- **backup of all eoscta instance buffer namespace**
  - including itself

Cumulated backup volume written to eosctabackup instance in 2022

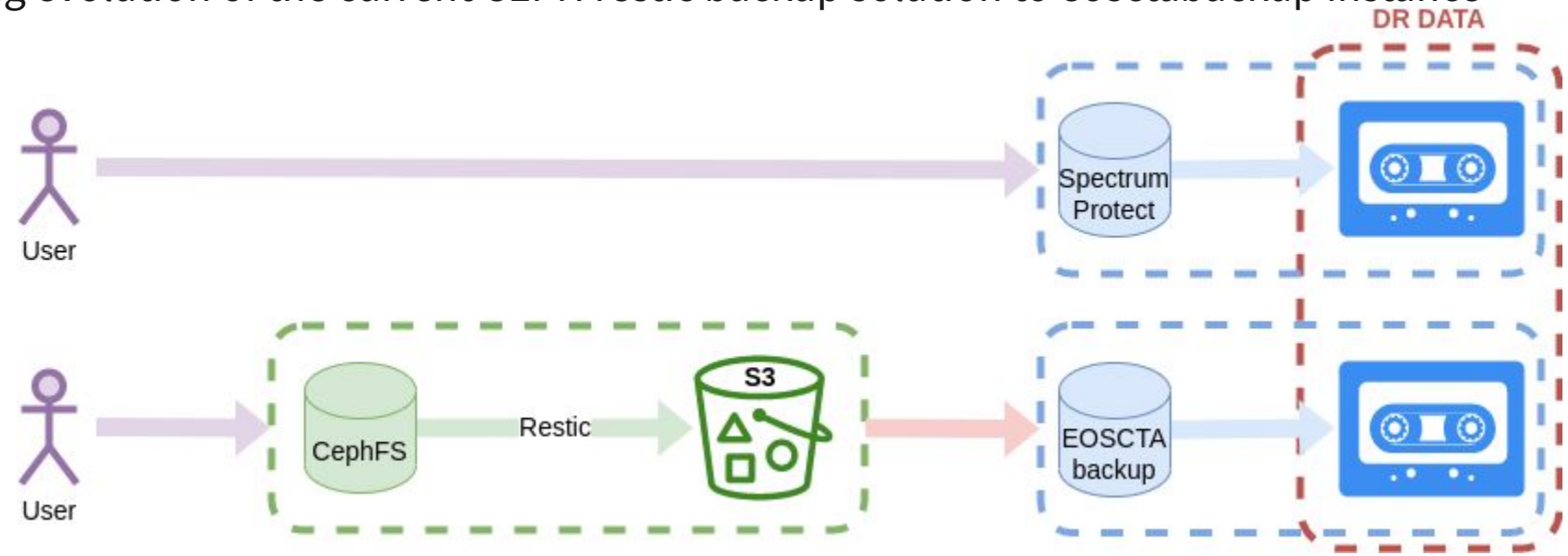


BC/DR: Storage group role as backend provider

Critical data are currently backed up to Spectrum Protect based Backup service and eosctabackup instance

