dCache Update

WLCG Grid Deployment Board

Tigran Mkrtchyan for the dCache collaboration

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Scientific Data Challenges

Ingest

- High data ingest rate
- Multiple parallel streams
- High durability
- Effective handling of large number of files

Analysis

- High CPU efficiency
- Chaotic access
- Standard access protocols
- Access control
- Local user management

Sharing & Exchange

- 3rd party copy
- Effective WAN Access
- In-flight data protection
- Identity federation
- Access control

Long Term Preservation

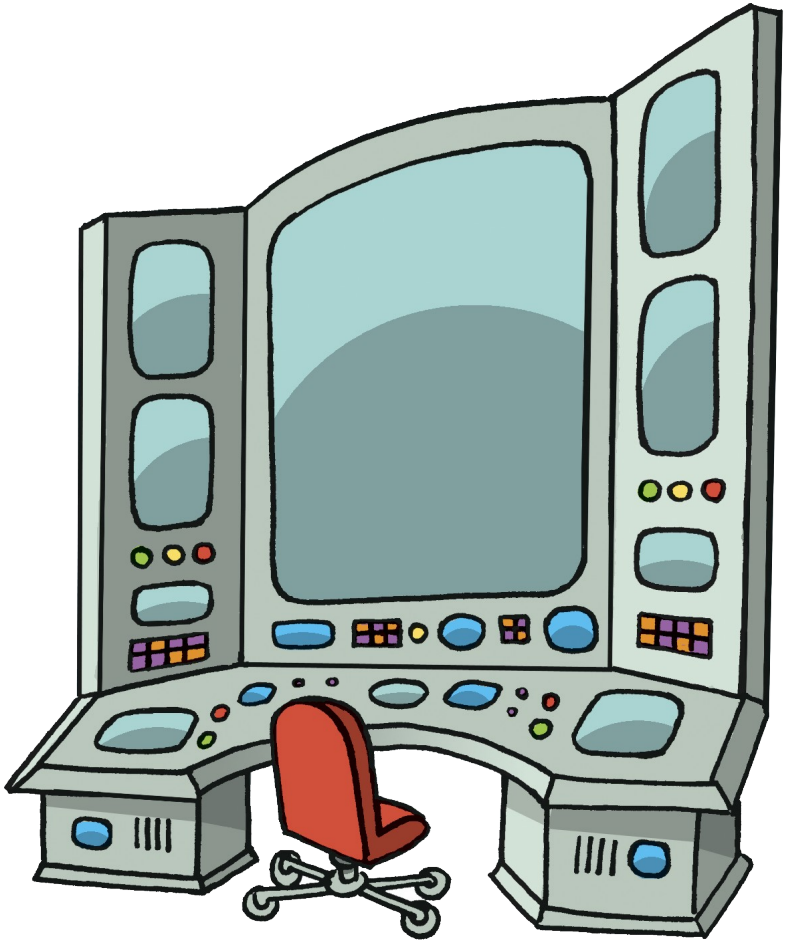
- High Reliability
- Self-healing
- Automatic technology migration
- Persistent identifier

Technical Directions

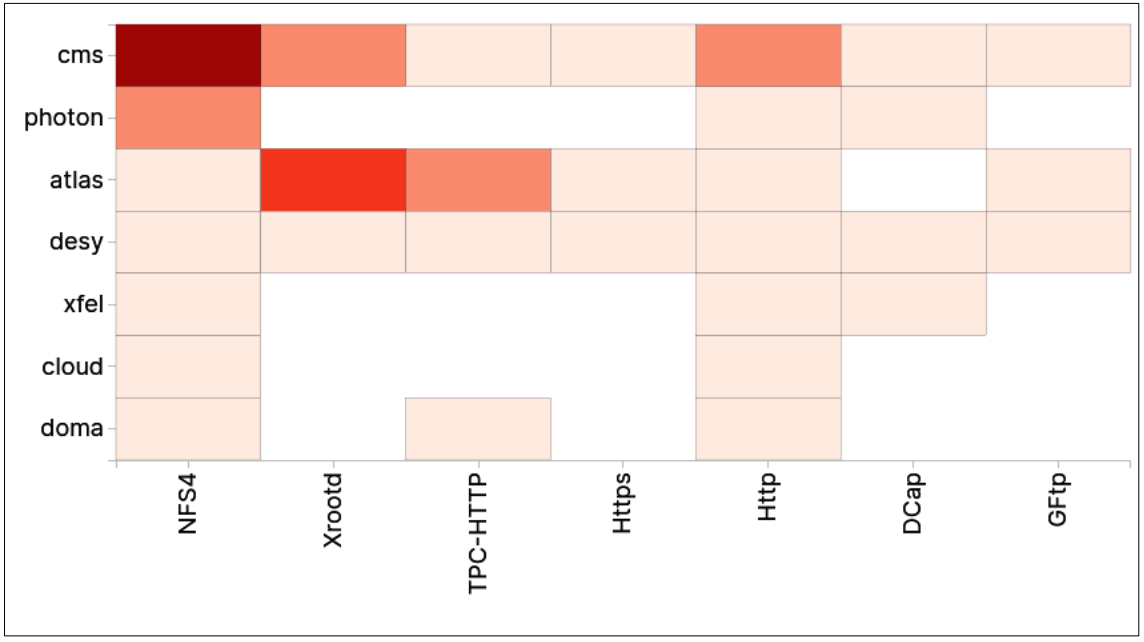
- Scaleout
  - Namespace
  - Number of pools (cells)
- Token-based Authentication
- Better *Analysis Facility* support
  - POSIX access and compliance
  - HPC workload support
    (DDoS protection)
- QoS
- Tape integration
- Green IT

