



dCache CTA integration: Status, Experience, Plans.

*Tigran Mkrtchyan
for the dCache collaboration
&
DESY tape team.*



Our thanks to CTA developers for support and prompt response when help is needed!

What is DESY (as storage)

Experiments/Community	Service
EuXFEL, Petra-III, ILC, Accelerator R&D, ...	Primary data site (Tier-0). Provides online, nearline and archival storage.
Belle-II, ...	Provides online and near-line storage.
Atlas, CMS, LHCb	Online only.
H1, Hermes, Hera-B, Zeus , ...	Provides online and archival storage.

Multiple Faces of Tape

At Tier-0

- High data ingest rate
- Multiple parallel streams
- High durability, multiple copies on different media
- Long-term nearline access
- Small file handling

At analysis facility

- Automatic data migration
- Bulk recall on periodic basis
- Long-term nearline access
- Recall prioritization

Data Archive

- Manual data migration
- Long-term preservation
- Automatic technology migration
- Self-healing

Technology in Place at DESY

Hardware

- 2x Oracle SL8500 - EOL
 - 26x LTO-8 drives
- 2x IBM TS4500 (since 2021)
 - 20x Jaguar
 - 18x LTO-9
 - Different buildings (500m)

Software

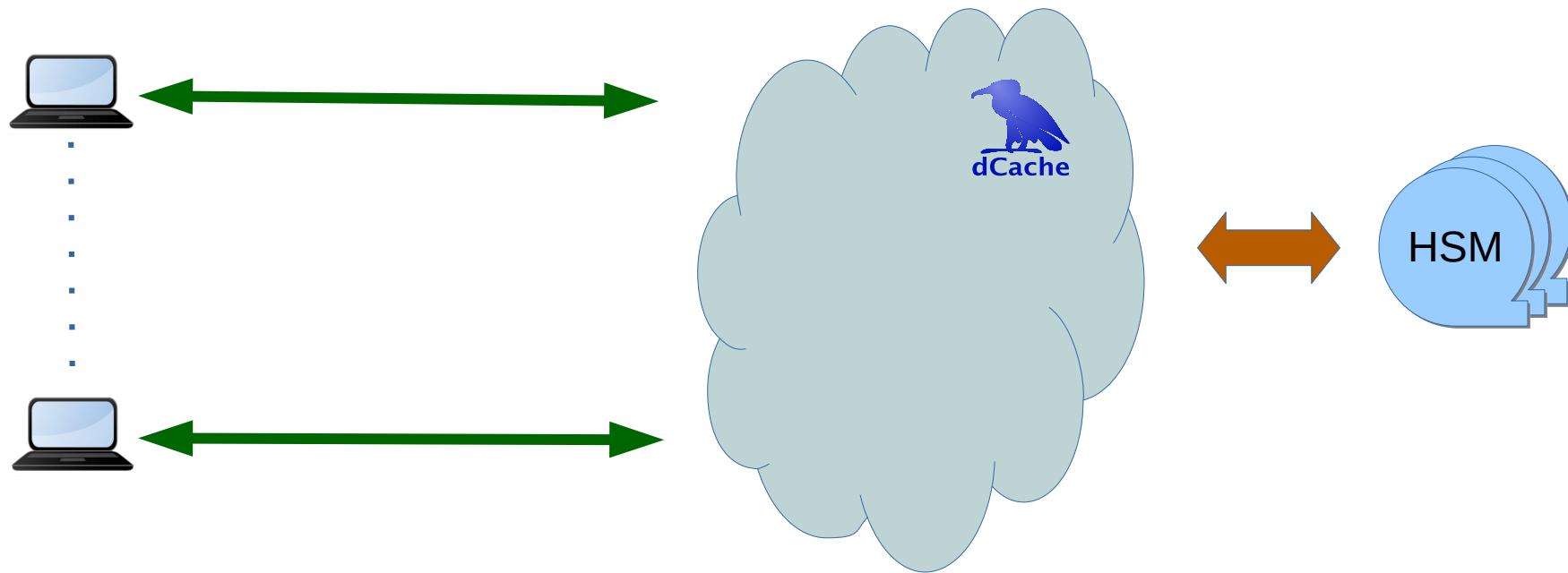
- TSM (IBM Spectrum Protect) – classic backup
- dCache – interface to HSM system
 - Scientific Data
 - AFS/Mail backup
- OSM (Open Storage Manager)
 - Since 1994
 - Proprietary software, maintained by DESY (~80% local modifications)
- CTA
 - Since 2022

Tape Software Requirements

- Maximize tape HW efficiency
 - Integration into DESY ecosystem
 - Integration with dCache tape interface
- Stable operation for a next decade
- Should be Open-source, adopting open standards
- Wide user and technology community



dCache+HSM Tandem (DESY)



All access to scientific data on tapes goes exclusively through dCache!

dCache Tape Connectivity



- Write-back / Read-through cache behavior
- Transparent for the users
- Available via all protocols (subject to authorization)
- Supports multiple HSM on a single instance
- Stores tape location as opaque data provided by HSM

Interfaces to HSM

- Execute external migration script
 - Stupid, Simple, Genius ...
 - Reference implementation of driver API
- Plugable driver Java API:
 - Suitable to create efficient HSM connectivity
 - ENDIT (*Efficient Northern dCache Interface to TSM*)

dCache HSM Interface

```
// dCache interface to tape system

public interface NearlineStorage {

    void flush(Iterable<FlushRequest> requests);
    void stage(Iterable<StageRequest> requests);
    void remove(Iterable<RemoveRequest> requests);

    void cancel(UUID uuid);

    // driver initialization methods

    ...
}
```

