dCache: storage system for data intensive science

26th CHEP

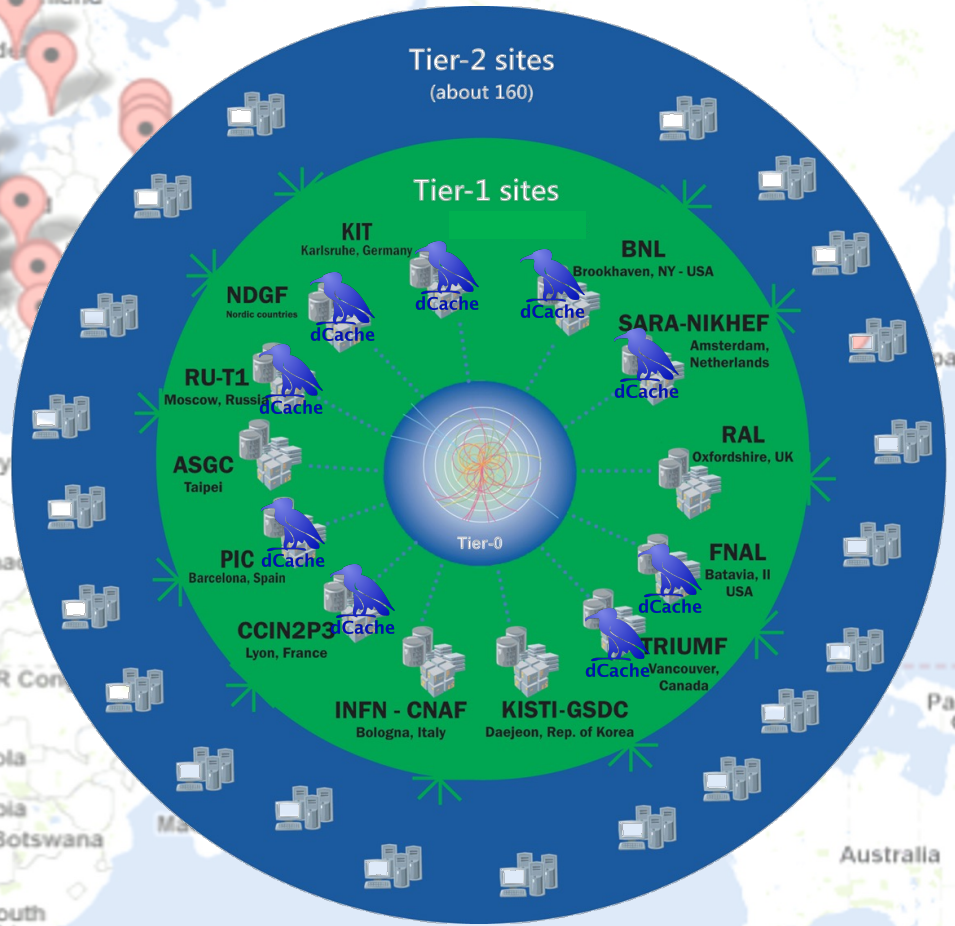
Dmitry Litvintsev for the dCache collaboration

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

The **Egyptian vulture** (***Neophron percnopterus***), also called the **white scavenger vulture** or **pharaoh's chicken**
https://en.wikipedia.org/wiki/Egyptian_vulture

Data intensive science

- Energy Frontier
  - LHC, Belle II
- Intensity Frontier
  - Neutrino physics
    - Dune, Icarus, Nova, Microboone…
  - Rare phenomena, BSM, precision SM tests
    - Mu2e, g-2 …
  - Photon Science
    - EuXFEL
- Cosmic Frontier
  - Dark Energy / Dark matter search
  - Astro-physics
  - Astro-particle physics
- Life sciences

