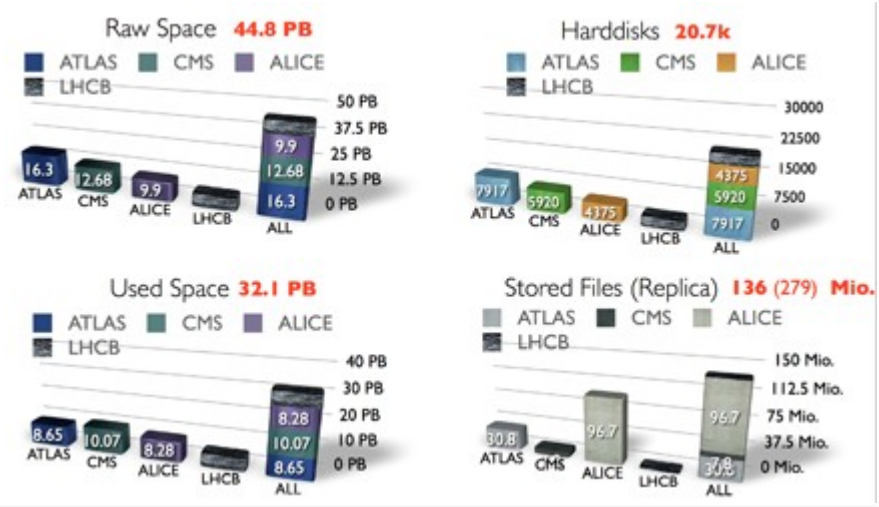
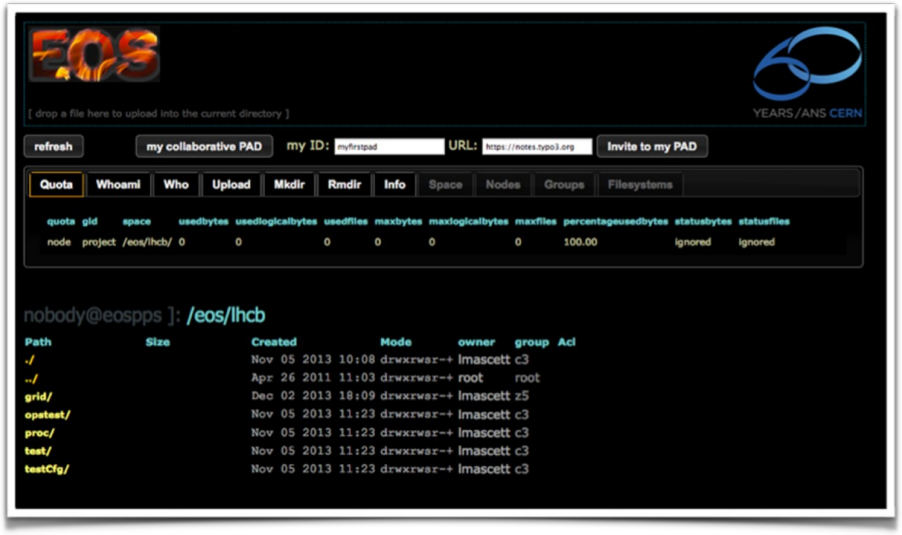
Areas of developments in storage at CERN

Massimo Lamanna

Source: A.J.Peters and FDO section

CERN IT Department
CH-1211 Genève 23
Switzerland





- EOS: Large disk farms for physics and beyond
  - Currently ~25 PB used quota
  - → 100 PB quota (@LHC Run2)
- Developed in CERN/IT (DSS)
- Original goal
  - Large scale (PBs for 100s/1000s independent scientists) analysis of LHC data
  - Arbitrary level of data durability via cross-node file replication or RAIN using commodity hardware
- Status
  - Open to non-physics use cases
  - NB: large number of protocols available!