Some (DESY) Numbers

- XFEL
  - Total capacity ~120 PB
  - ~400 physical hosts (~4000 dCache pools)
  - 20-40 GB/s injest
- Photon
  - DB size – 2.5TB
  - ACL table 600GB
  - Directories with 3 10^6 files
  - 1.2 10^9 file system objects
  - 100K files in the flush queue
  - Two tape copies, different media type
- ATLAS
  - dir/file → 1/3
- NextCloud
  - File lifetime < 1s

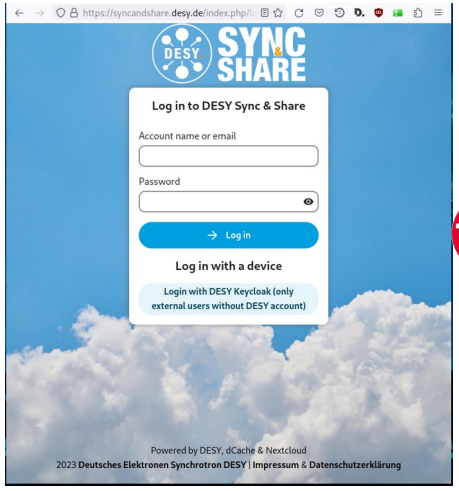


Data Access Variety (at DESY)

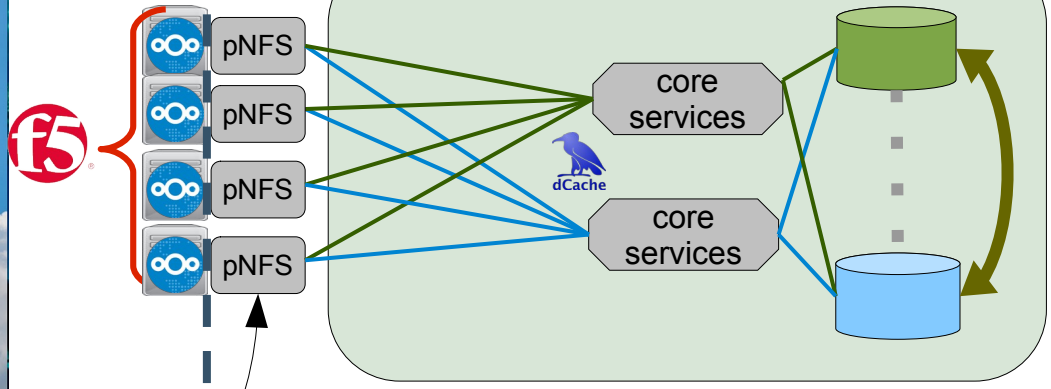


- ROOT-IO
- Non-HEP tool chain
  - Active use of Jupyter Notebooks
  - Non-ROOT data formats
- Industry standard AuthN
  - Tokens based authentication
  - Federated IdP
- Use of private clouds
  - Data access from a container
- Use of HPC resources