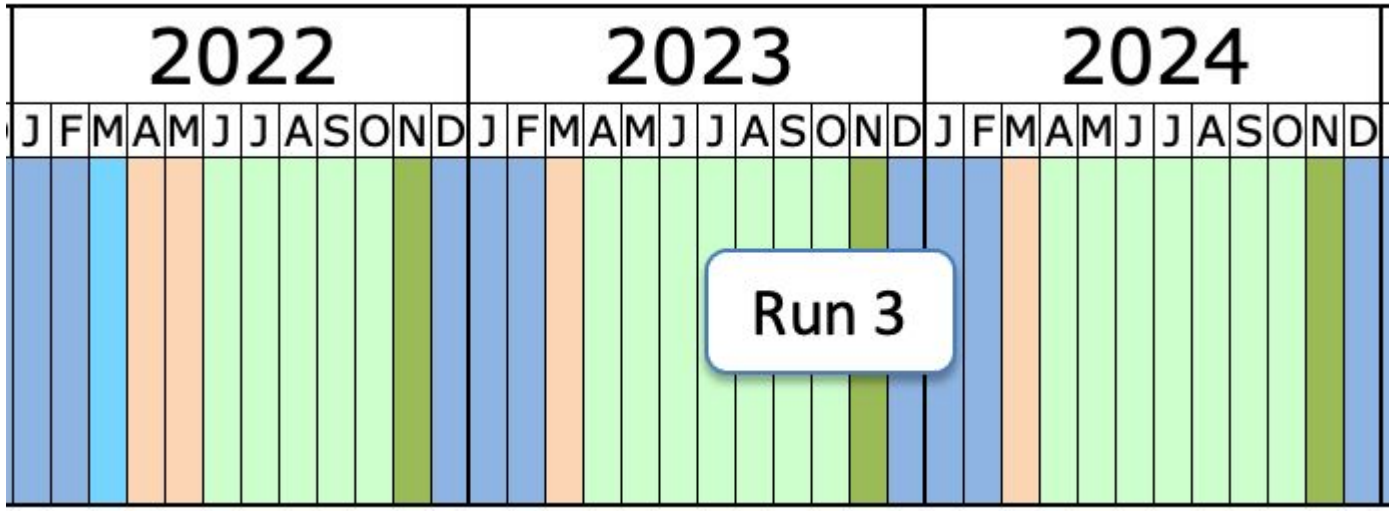
Current DR review in CERN IT will increase backup volume

- **Tape storage can provide *air gap* layer needed for various disaster scenarios**
- **Ceph S3 is already a user backup backend**
  - Planning evolution of the current CEPH restic backup solution to eosctabackup instance