dCache HSM ⇔ Link

- Files belongs to storage classes

xfel:SQS-2019@osm

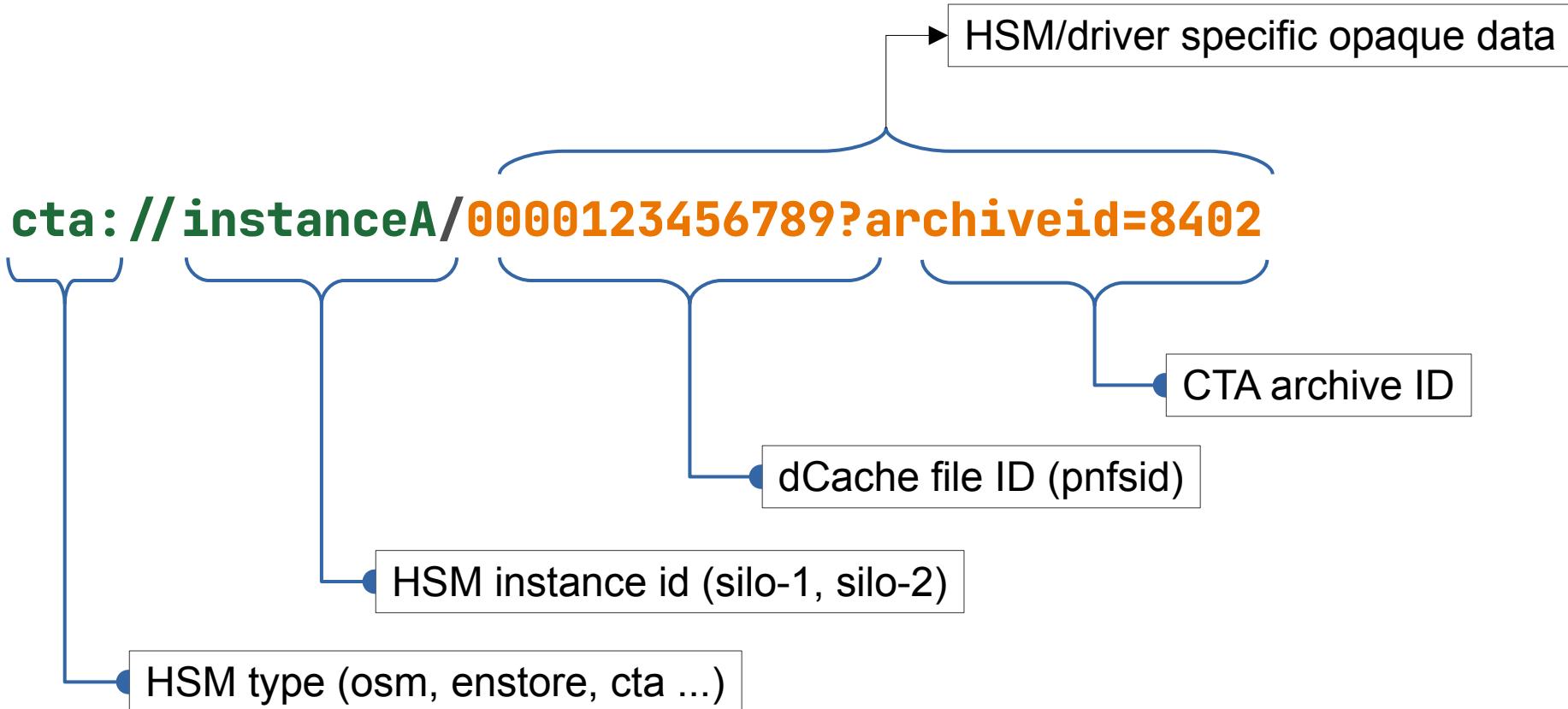
- Configure HSM connectivity on the pool

```
hsm create osm siloA script \  
-command=hsmcp.py
```

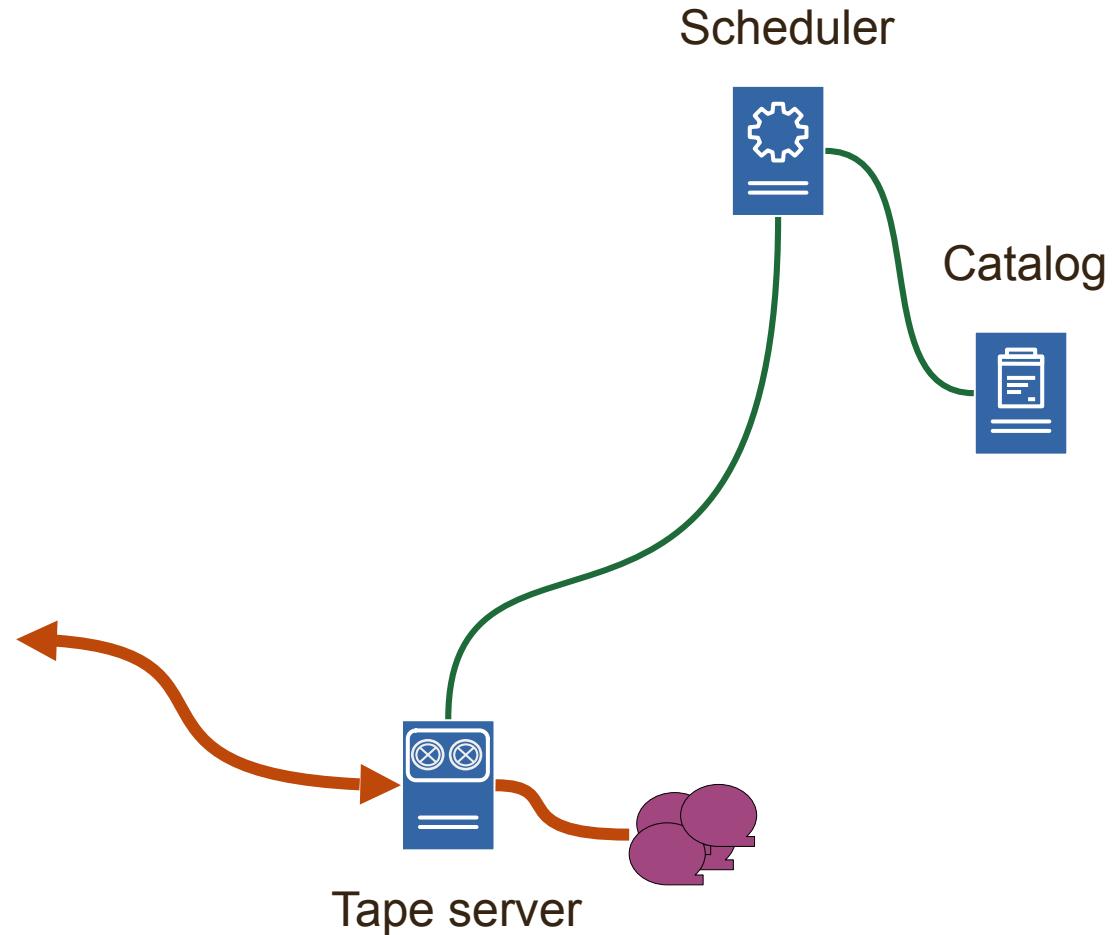
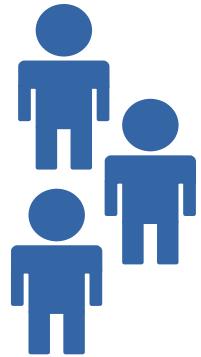
- Tape location stored in namespace as URI