NextCloud Instance @ DESY



NextCloud-25

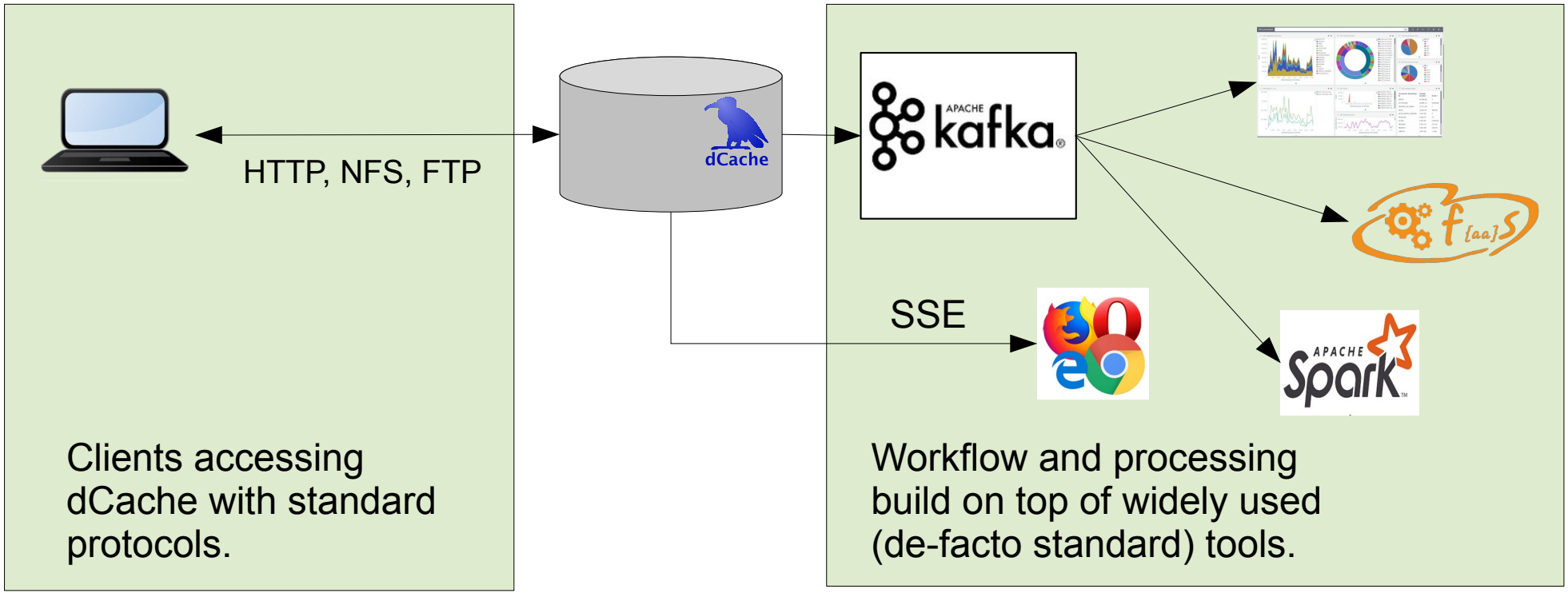


dCache-7.2

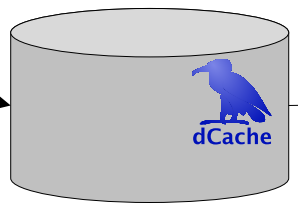
dCache as a Storage Backend

- PB-scale storage system
- No changes in Nextcloud required
- Unique functionality
  - Tape integration
  - File ownership preservation
  - NFS export to selected users
  - Storage events
  - Data visible by all protocols and security flavors

Standards Everywhere…



Automatic Metadata Population



Event: new data

extract metadata

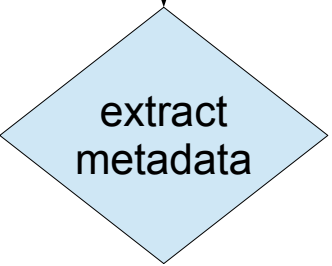
Automatic catalog update

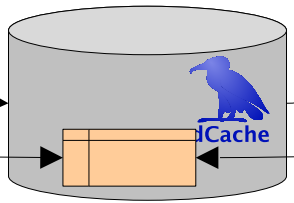
metadata catalog

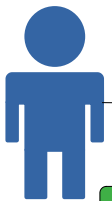
#Anna Bertha Ludwig

#hand #1895 #x-ray

Metadata Population







Manual Labeling

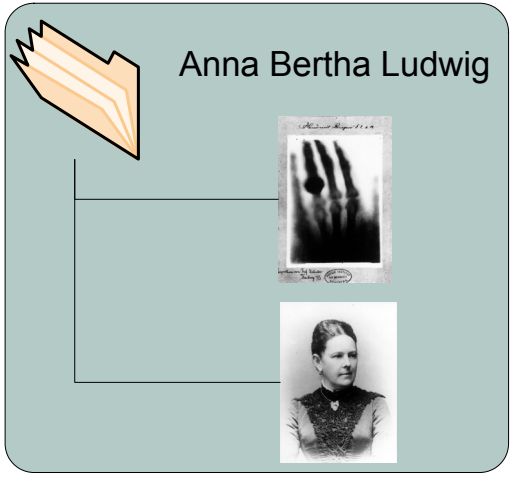
Automatic labeling

#Anna Bertha Ludwig

#hand #1895 #x-ray

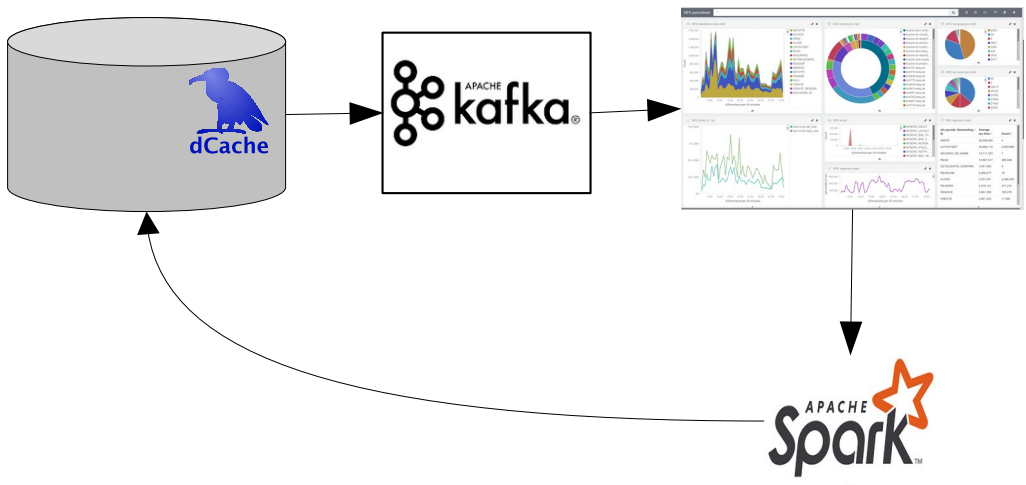
User Metadata/Labeling in dCache

- Extended attributes
  - Exposed via NFS, WebDAV, REST
- Label-based virtual **read-only** directories
  - List all files with a given label
- *dCache rules applies*
  - *Visible through all protocols*



Self-Adaptive dCache

- Joint project with Hamburg University on Applied Science
  - MAPE-Loop
    - Automation of large deployments
    - Hotspot detection and re-balance
    - Self-healing load optimization



SRM Still Here ... *but not too long*

- Two main gaps to fill
  - Space allocation
  - Tape operation
- Two alternatives to replace
  - User and Group based Quota system
  - WLCG tape recall API

Tape REST-API v1 (like SRM, but different)

STAGE

- Request to stage many files at once

CANCEL

- Cancel bulk request

DELETE

- Cancel bulk request + clear history/status

EVICT

- unpin cached copy

PIN

- Pin cached copies with a lifetime

FILEINFO

- Request status many files at once (locality, checksum)