Physics Run3 update

Plan



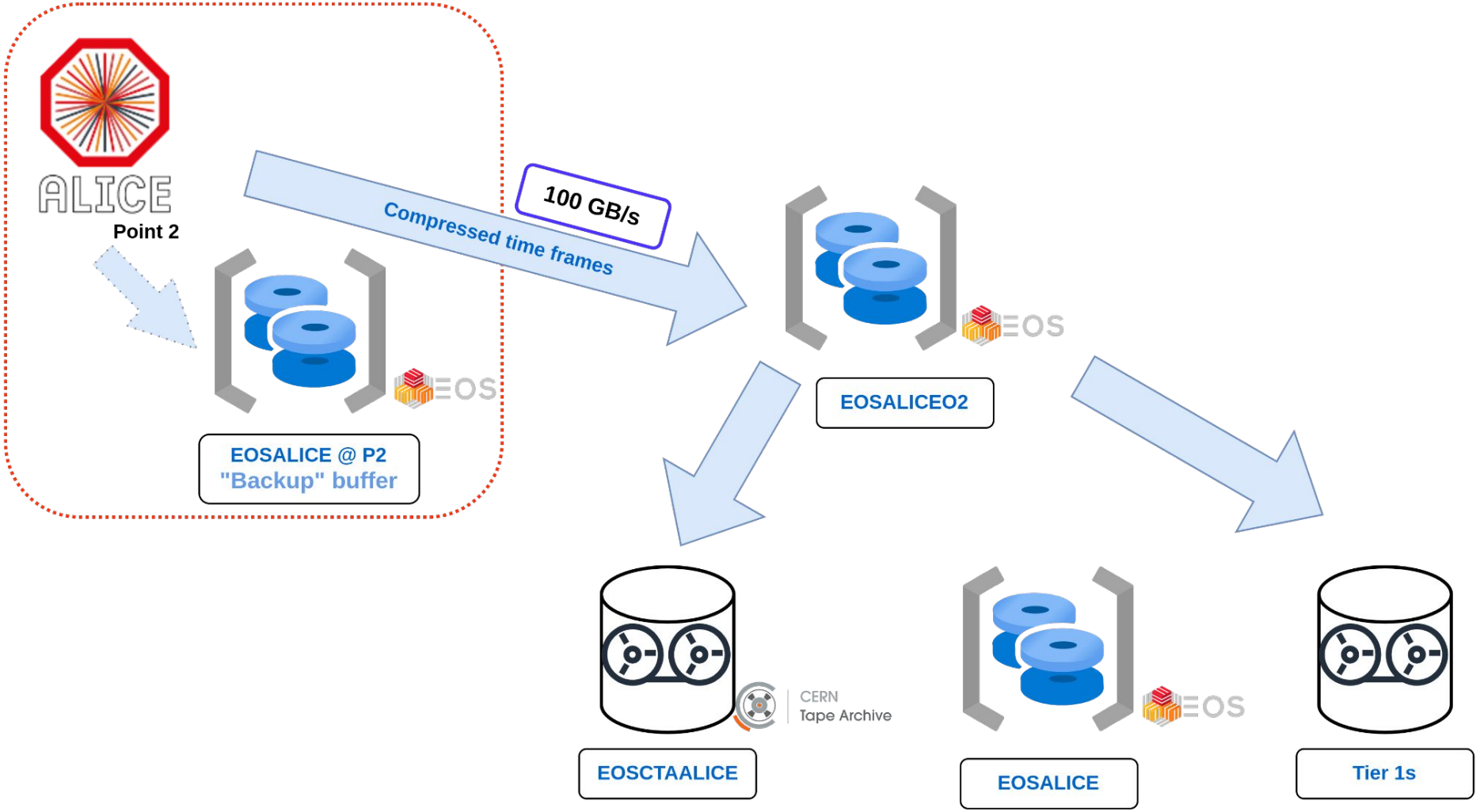
In practice it is more complicated…

- **Shorter overall run expected during 2023**
- **All experiments confirmed to stick to agreed number presented during HEPIX fall 2022…**

  …for proton run

- **First Heavy Ion run during Run3**
  - expected to last longer than initially planned to compensate HI absence during 2022
  - several uncertainties will require additional safety margins from storage physics infrastructure

First HI run during Run3: ALICE workflow and infrastructure review



First HI run during Run3: ALICE workflow and infrastructure review

- **+50 PB capacity (+50%) approved by IT-Steering Committee**
  - our estimations: 6 weeks of HI run at 150GB/s (modulo duty cycle) will generate 150PB of data
- **New capacity added to the ALICEO2 instance: +20PB (+20%) (March 2023)**
  - Total raw capacity: 137PB raw
    - Logical capacity: 114PB in EC 10+2
  - Total number of nodes: 100
    - single 100Gbps NICs, mostly AMD CPUs
    - running alma 8
  - Total number of disks: 9600 (mix of 12TB, 14TB and 18TB)
  - Scheduling groups: 384 (24/26 disks per group, ~360TB / group on average)
- **+30PB will be added toward HI run**

ALICE O2 March 23 new capacity benchmarks

READ or WRITE:

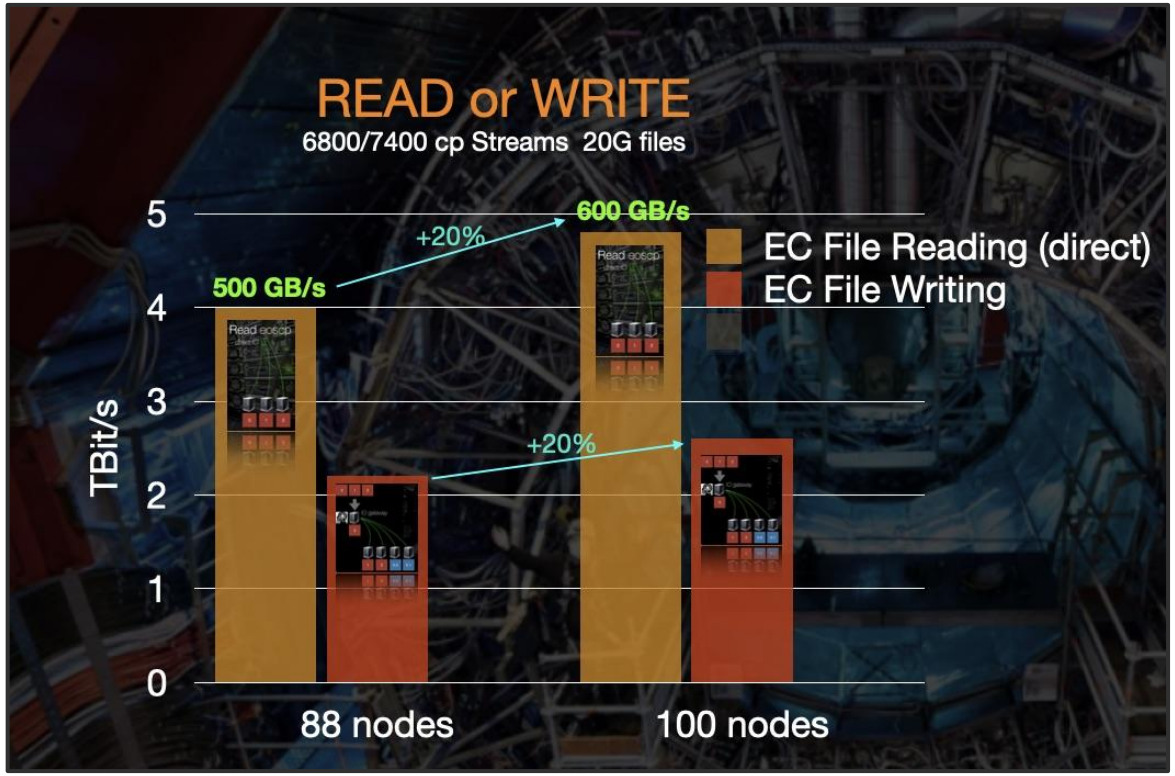
- **600GB/s read or 300GB/s write**

READ and WRITE performance was:

- **250GB/s read + 200GB/s write**

Adding 20% capacity allowed similar BW gains:

- **300GB/s read + 240GB/s write**
  - Write bandwidth matches incoming network speed 24x100Gb/s