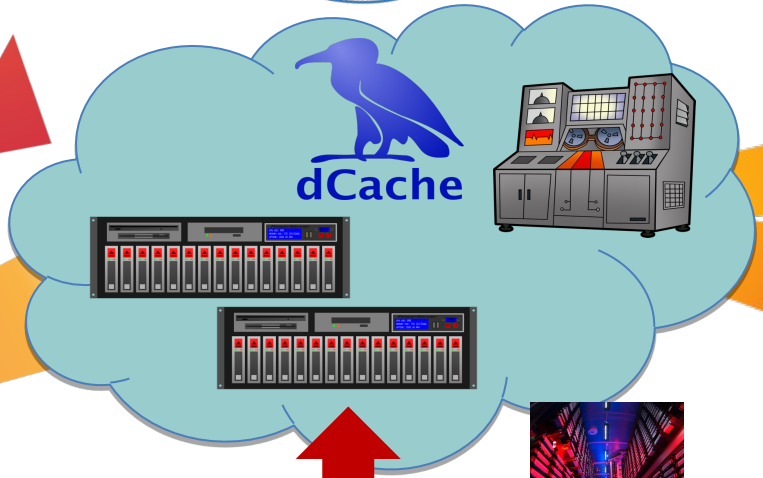


dCache: typical data flows and challenges

High speed
data ingest

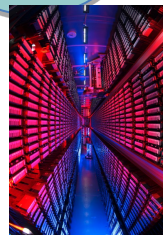
Data management
flow control

Batch processing



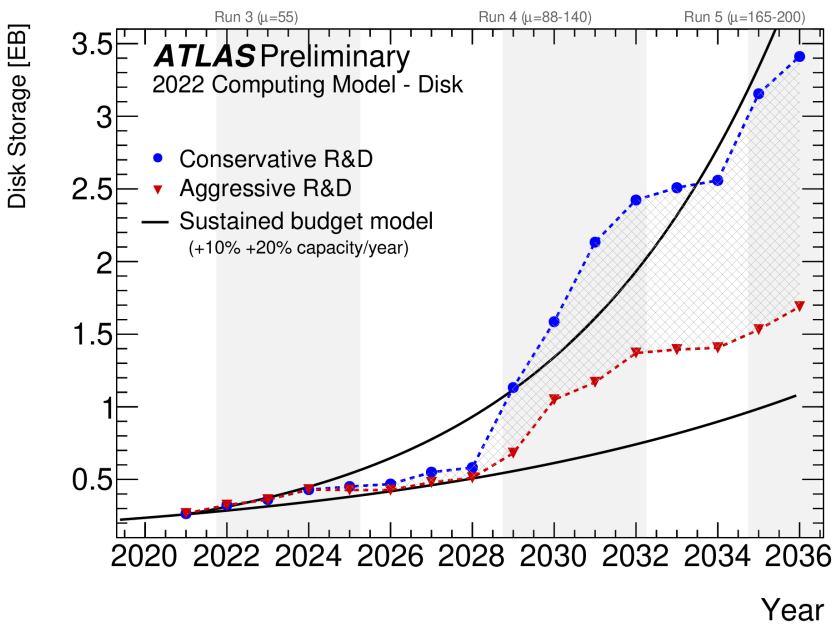
Interactive analysis

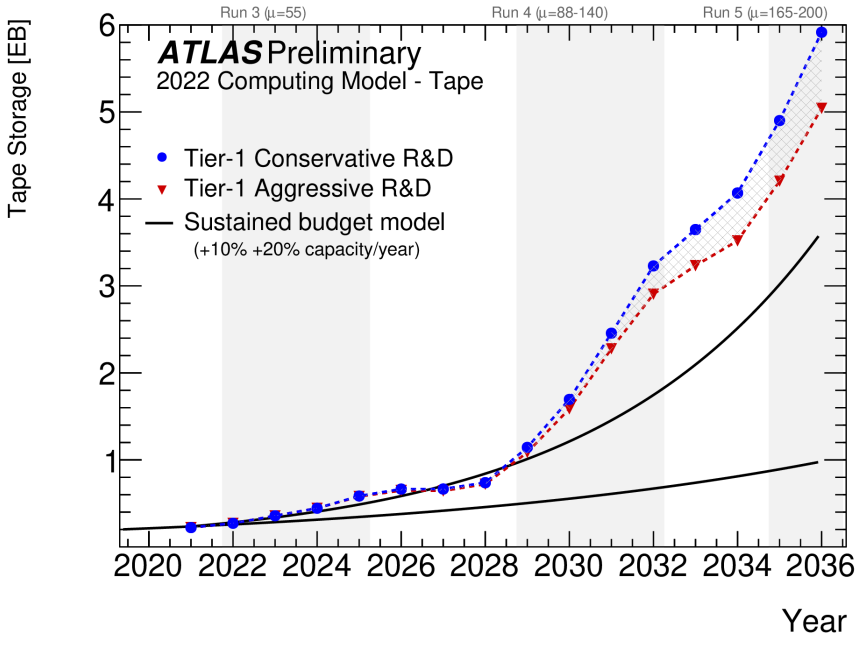
Tape archive



Wide Area Transfers

The Challenge





Source: ATLAS Software and Computing HL-LHC Roadmap
https://cds.cern.ch/record/2802918

- Data intensive science is expected to be … even more data intensive
- Flat budgets => reliance on tape for long term storage.

Takeaway for dCache

- Cache / tape ratio is expected to drop.
  - Fosters more managed sequential bulk data processing cycle (“tape carousel”).
  - Optimization of tape access on storage back-end becomes paramount:
    - Tape access is intrinsically sequential.
    - Massive tape recalls accessing many tapes in parallel should be implemented in a way that:
      - Access to one tape does not block other requests to other tapes.