osm://siloA/xxxxxxxxxxxxxx

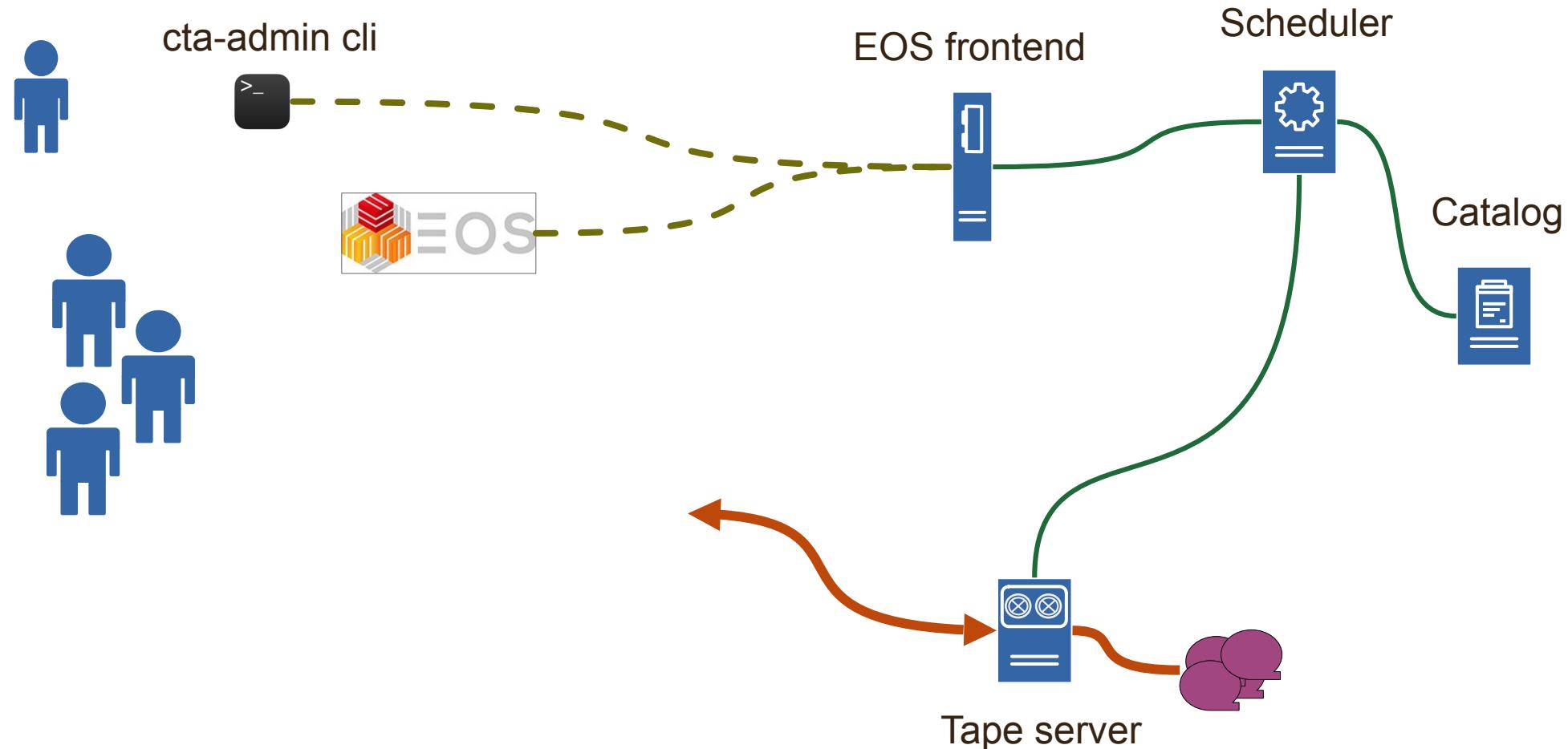
dCache HSM ⇔ Link



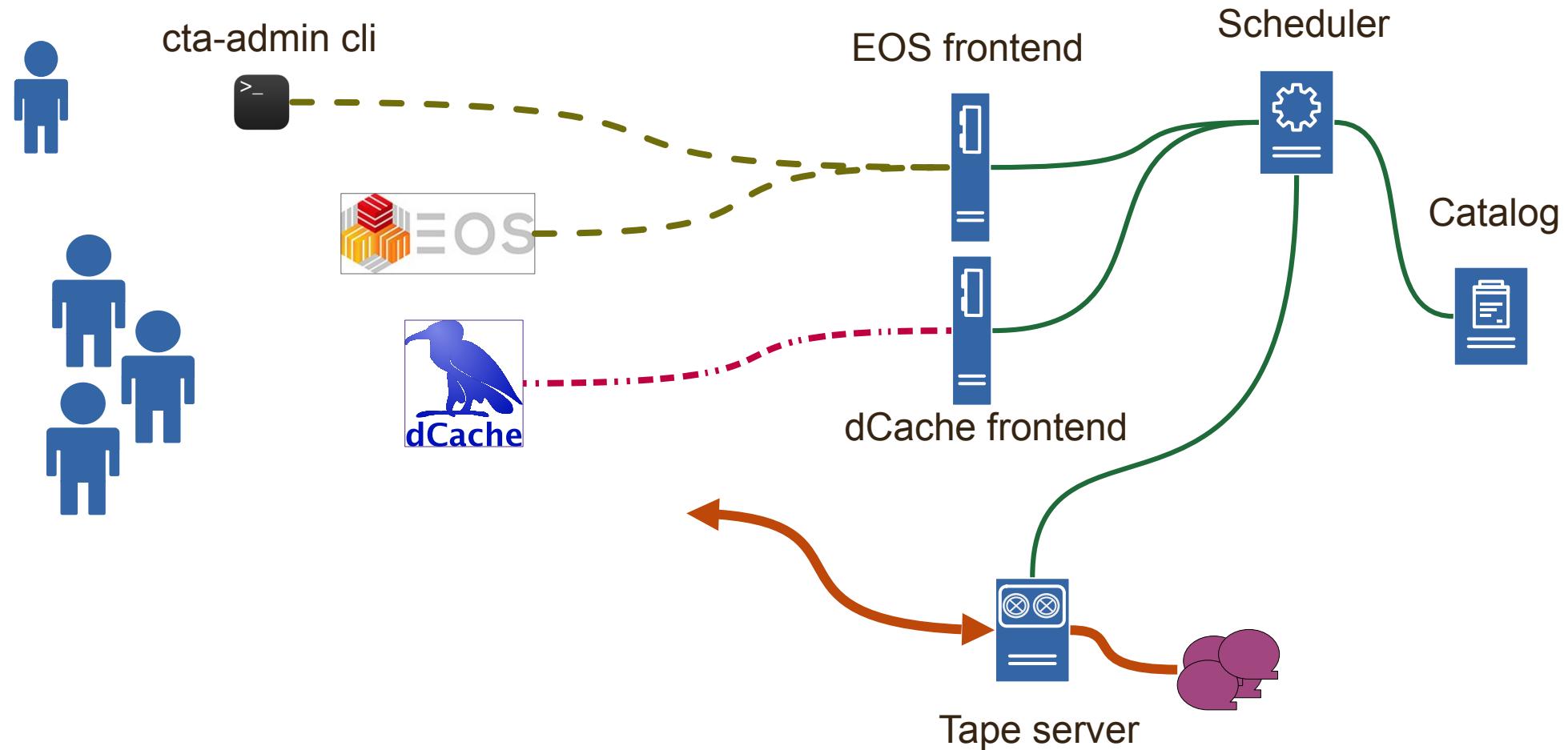
(Extremely) Simplified CTA design



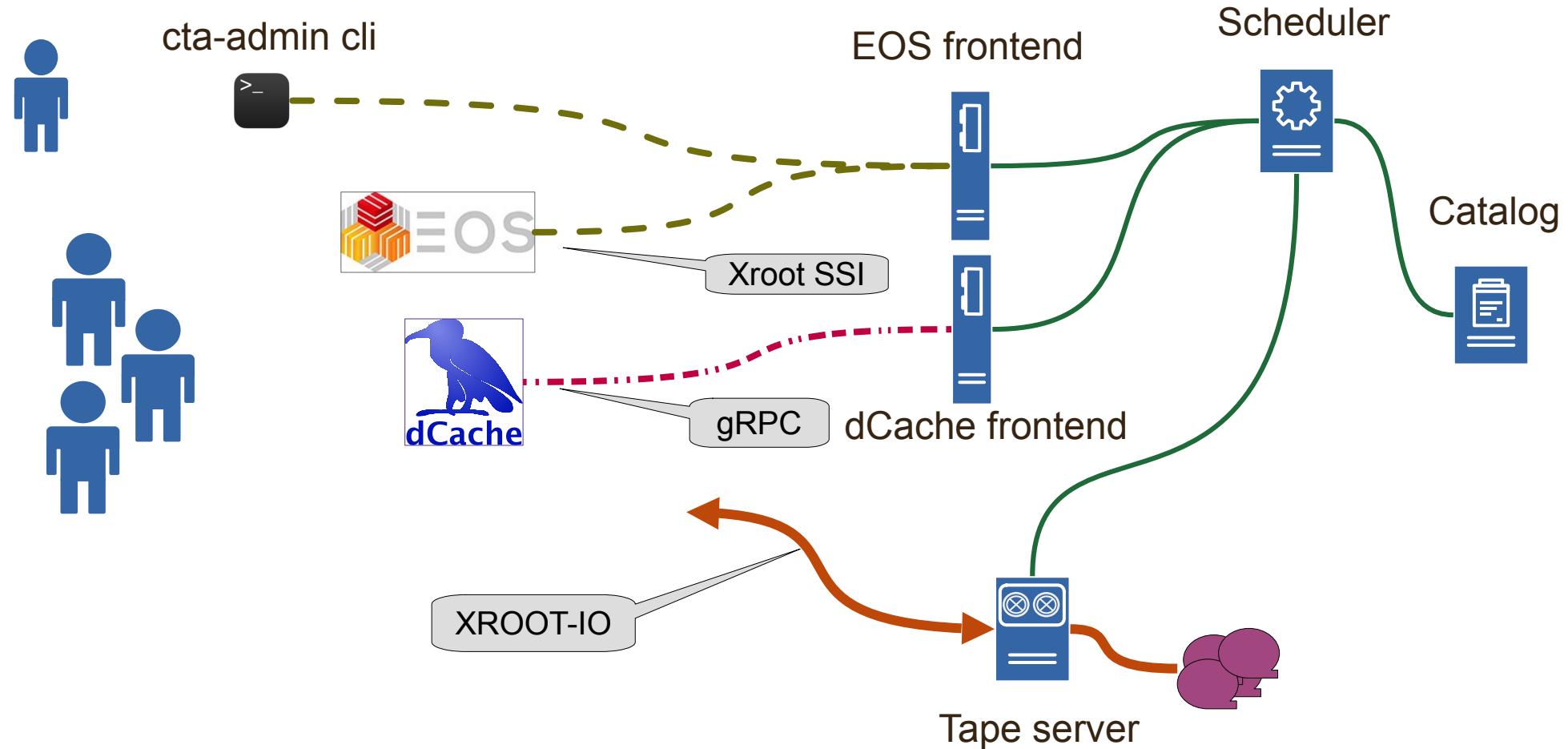