StoRM

Tape rest API

dCache bulk API

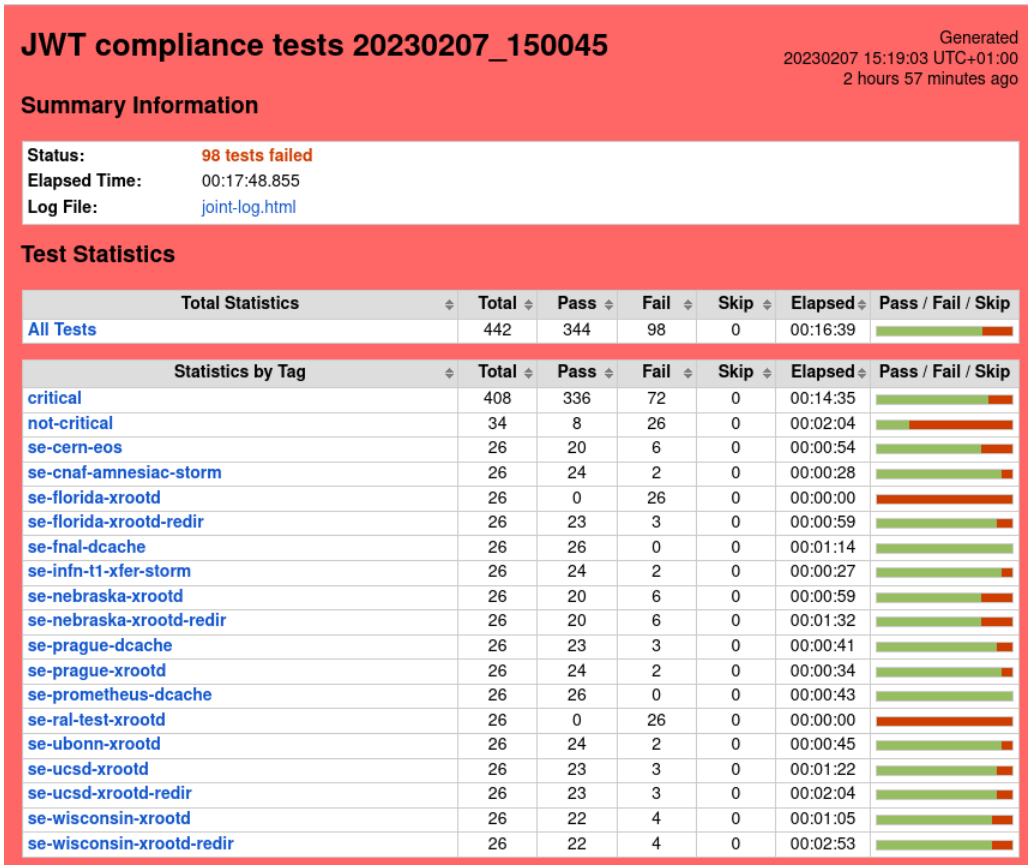
WLCG Tape API

User/Group Quotas

- **Quota ≠ Space reservation**
- Lazy, based on periodic scans
  - Users might overrun
  - Removed space not reclaimed immediately
- Global per file system
  - No quota per directories
- Respects Files Retention policy
  - Separate for 'disk' and 'tape' files
- Available since 7.2, enabled by default since 8.2

Token-based AuthN

- Core functionality is there
- Sites have started deployments
  - New use-cases popping up!
- Documentation update in progress



JWT compliance tests 20230207_150045

Generated
20230207 15:19:03 UTC+01:00
2 hours 57 minutes ago

Summary Information

Status:	98 tests failed
Elapsed Time:	00:17:48.855
Log File:	joint-log.html

Test Statistics

Total Statistics	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
All Tests	442	344	98	0	00:16:39	

Statistics by Tag	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
critical	408	336	72	0	00:14:35	
not-critical	34	8	26	0	00:02:04	
se-cern-eos	26	20	6	0	00:00:54	
se-cnaf-amnesiac-storm	26	24	2	0	00:00:28	
se-florida-xrootd	26	0	26	0	00:00:00	
se-florida-xrootd-redir	26	23	3	0	00:00:59	
se-fnal-dcache	26	26	0	0	00:01:14	
se-infn-t1-xfer-storm	26	24	2	0	00:00:27	
se-nebraska-xrootd	26	20	6	0	00:00:59	
se-nebraska-xrootd-redir	26	20	6	0	00:01:32	
se-prague-dcache	26	23	3	0	00:00:41	
se-prague-xrootd	26	24	2	0	00:00:34	
se-prometheus-dcache	26	26	0	0	00:00:43	
se-ral-test-xrootd	26	0	26	0	00:00:00	
se-ubonn-xrootd	26	24	2	0	00:00:45	
se-ucsd-xrootd	26	23	3	0	00:01:22	
se-ucsd-xrootd-redir	26	23	3	0	00:02:04	
se-wisconsin-xrootd	26	22	4	0	00:01:05	
se-wisconsin-xrootd-redir	26	22	4	0	00:02:53	

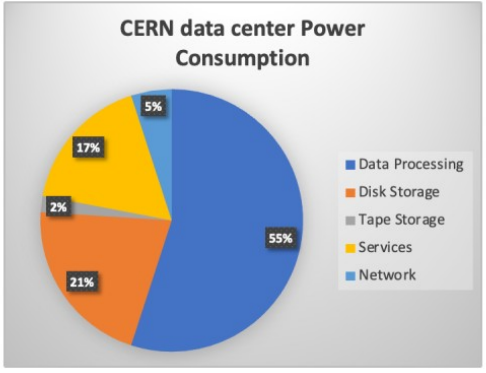
New Tape Hype?