      - Files recalls are grouped and fed into HSM in a way that minimizes volume mounts and seeks
        https://inspirehep.net/files/6ed866ef28a432cb2a3895634a54ce28
- Access to cached data for bulk processing over low latency network should not suffer from interference from slower WAN transfers and (more) chaotic user analysis activity.
- Redundancy of dCache services should eliminate or minimize downtimes so that bulk data processing is not impacted as any minute will count more as we go into the future.

How do we meet the challenges

- Scale out
  - Namespace
  - Number of pools
- Federated identity
  - Token based AA
- Improve *Analysis Facility* support
  - POSIX access and protocol compliance
  - HPC workload support
- Storage QoS and data lifecycle
- Tape:
  - WLCG Tape REST API
  - Integration with CTA

Standards Everywhere…

HTTP, NFS, FTP

Clients accessing dCache with standard protocols.

SSE

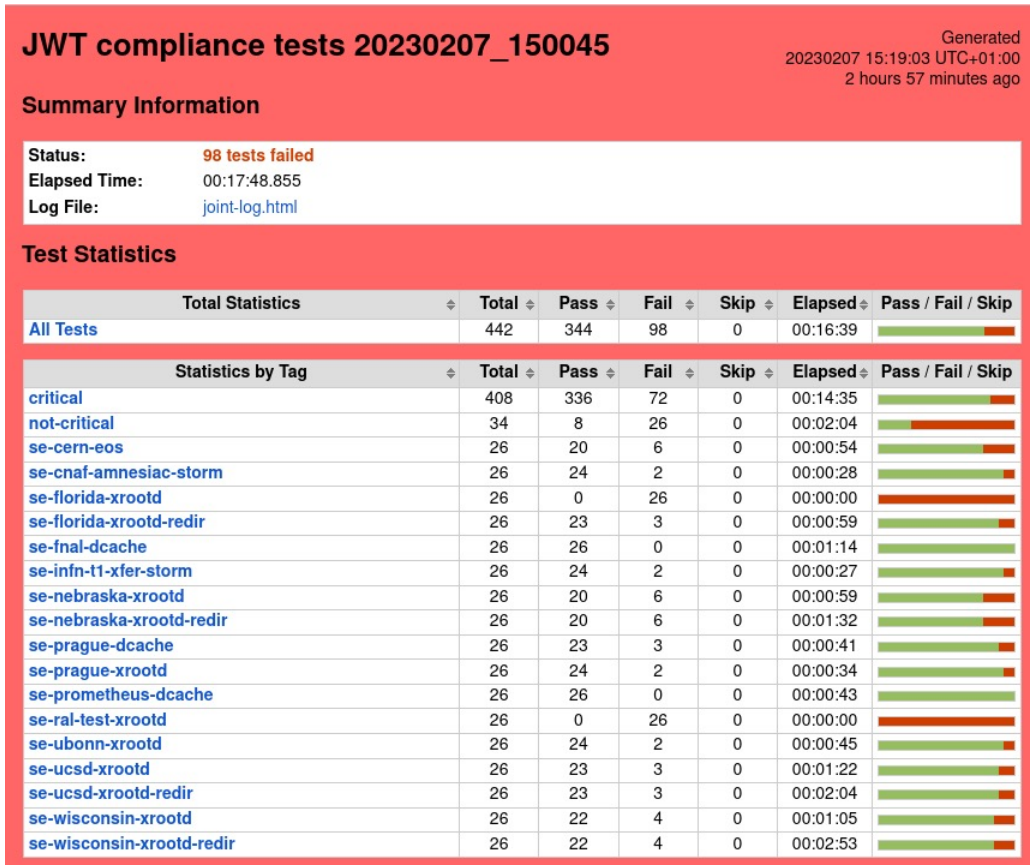
Workflow and processing build on top of widely used (de-facto standard) tools.