EOS highlights

- No need for a central relational database
  - In memory hashtable
- Designed for infinite scalability and arbitrary reliability
  - Replica or Erasure Code (ReedSalomon, LDPC...)
- Disantagle physical and logical view
  - Disks keep file replicas, MGM manages them (no disk – service link)
  - Easy to manage realistic hardware (heterogeneity)
- Multi site federation support
  - Wigner / Geneva > 1000 Km distance

In-memory catalogue (MGM)

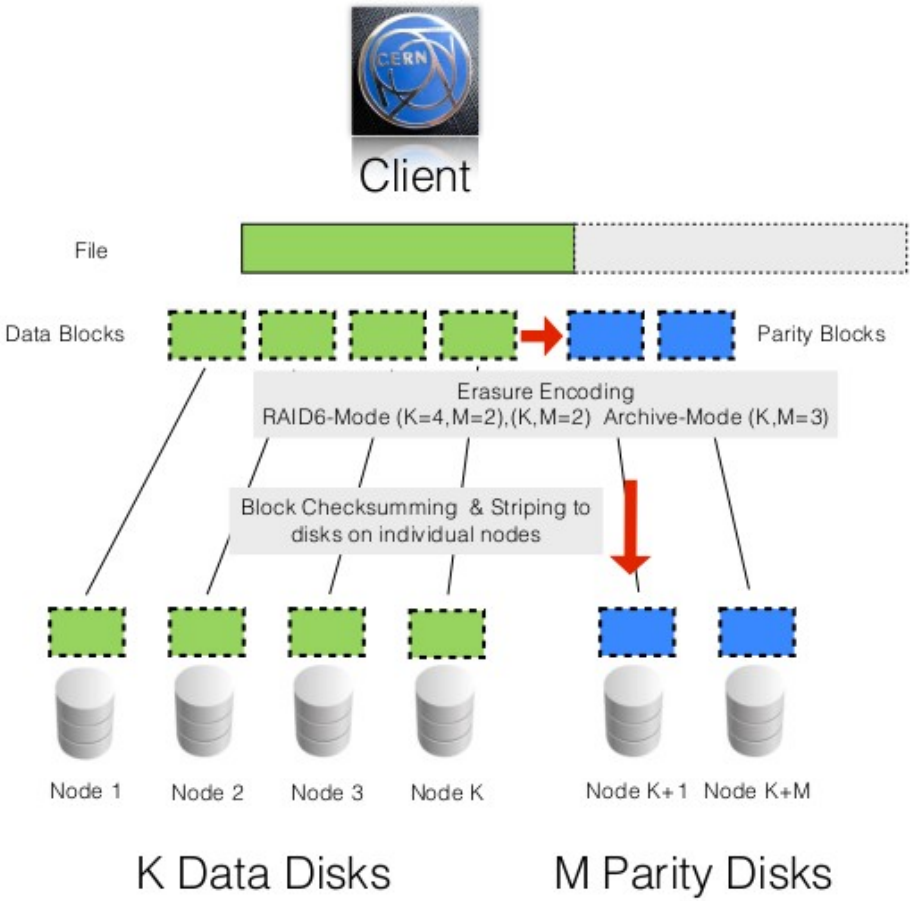
- Cornerstone of the architecture
- Initial concern
  - Run out of memory?
  - Durability
- Where are we?
  - Low-latency file access with in-memory namespace ~ 200 M files
  - Stable and durable (Master + Slaves)
  - Demonstrated by doing it
- <u>Investigation ares</u>
  - Stability above the 1B-file area
  - EOS Diamond
    - EOS Ceph inbreeding

DSS

RAIN & Erasure Encoding in EOS

Redundant Array of independent nodes

CERN IT Department



RAIN

- Block Striping allows parallel IO
  - boost single file performance

e.g. with K=4 280 MB/s streaming write & 400 MB/s streaming write

- Erasure Encoding + Block Checksumming allows
  - error correction on the fly
  - up to M concurrent disk failures without data loss

- disk space savings compared to file replication

e.g. 25% less space needed for (4+2) vs. 2 replica

Not deployed at large scale: several scenarios to be investigated:

- Interplay replica/fragments and self-healing
- New modes of operations (large n of fragments)
- ...