WLCG data centers power consumption

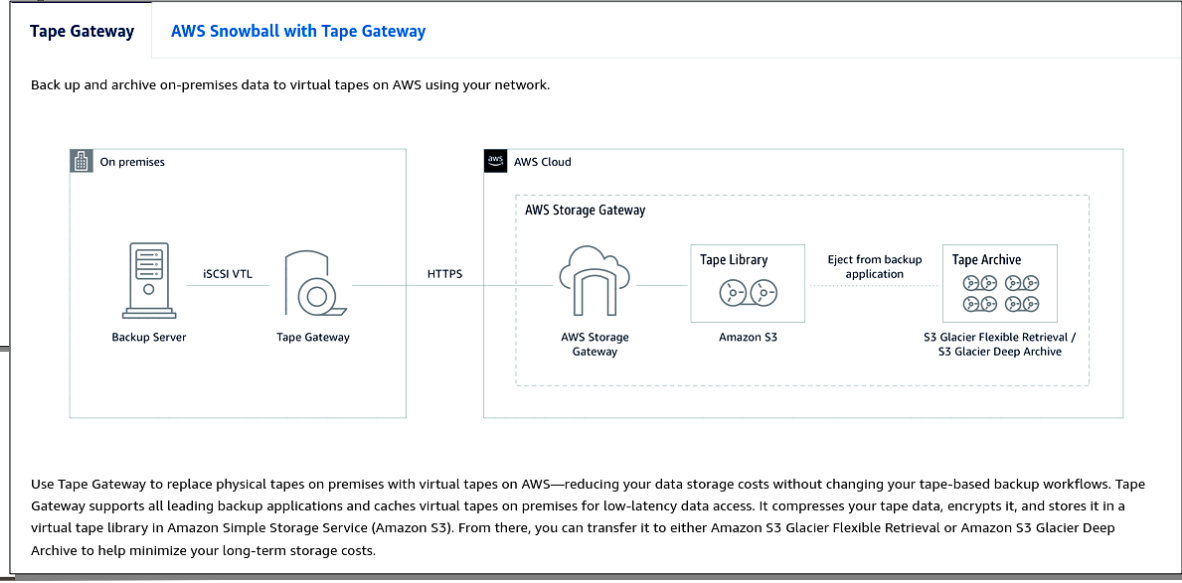
The pie chart shows the breakdown of the power consumption at the CERN data center

Most of the power is consumed for data processing (CPUs). Large part of the “services” are in fact CPUs

In this study we will focus on the energy needs for CPUs



Shameless stolen from Simone Campana



dCache ⟺ CTA Integration

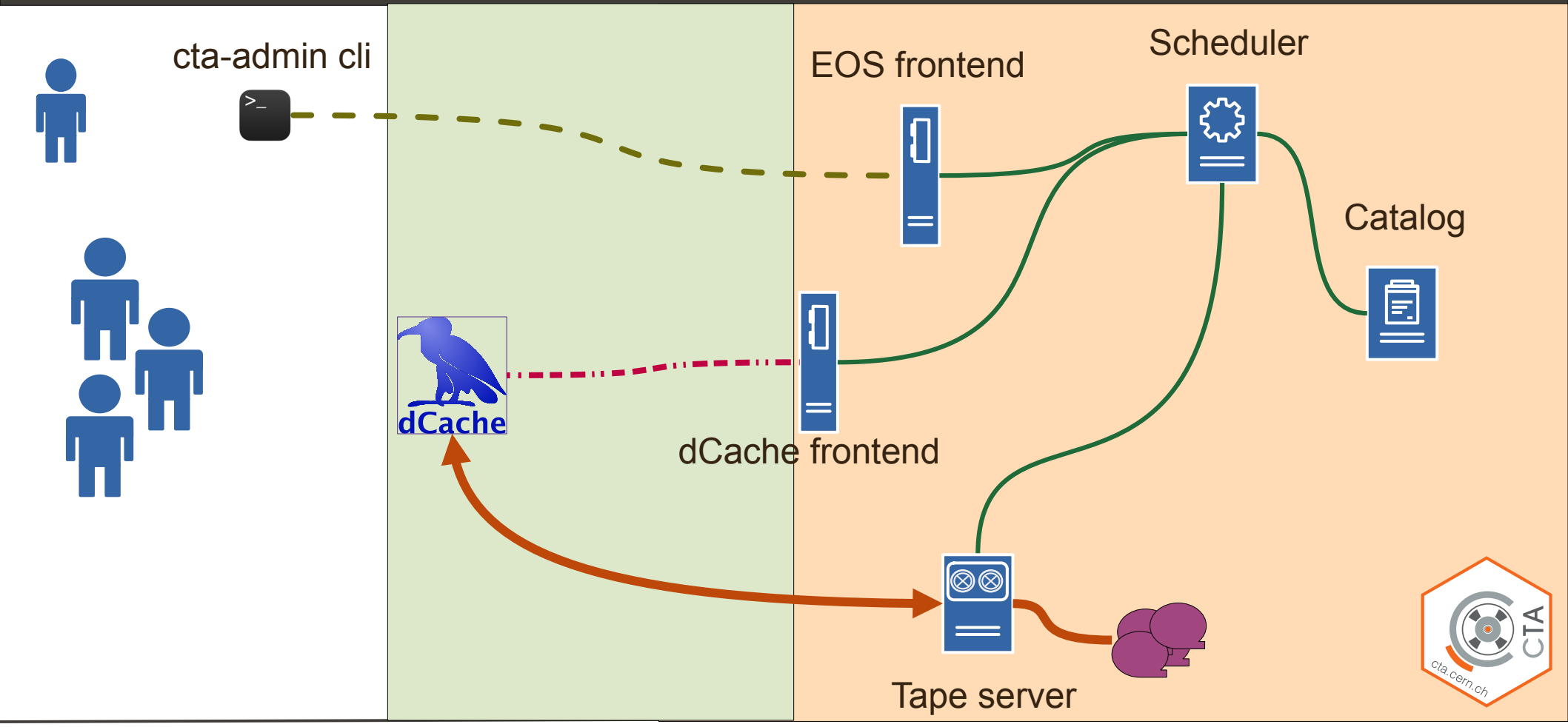
dCache (≥ 7.2)