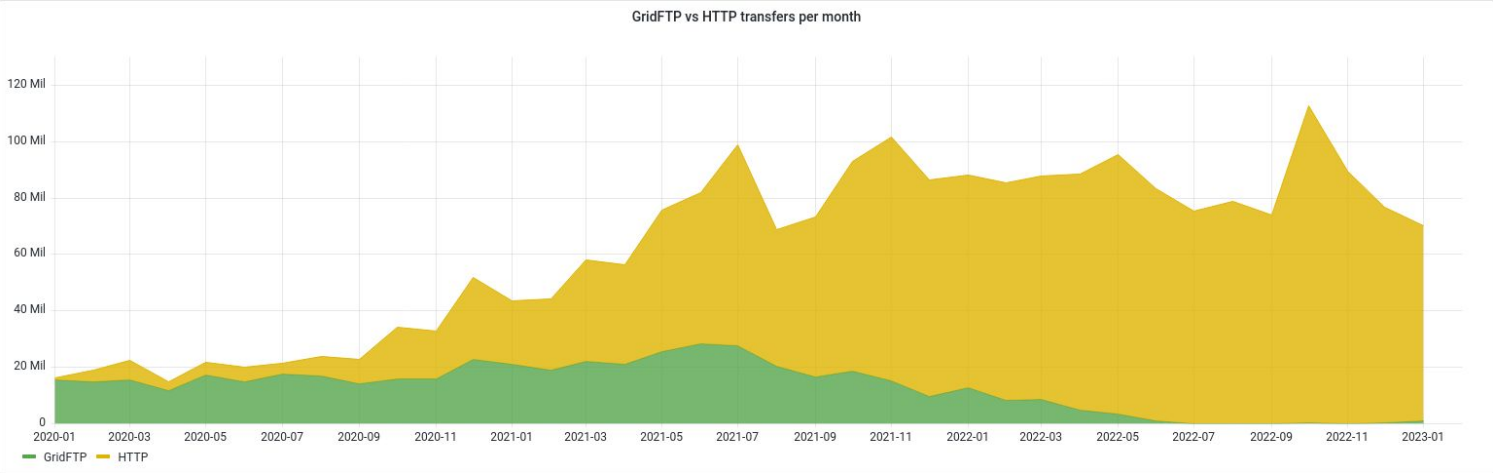


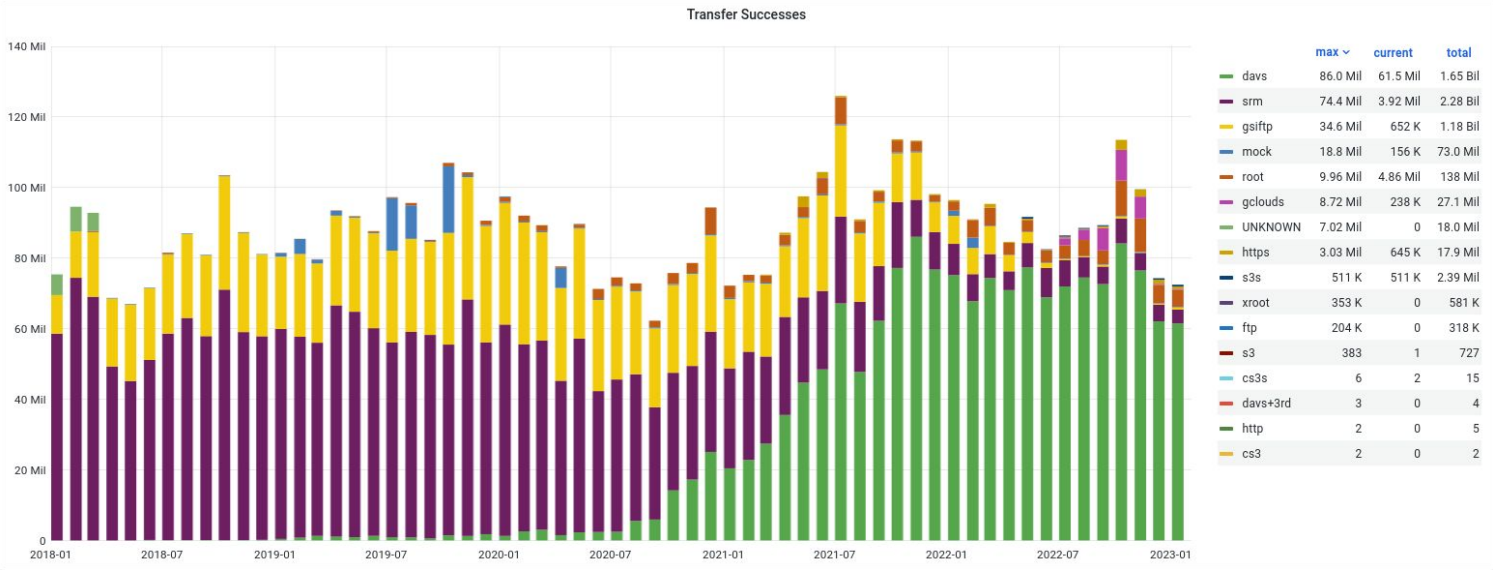
Physics workflows: protocol consolidation HTTP

HTTP is replacing gridftp and SRM traffic for grid transfers.