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

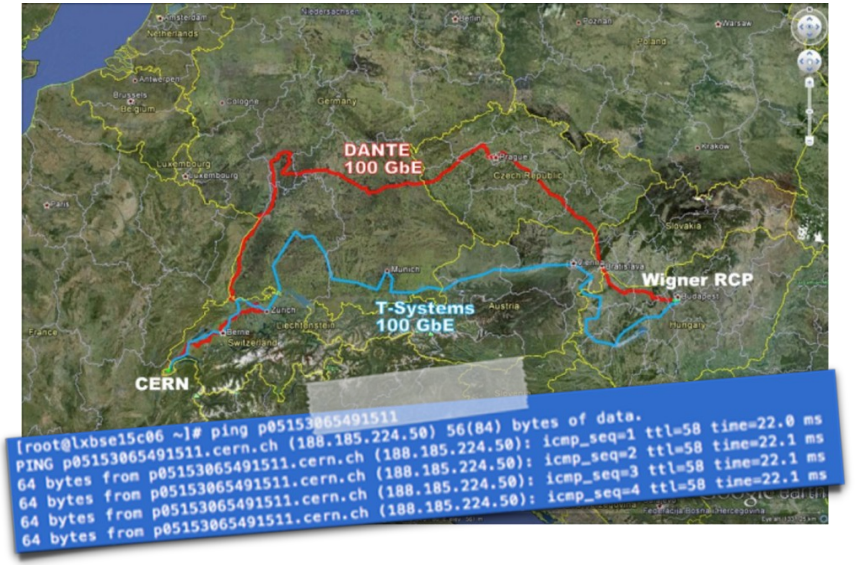
EOS in production

- Six multi-PB installations
  - few thousand disks per instance
  - Tape storage with CASTOR
  - New disks going mainly to EOS
  - **Operated by the same team**
- Simplified life-cycle management workflows for on-going replacement/repair of hardware
  - JBOD disks (no RAID controller) using software RAIN
  - Low-latency file access with in-memory namespace ~ 200 M files
  - Fine-grained access control and quota management
- GRID and local storage element
  - Accessed from thousands of CERN-local and remote batch nodes with Kerberos and GSI authentication
  - Full support for XRootD, GridFTP, HTTP(S) & WebDav protocol - and – mountable with FUSE as a remote file system
    - notably users love fuse...

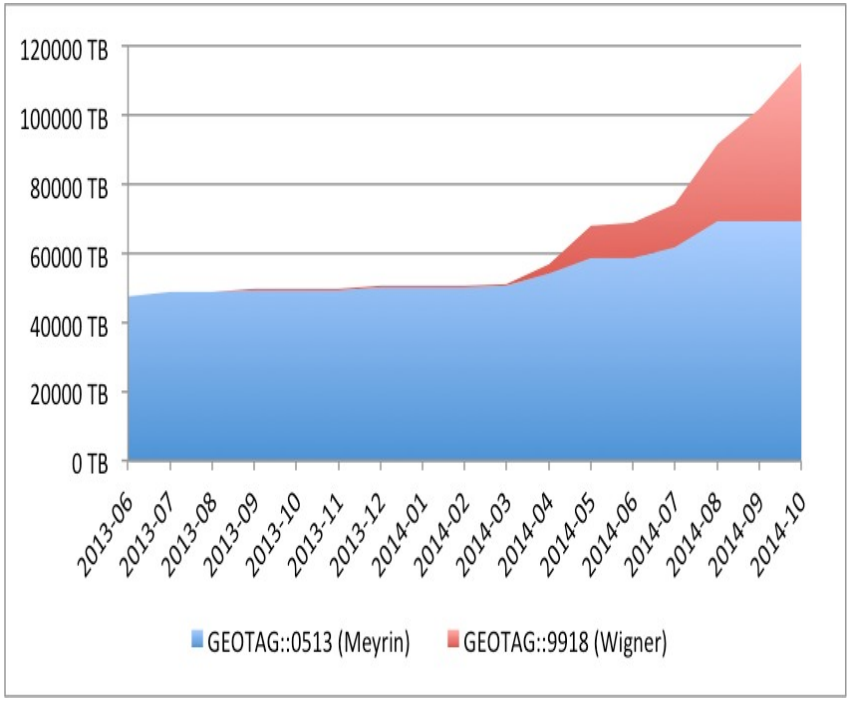
EOS in production (some challenges)