- Nearline driver to add
- Can run in parallel with other HSMs
- Pre-scheduling on pools should be disabled/reduced
- File path, uid, gid not preserved

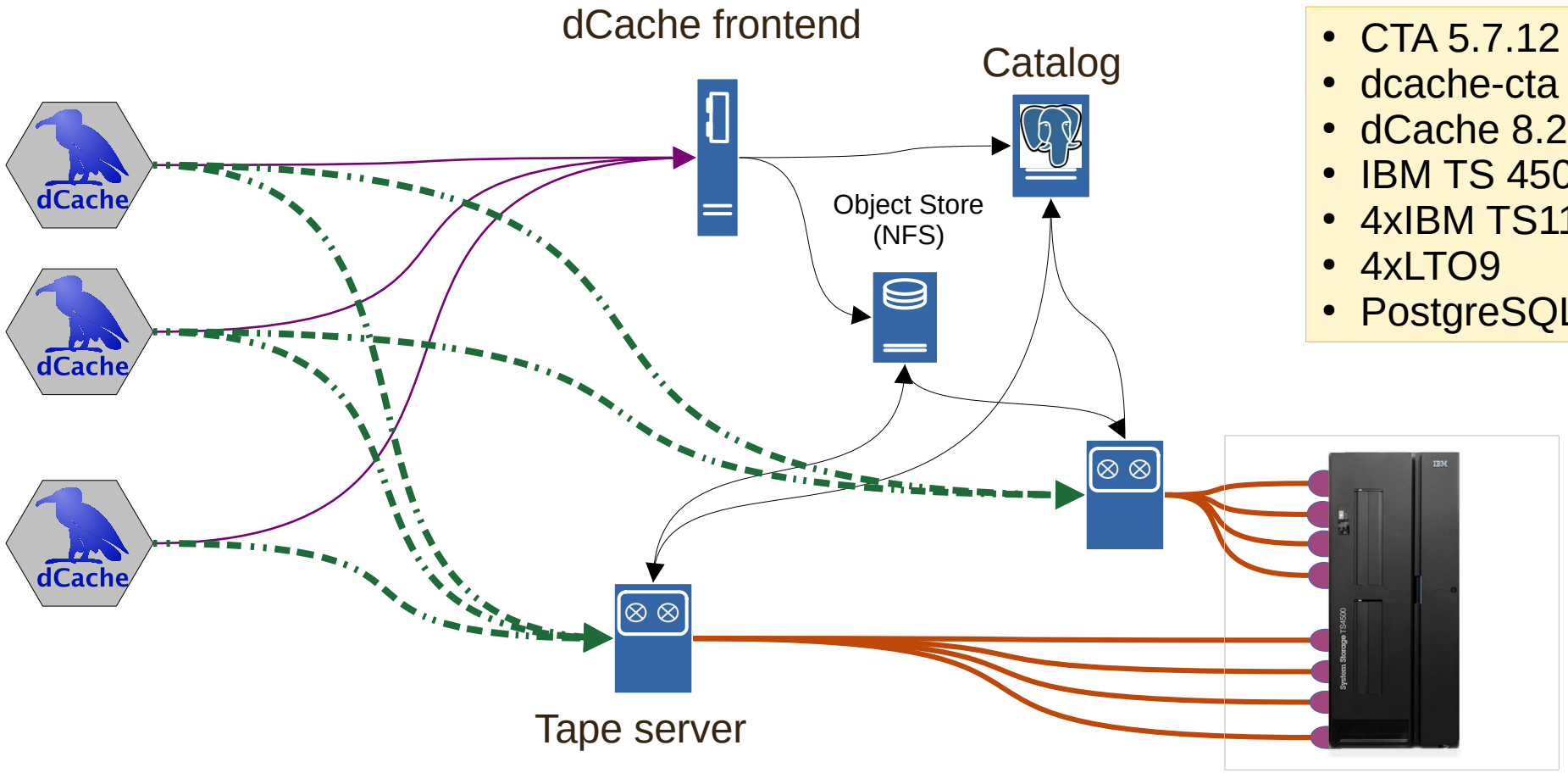
CTA (≥ {5,4}.7.12)

- Additional *cta-frontend-grpc* frontend service (packaged as own rpm)
- Limited to dCache required minimal functionality
  - Not dCache specific
  - *cta-frontend still needed for admin commands*