(Extremely) Simplified CTA design



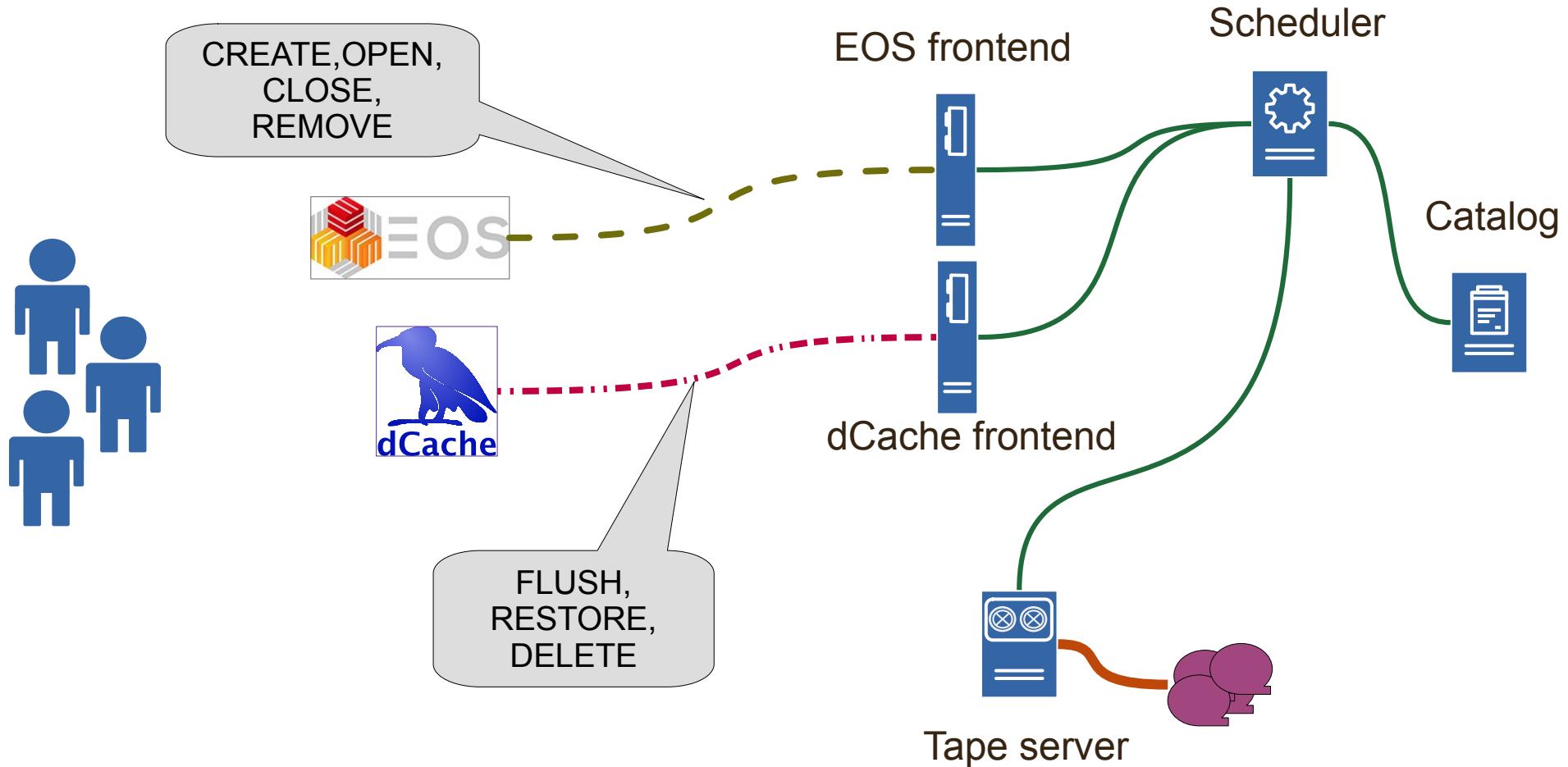
(Extremely) Simplified CTA design



(Extremely) Simplified CTA design



(Extremely) Simplified CTA design



dCache CTA gRPC

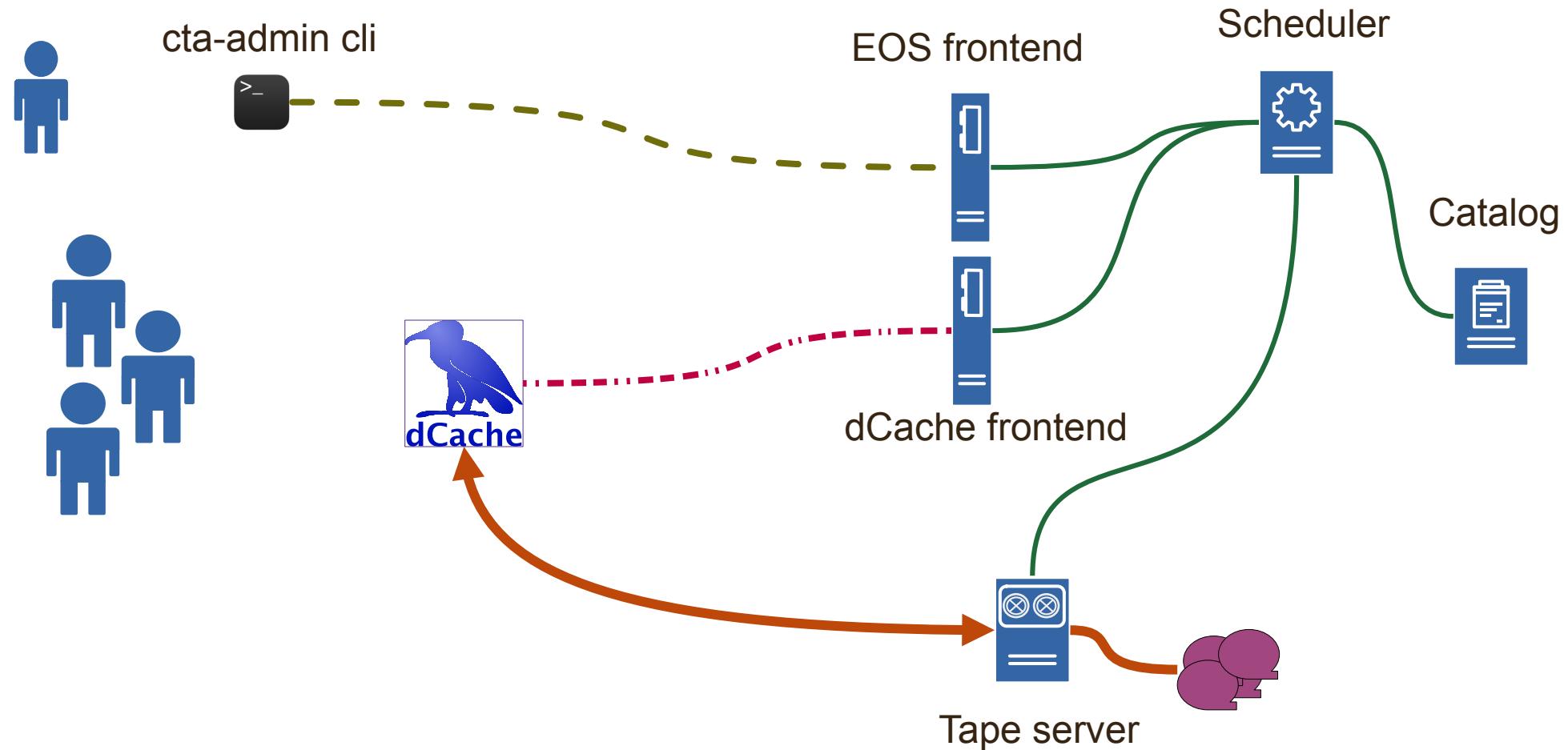