HTTP TAPE REST API deployed at CERN on all EOSCTA LHC instances

Starting project to support WLCG token based authentication for physics data transfer



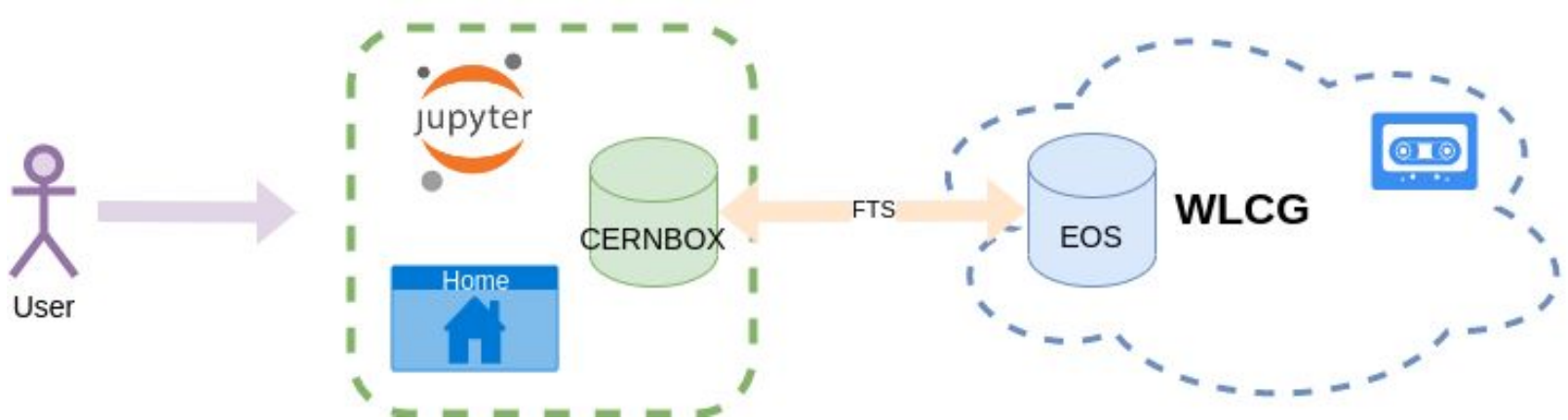


CERNBox update

- **CERNBox preparing for EOS5 migration: currently migrating from libmicrohttp (deprecated in EOS5) to xrdhttp**
  - Will allow CERNBox to participate in Rucio/FTS driven third-party copy data distribution

Bridging Physics community and user driven analysis workflows with HTTP

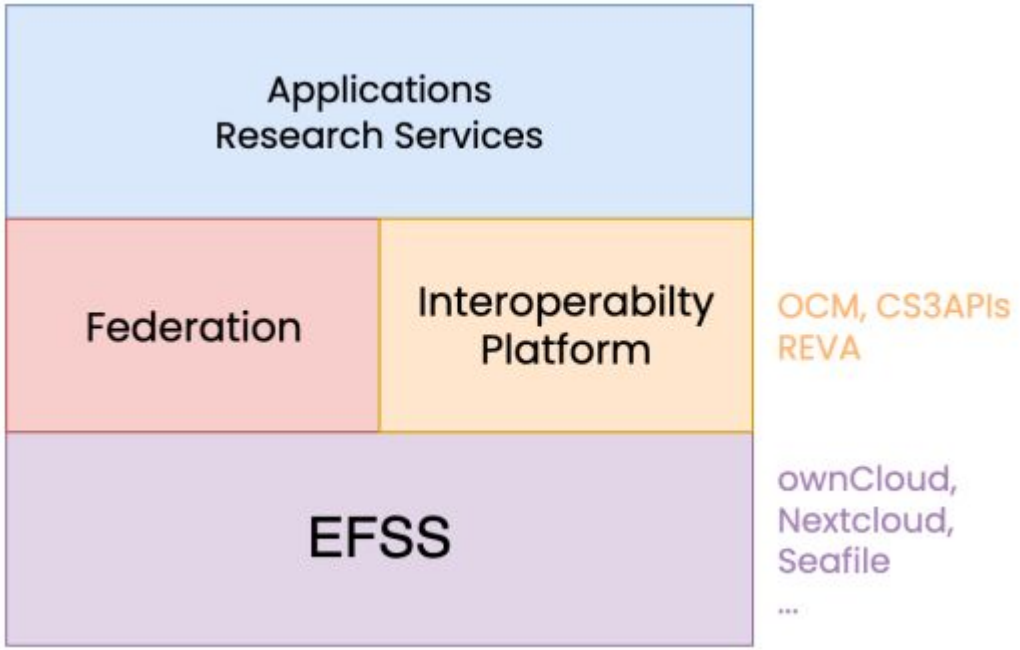
- **Physics user workflows transferring data between CERNBox and physics storage on WLCG**
  - Using current WLCG X509 authentication
  - HTTP file transfers with FTS Between physics EOS instances and user CERNBOX home directory
    - user data processing workflows, analysis notebooks, sync and share...