- What happens above ~ 1B files?
  - OK if static
  - What if the rate of changes goes up?
- CERNBOX
  - Marry EOS (EOSUSER) with your laptop
    - Flexibility of the sync client
    - Power of a file system (e.g. LXBATCH jobs using the EOS asa file system
    - Both views coincide!

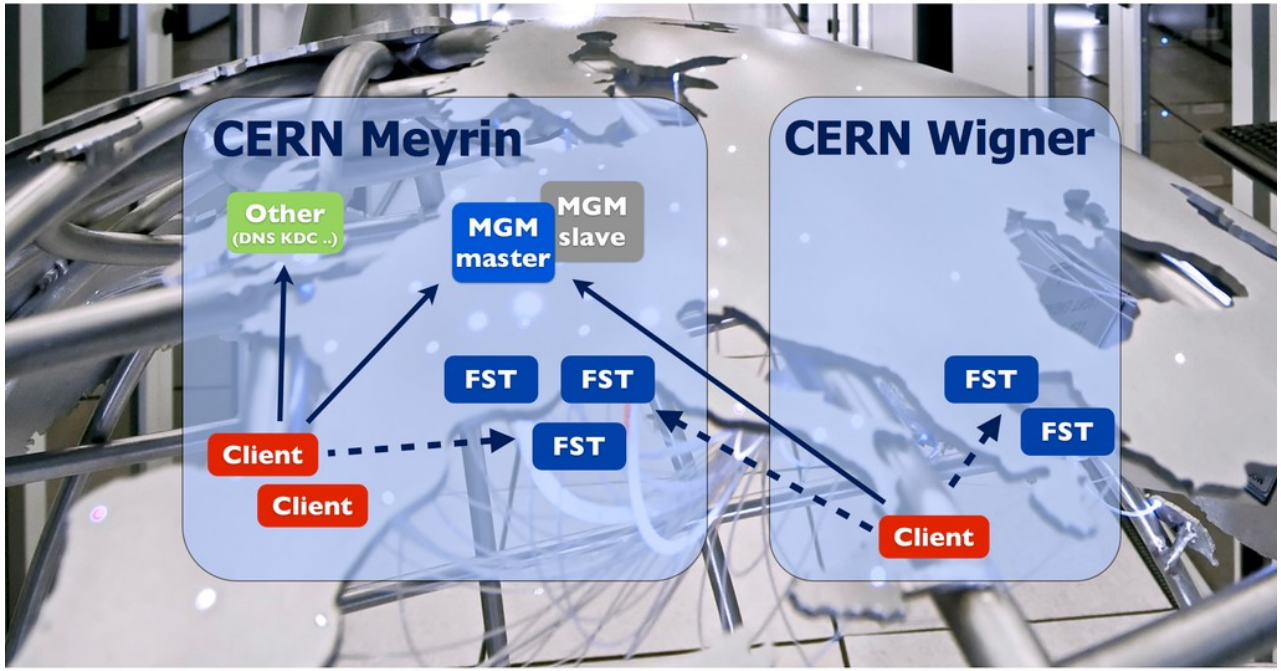
EOS installed across the CERN computer centres