User/Group Quotas

- dCache was conceived as a disk cache buffer in front of tape HSM.
  - Cache is managed based on LRU.
  - Capacity of HSM system was assumed to be infinite.
- => There has been no provision for user quotas.
- As system is also used in disk-only (no HSM back-end) setups, the need in user and group quotas has been recognized.
- We have added user and group quota support in dCache 7.2 release. See poster session https://indico.jlab.org/event/459/contributions/11352/ at this conference.
- Features:
  - Lazy quota updates, based on periodic namespace scans.
    - Users are allowed to go over quota until next scan is run.
    - Removed space is not reclaimed instantly.
  - Global per file system.
    - No quota per directory tree.
  - Respects file Retention Policy:
    - Allows to limit ”disk” and “tape” used space by UID/GID

Token-based Authn/Authz

- Core functionality is there.
- Sites have started deployments.
- 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	

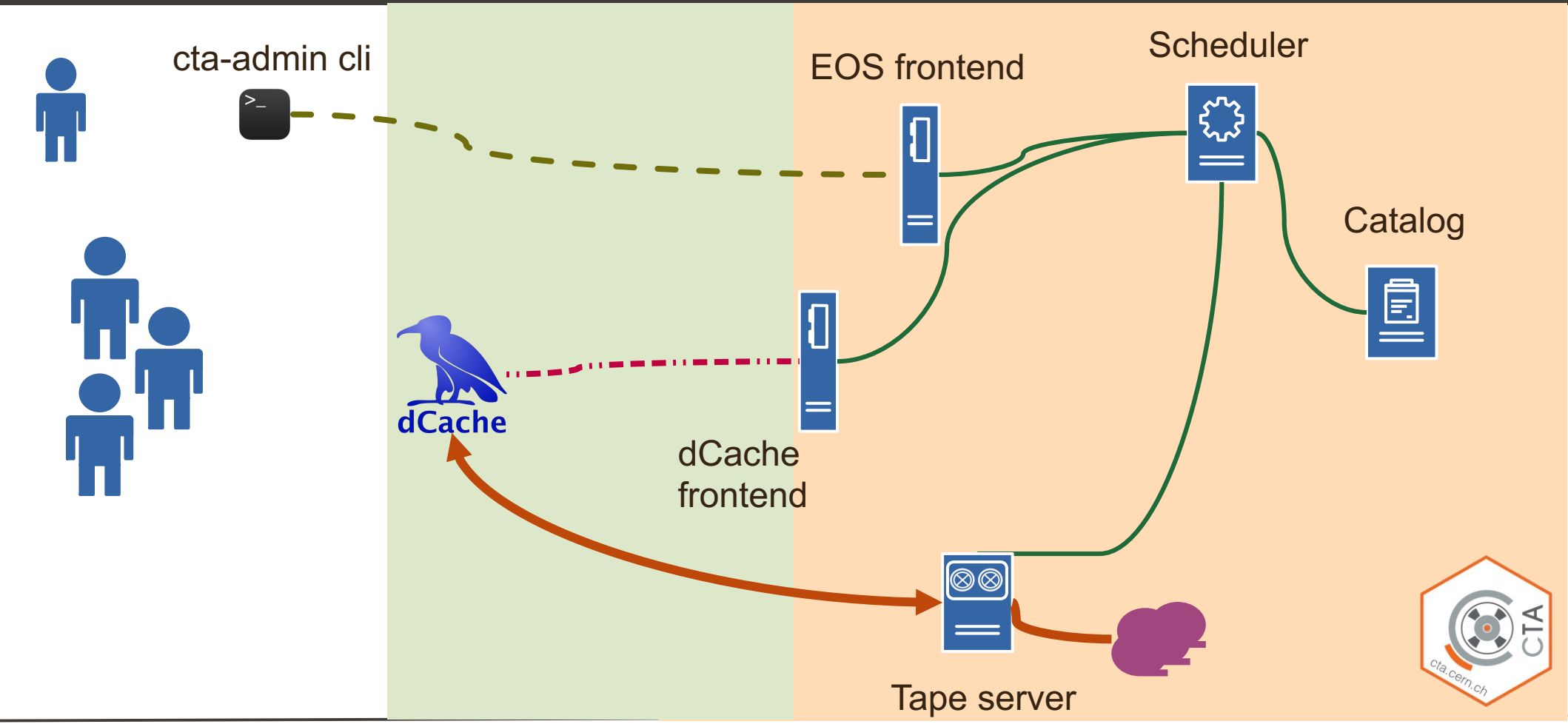
Tape rest API

https://example.org:3880/api/v1

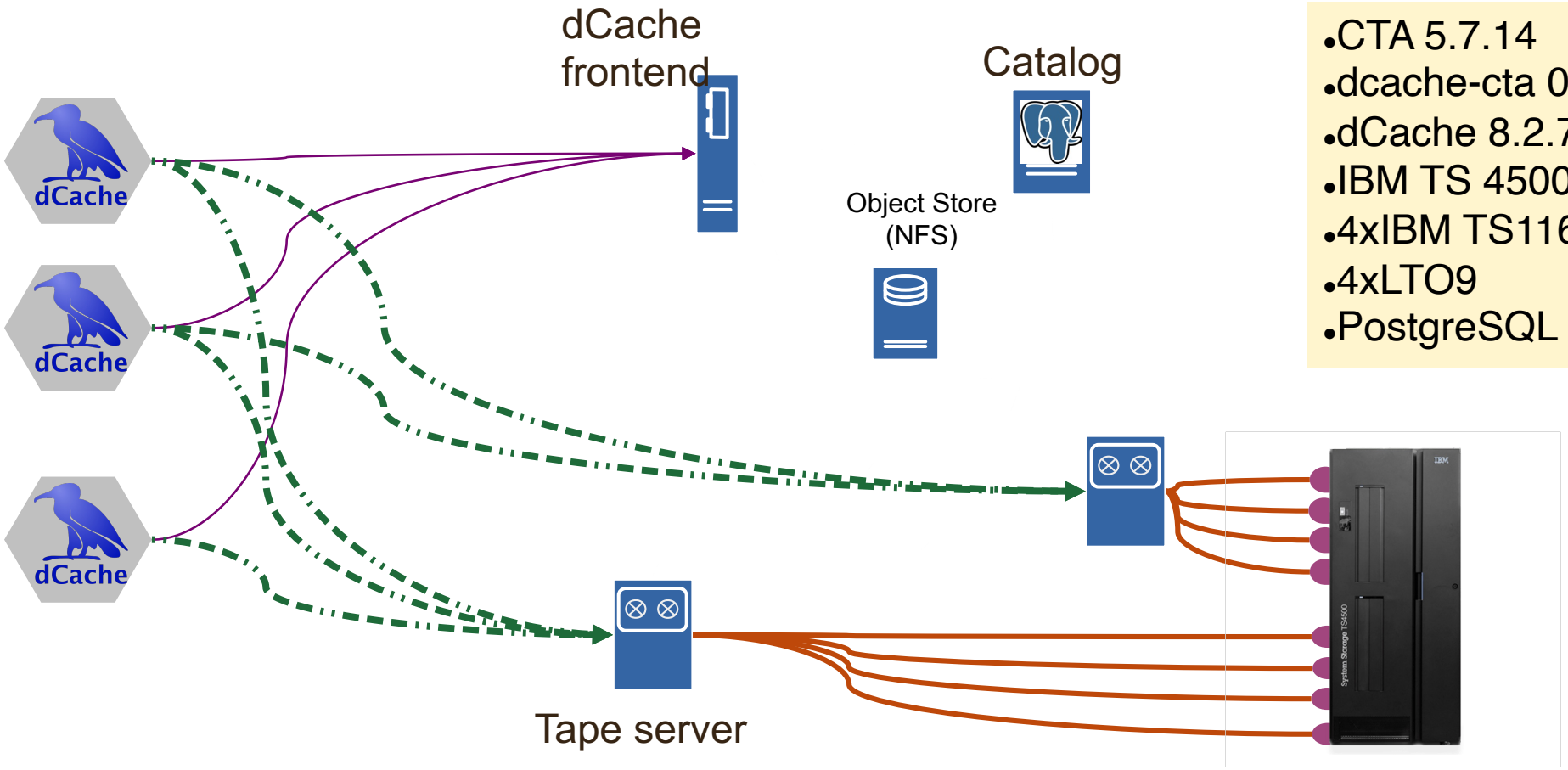
dCache bulk API

WLCG Tape API

Integration with CTA



Production Deployment at DESY

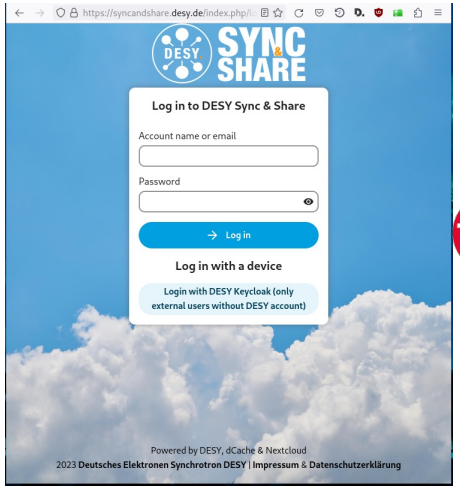


- CTA 5.7.14
- dcache-cta 0.9.0
- dCache 8.2.7
- IBM TS 4500
- 4xIBM TS1160
- 4xLTO9
- PostgreSQL 14.1

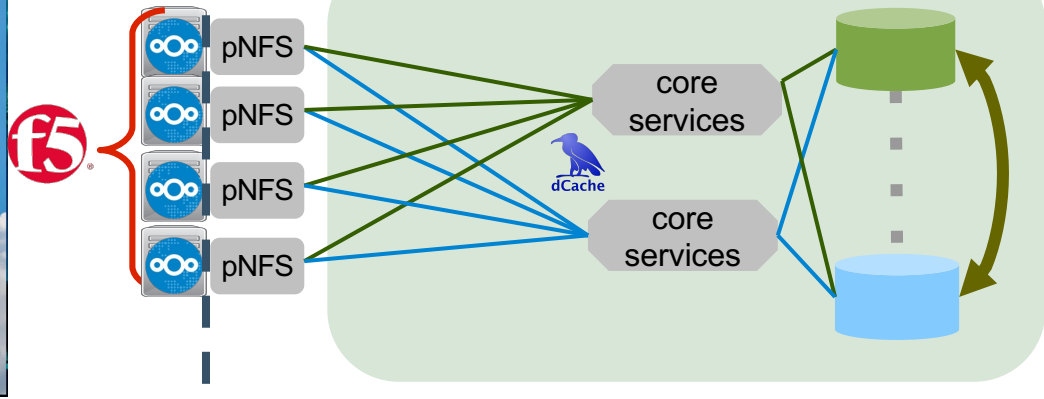
dCache+CTA Status

- Seamless integration with dCache is merged into upstream CTA code at CERN.
  - The latest official CERN releases provides dCache required functionality.
  - The proposed dCache interface utilizing gRPC is being adopted for EOS as well.
- The existing ENSTORE/OSM tape format is supported for READ.
  - The ENSTORE/OSM tape catalog conversion procedures are successfully tested at DESY and Fermilab.
  - All HERA experiments and BELLE-II at DESY migrated to CTA (5.4 PB)
  - EuXFEL migration will take place during summer shutdown (99 PB)
- dCache+CTA deployment at other HEP sites.
  - Fermilab and PIC Barcelona have successfully replicated DESY setup in test environment (production currently uses dCache + ENSTORE).
  - RAL in UK plans to migrate to PostgreSQL from ORACLE based on our experience.