Interoperable federation of sync and share services

CERNBox: demonstrated the possibility to share storage resources and applications across organisations and institutions.

- Based on an open interoperable protocol called **OpenCloudMesh** (**HTTP Rest API**), which has been adopted by CERNBox and other popular cloud storage services.
- **ScienceMesh** is an emerging federation of EFSS services, presented at CS3 2023
- More information on CS3 – Cloud Services for Synchronisation and Sharing Conference Series, co-organised by IT Storage group



Protocol and authentication consolidation trends

User workflows:

- **WEBDAV data access for application research services**
- **OIDC token authentication**
- **OCM sharing**

Physics workflows:

- **HTTP TPC data transfers**
- **WLCG token authentication**

More bridges are built between the 2 communities, consolidating towards same protocols, authentication technologies…

Linux status in Storage infrastructure

CentOS 7 and CentOS Stream 8 EOL in 2024

Linux plans for CERN exposed in December 2022:

- **CERN IT supported distributions: Alma9/RHEL9**

Linux distribution status in IT storage

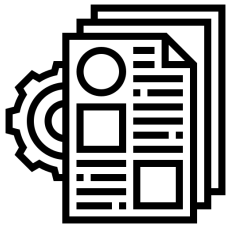
- **ALICE O2 diskservers previously on CentOS Stream 8, migrated to Alma 8**
- **Other servers running CentOS7/RHEL7**

Next steps:

- **Build all Storage software for alma9**
- **Alma 9 deployment in production during 2023 YETS**

Observability guild in storage group

Collaborate and share ideas



Document specific procedures or topics

Harmonize IT-SD monitoring projects technos

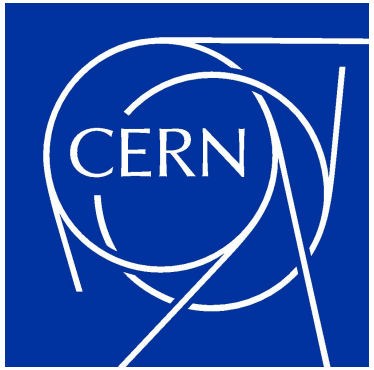
Synchronization with MONIT team

Conclusion

- **Consolidations at multiple levels will provide opportunities to optimize storage resources**

- **On the physics side, year 2022 confirmed in production nominal performance of all storage services**
  - expecting more records to be broken in 2023
  - consolidation of protocols in experiment workflows will continue in 2023

- **EOS software is key to all CERN Storage sections**
  - 2023 is EOS5 migration year for all CTA and CERNBOX EOS instances
    - migrating to a common *cerneosserver* puppet module
    - more details with developer insight at EOS 2023 Workshop 24-27 April 2023 at CERN



home.cern