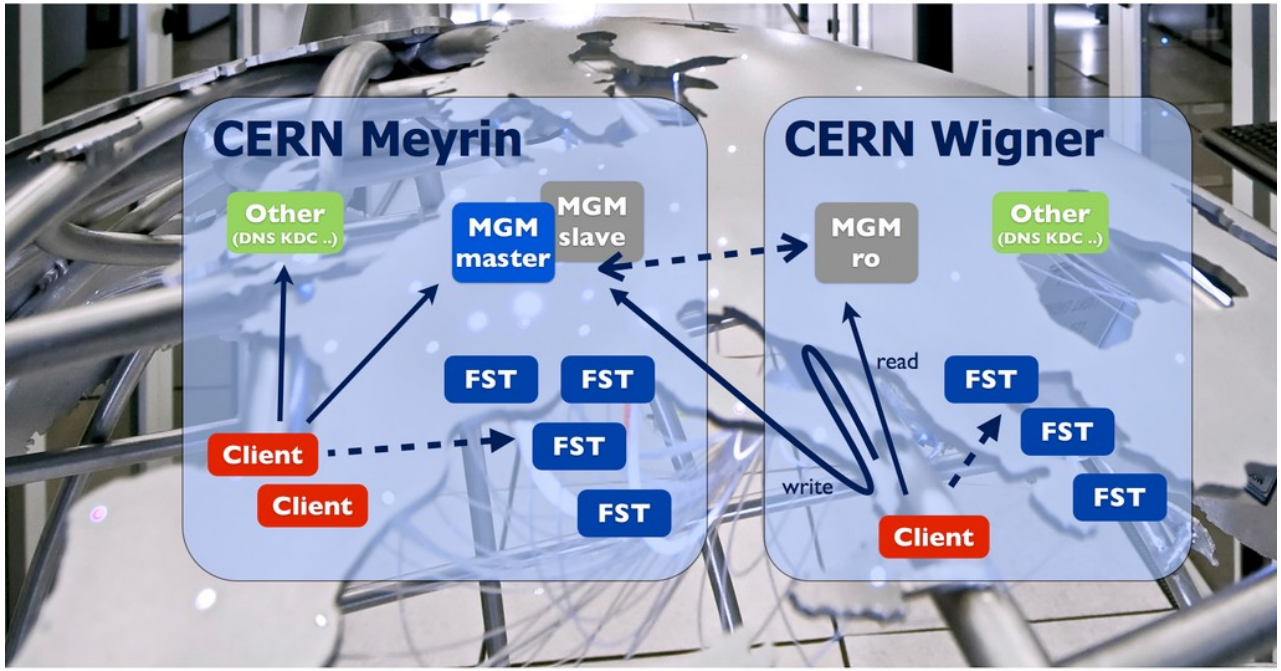
- EOS takes advantages of the two CERN computer centres
  - Coping with ~20-ms latency
  - Distributing copies across the two sites for dependability and performance
- Status
  - As today we are crossing the 40% mark



EOS 2014 Deployment



EOS 2015 Deployment



Challenges in distributed installations

- 2 sites across 1000 km: OK
- Can we do more?
- Locality in a several-site installation
- EOS hit-the-road-jack kit :)
- ...

Workshop on Cloud Services for File Synchronisation and Sharing

All material is online (slides and presentations recording)