Integration with CTA



Deployment at DESY



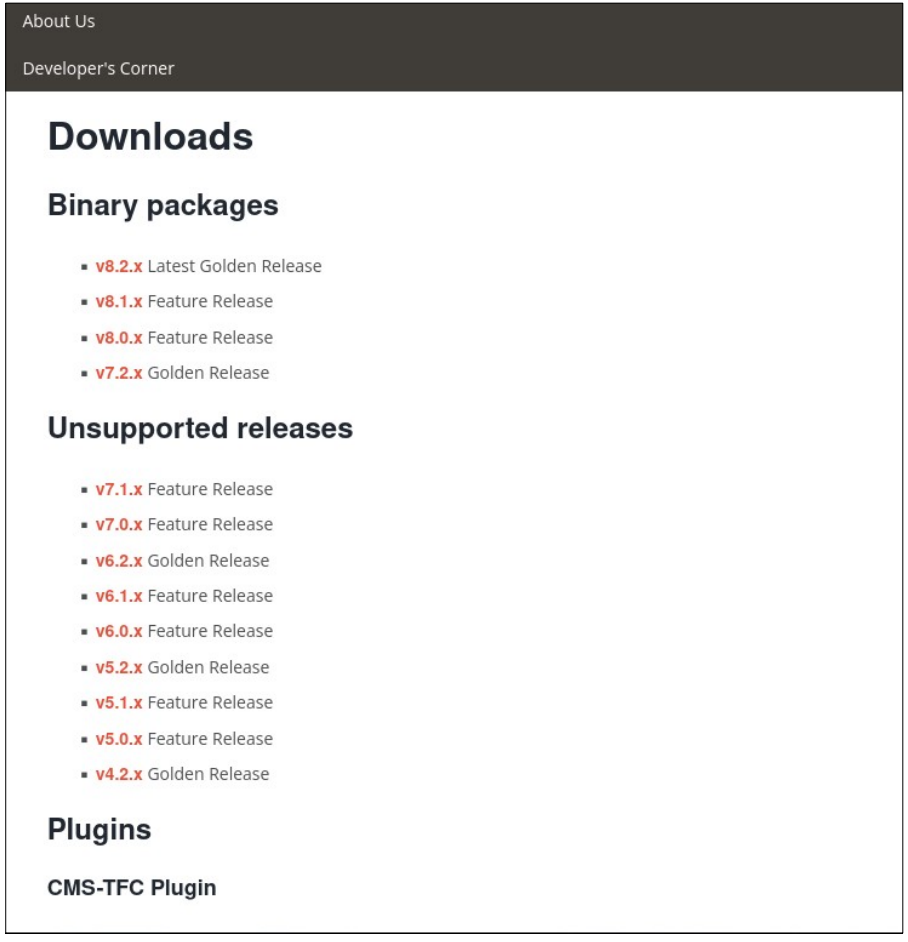
- CTA 5.7.12
- dcache-cta 0.8.0
- dCache 8.2.7
- IBM TS 4500
- 4xIBM TS1160
- 4xLTO9
- PostgreSQL 14.1

dCache+CTA Status

- Seamless integration with dCache is merged into upstream CTA code at CERN
  - The latest official CERN releases 5.7.14 is deployed at DESY.
  - The proposed dCache interface is under adoption by EOS.
- The existing OSM tape format is supported for READ
  - The code changes are adopted by Fermilab data management team for ENSTORE tape format.
  - The OSM tape catalog conversion procedure is ready and exercised multiple times. Final migration expected by Q1 2023 (a.k.a now).
- Our deployment replicate to by other HEP sites
  - PIC Barcelona have successfully replicated our setup (currently dCache + ENSTORE).
  - Fermilab is planning in Q2 2023 (currently dCache + ENSTORE).
  - RAL in UK plans to migrate to PostgreSQL from ORACLE based on our experience

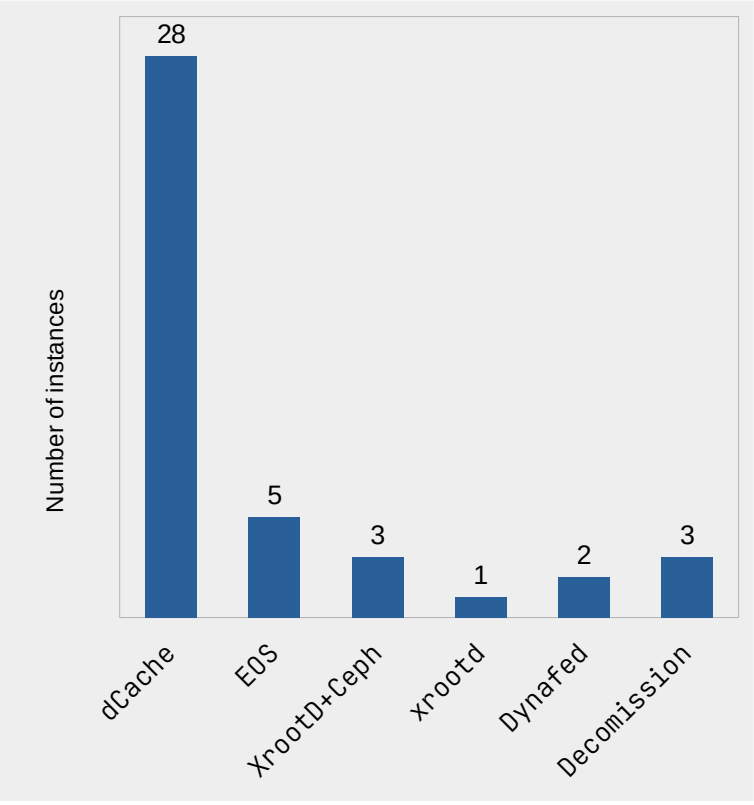
Supported OS platforms

- 6.2 - 8.2
  - RHEL 7, 8, 9
  - JVM 11
- 9.0 (Feb. 2023)
  - RHEL 7, 8, 9
  - JVM 11, 17
- 10.0 (~ 1Q 2024)
  - RHEL 8, 9
  - JVM 17