NextCloud Instance @ DESY



NextCloud-25



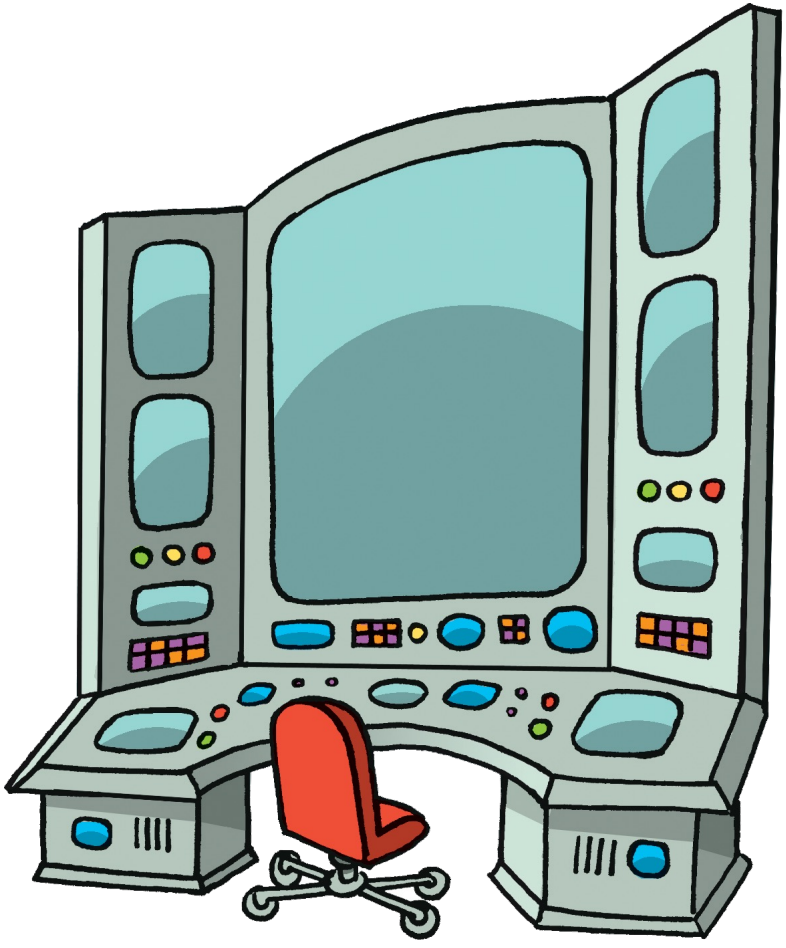
dCache-8.2

dCache as a Cloud 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.