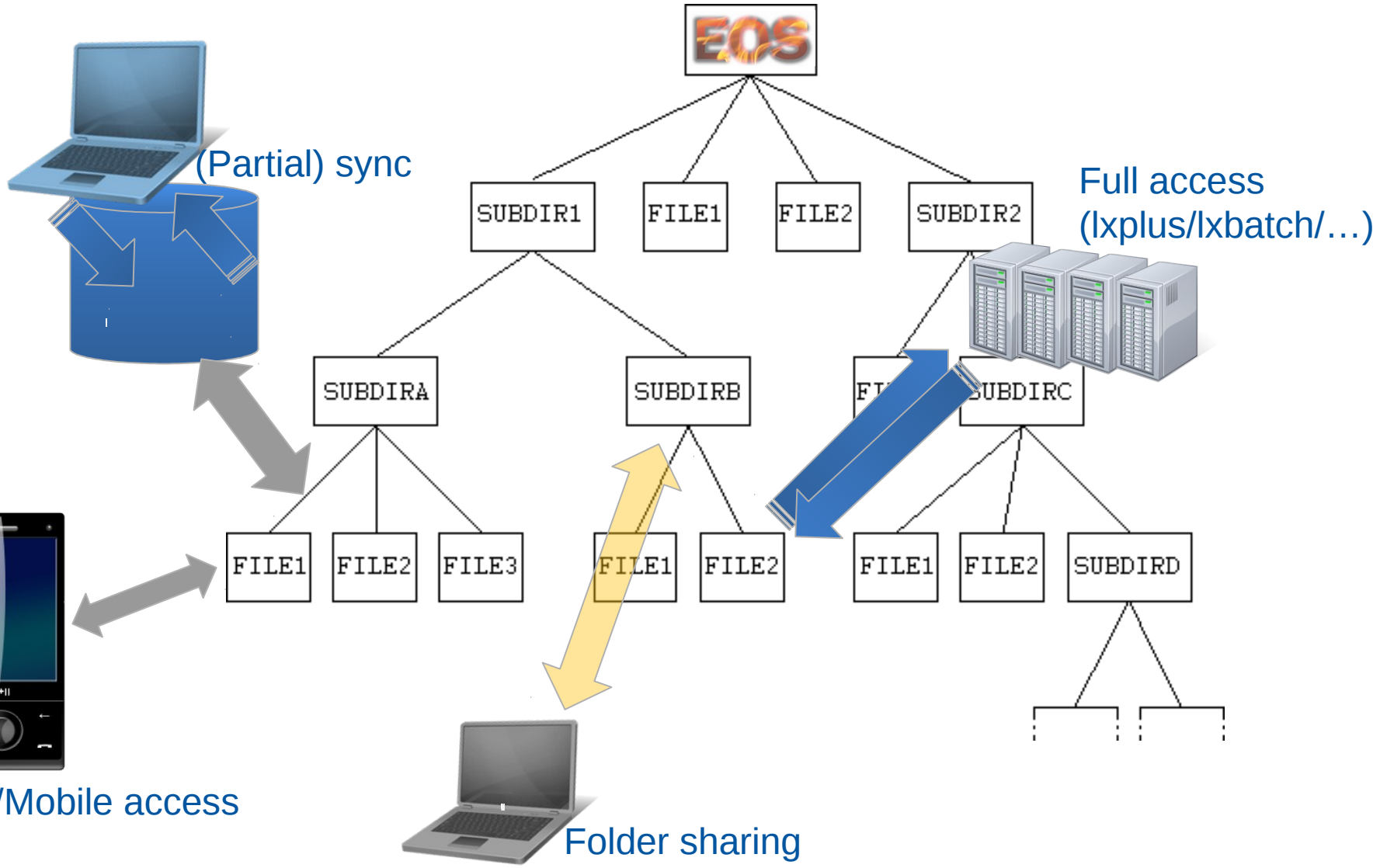
>80 participants

Broad interest (large participation outside our “traditional” community)

Several companies

CERNBox presented

CERNBox


EOS
SUBDIR1
FILE1
FILE2
SUBDIR2
SUBDIRA
SUBDIRB
SUBDIRC
FILE1
FILE2
FILE3
FILE1
FILE2
FILE1
FILE2
SUBDIRD
(Partial) sync
Full access
(lxplus/lxbatch/…)
Web/Mobile access
Folder sharing

Conclusion

- EOS is our flagship project
- Essential for the LHC Run 2
  - Revised utilisation of all DSS projects
- EOS evolution
  - Number of challenging questions
  - Important for LHC and non-LHC usage