DPM Migration

- Spike of new users
- Series of tutorials
- Help from EGI
  - Thanks to Petr Vokac



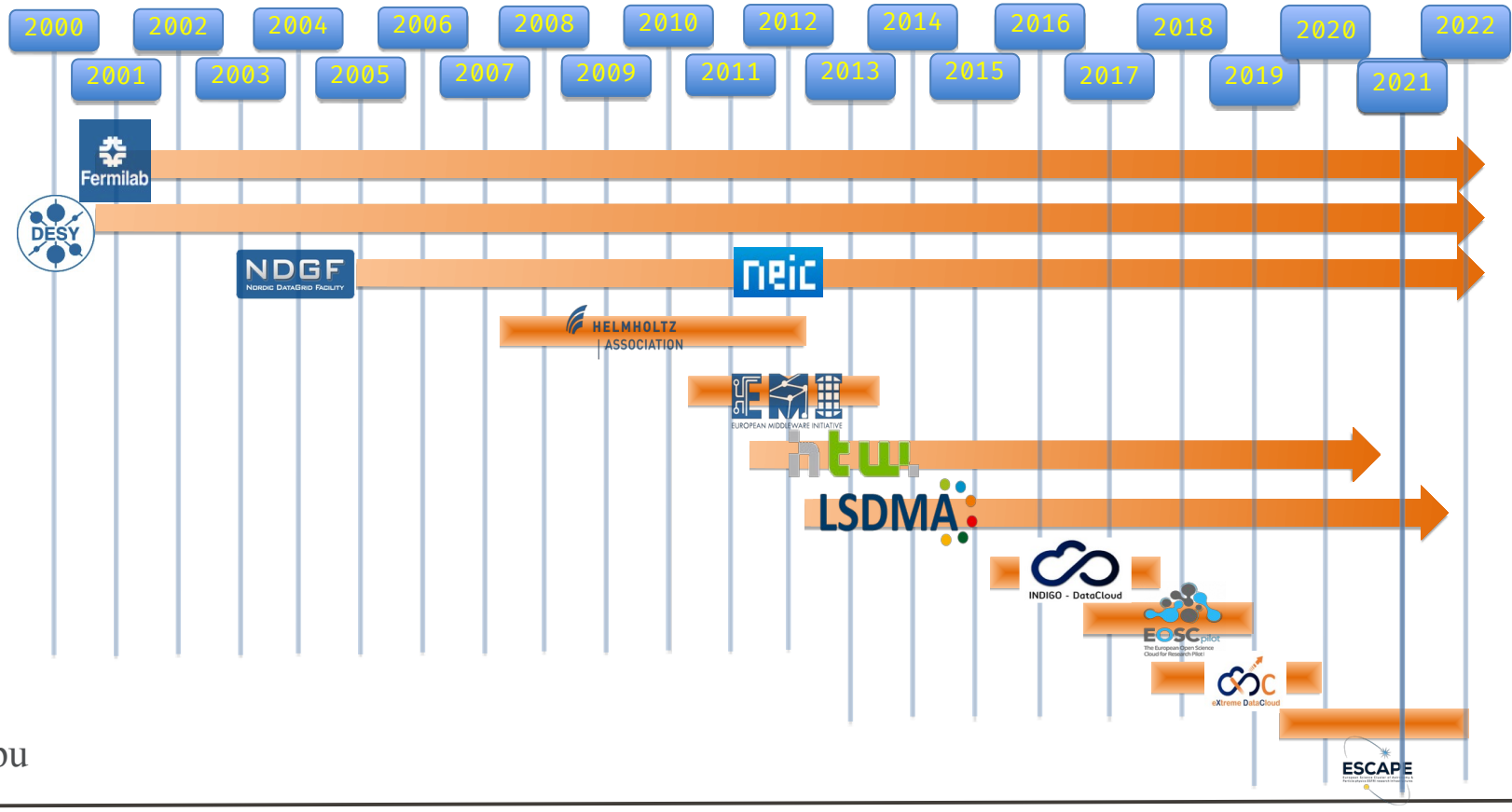
https://docs.google.com/spreadsheets/d/1KDVAJ9JzlycA3Wrz1iY2fQxZndWdAezFnLaDAxXIpUs/edit

Prominent Changes

- BULK Service
- TPC improvements
- NFSv4.1/pNFS improvements
- XROOT evolution (TLS, tokens, TPC)
- HSM connectivity

Project Funding & Team

- **DESY**
  - Svenja Meyer
  - *Paul Millar**
  - Tigran Mkrtchyan
  - Lea Morschel
  - Marina Sahakyan
- **FermiLab**
  - Dmitry Litvintsev
  - Albert Rossi
- **NeIC**
  - Krishnaveni Chitrapu



Non Functional Changes

- Documented release/test process
- Shareable build pipelines
  - Can be replicated at sites
- Transparent release process
- K8S based deploument
- Code will stay on Github

Thank You!

More info:

https://dcache.org

To steal and contribute:

https://github.com/dCache/dcache

Help and support:

support@dcache.org, user-forum@dcache.org

Developers:

dev@dcache.org

© HTW Berlin/Alexander Rentsch

17th International dCache Workshop May 31 – June 1 HTW-Berlin

https://indico.desy.de/e/dcache-ws17