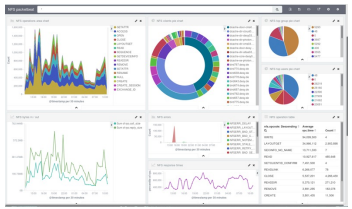
Scale out - some numbers

- 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



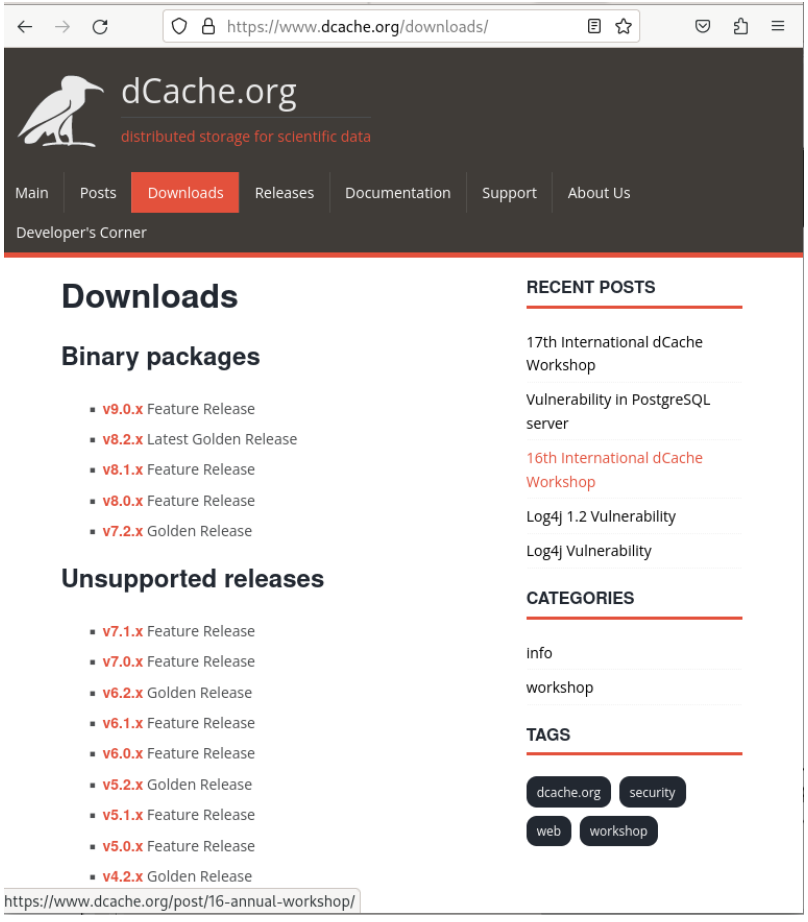
Self-Adaptive dCache

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



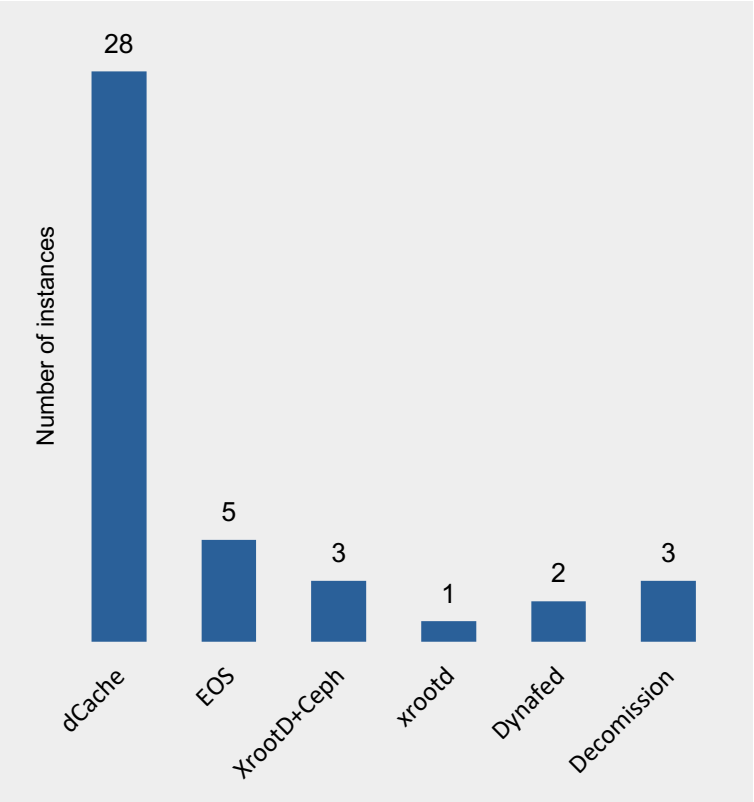
Supported OS platforms

- 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

- Influx of new users.
- Series of tutorials.
- Help from EGI.
  - Many thanks to Petr Vokac for facilitating the transition.



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

Summary of major changes in dCache

- BULK Service.
- WLCG Tape API.
- WLCG/Scitokens support.
- TPC improvements.
- NFSv4.1/pNFS improvements.
- XROOT evolution (TLS, tokens, TPC, proxy-IO).
- Namespace performance improvements.
- User/Group quota.
- CTA integration.

More info:

https://dcache.org

Contribute/complain:

https://github.com/dCache/dcache

Help and support:

support@dcache.org,

user-forum@dcache.org

Developers:

dev@dcache.org

Thank You!

Come to Berlin this summer

© HTW Berlin/Alexander Rentsch

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

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