```
// gRPC definition of dcache-cta interface

service CtaRpc {
    rpc Version (google.protobuf.Empty) returns (cta.admin.Version) {}

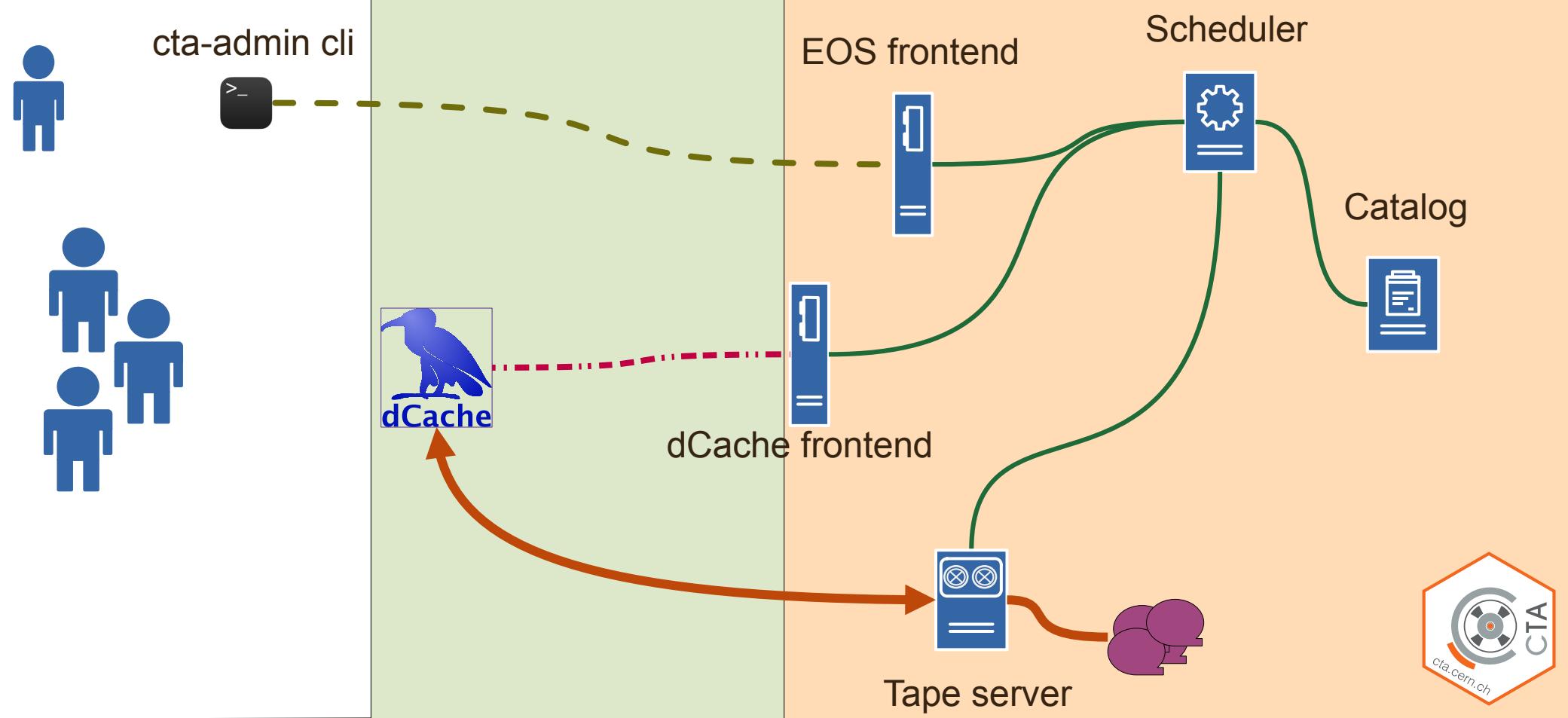
    rpc Archive (ArchiveRequest) returns (ArchiveResponse) {}
    rpc Retrieve (RetrieveRequest) returns (RetrieveResponse) {}
    rpc Delete (DeleteRequest) returns (google.protobuf.Empty) {}
    rpc CancelRetrieve (CancelRequest) returns (google.protobuf.Empty) {}

}
```

(Extremely) Simplified CTA design



Deployment at DESY



dCache ⇔ CTA Integration

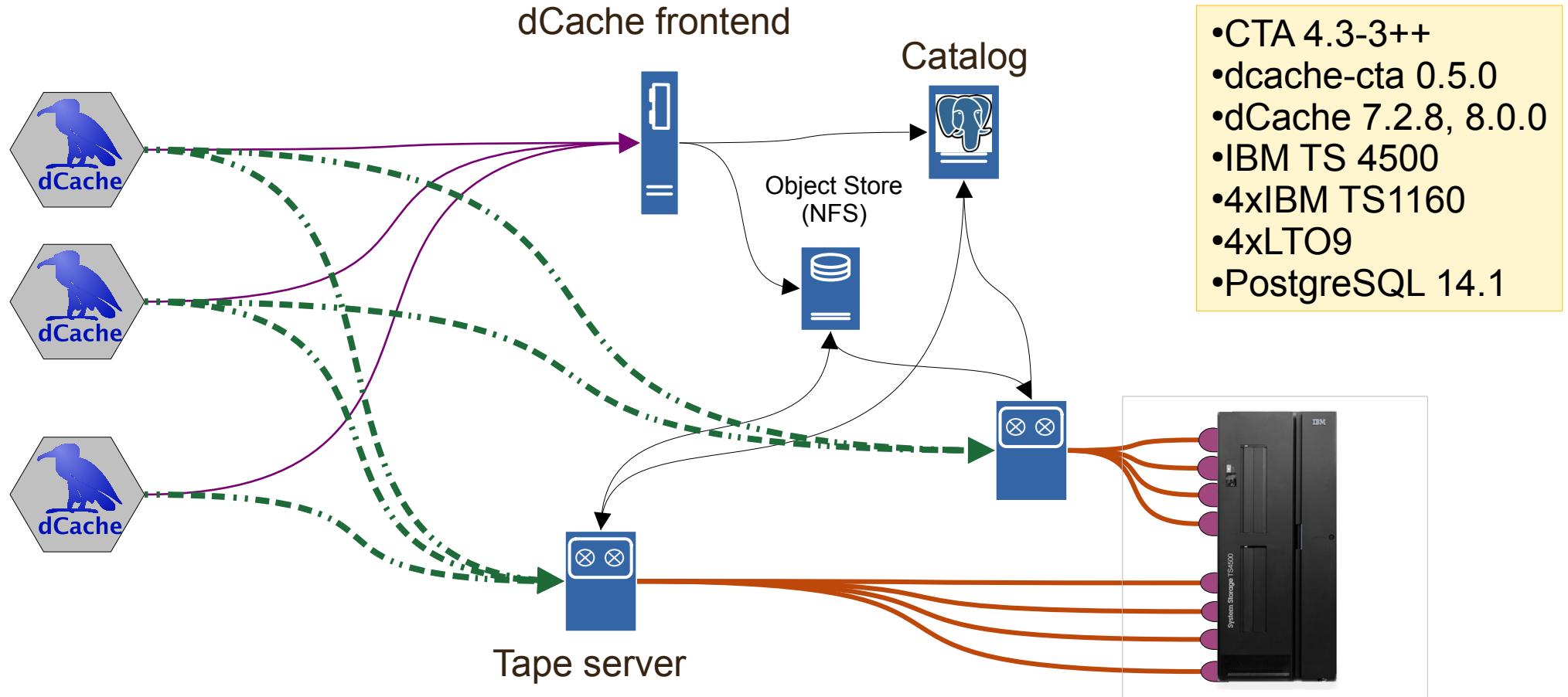
dCache

- Nearline driver to add (dCache $\geq 7.2.2$)
- Can run in parallel with other HSMs
- Pre-scheduling on pools should be disabled/reduced
- File path, uid, gid not preserved

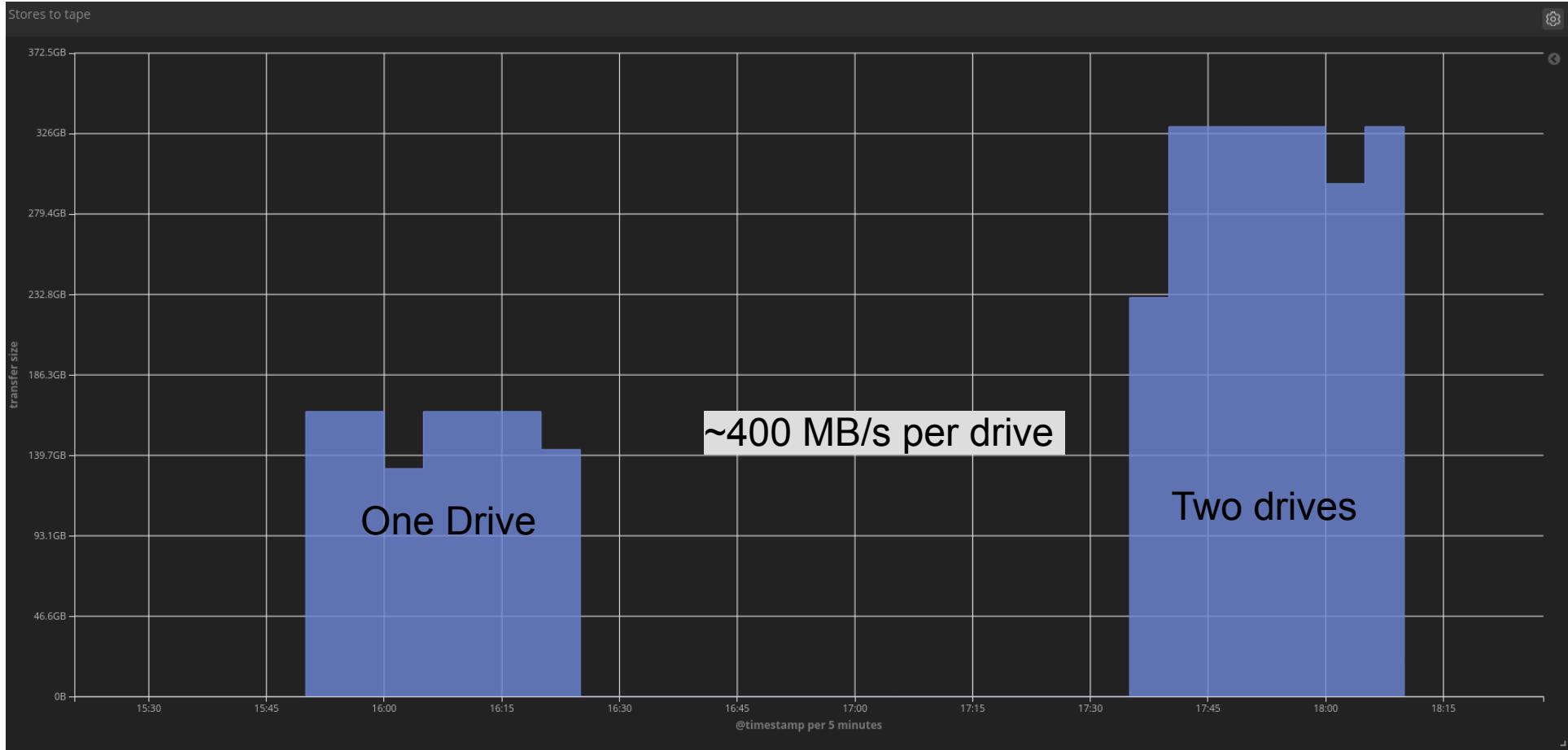
CTA

- Additional *cta-dcache* service (packaged as own rpm)
- Limited to dCache required minimal functionality
 - *cta-frontend* still needed for admin commands

Deployment at DESY



Some Test Results



Tested Use Cases

- Multiple drives on the same host
 - We plan 4 drives per node
- Multiple tape copies
- Multiple storage classes /file families
- JAG + LTO9 tapes in parallel
- PostgreSQL



Missing Parts (for DESY)

- Multiple tape formats are not supported
 - We still need (at least) to read old files
 - Prototype is under testing
- OMS ⇒ CTA catalog migration
- dCache support out-of-box
 - No custom builds
- Puppetization
 - Thanks to CERN for sharing manifests!

Multiple Tape Formats

- Migration should allow reading of existing tapes
- ENSTORE and OSM have their own tape formats
- Fermilab and DESY works on a solution(s)
 - Auto-detection vs field in tape catalog
- Stay tuned
 - The BoF right after this talk!



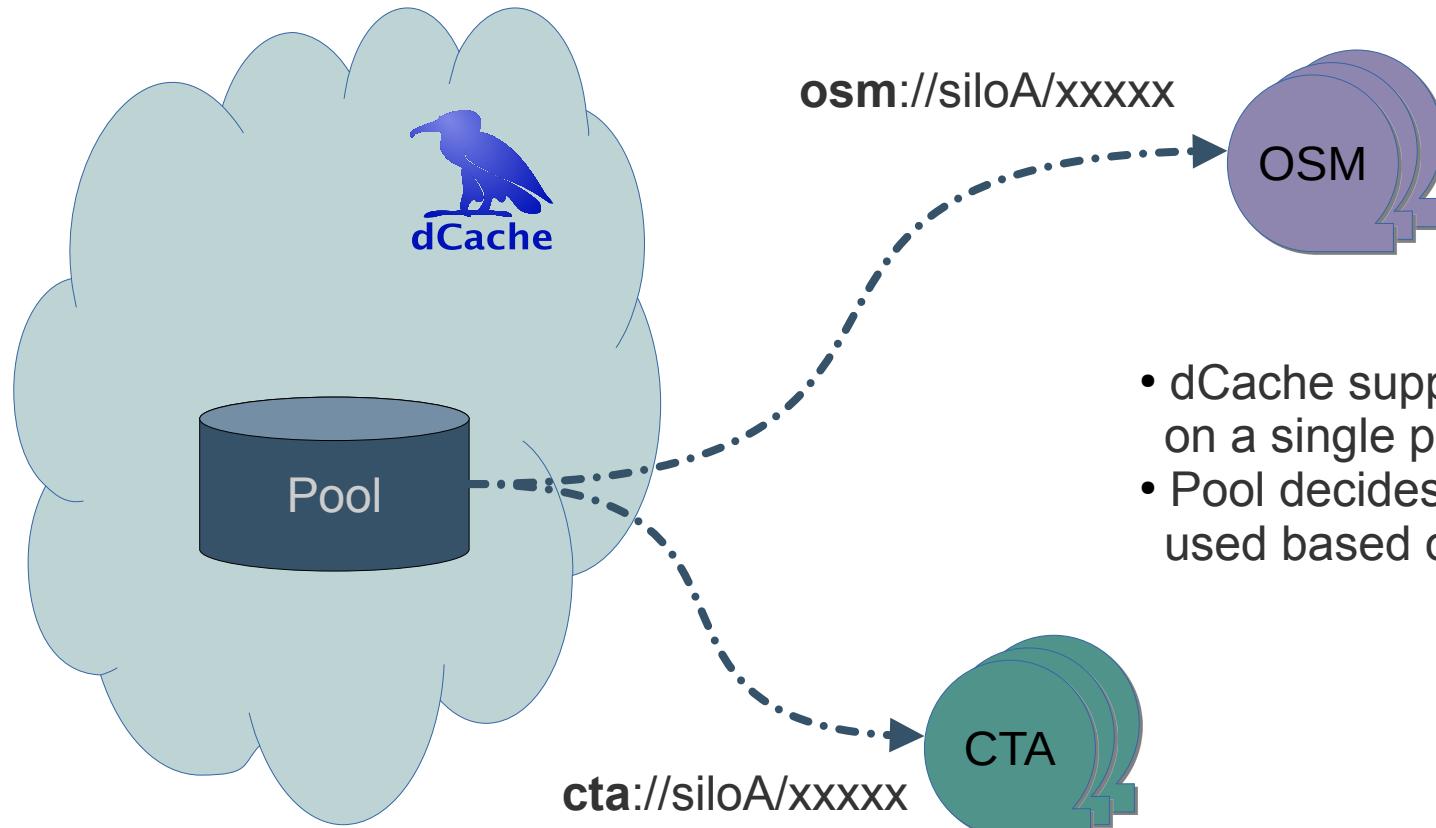
Next Steps at DESY

- Better operational experience
 - More FTE on CTA integration, support, development
- Pilot deployment planned for next week
 - More real-life loads
 - Controlled environment
- Migration path
 - ~~Copy or DB migration~~

Summary

- Tape is an essential part of IT-Services at DESY
- dCache is the only interface to scientific data
 - Tape connectivity dominates the local development
- Enstore and CTA are evaluated as HSM solution
 - Both require on-site development
 - Commercial alternatives are not excluded !
- We expect new system to be in place in 1Q 2022
 - ~6 months to make a decision

Pilot Deployment (DESY)



- dCache support multiple HSMs on a single pool
- Pool decides which driver to used based on HSM type

Summary (DESY)

- Tape is an essential part of IT-Services
- dCache is the only interface to scientific data on tape
 - Tape related activities dominates the local developments
- We see CTA as the preferred tape software at DESY
 - The architecture matches our demands
 - Seamless integration with dCache
- Pilot deployment expect next week(s)

More info

- CTA branch with dCache support
<https://gitlab.cern.ch/cta/CTA/-/tree/cta-dcache>
- dCache-cta HSM driver
<https://github.com/dCache/dcache-cta>
- Documentation
<https://confluence.desy.de/display/ITSC/dCache-CTA>
- Contacts
support@dcache.org

Packages (rpm)

[dcache-cta]

name=dCache CTA repository

baseurl=https://download.dcache.org/nexus/repository/cta/
el\$releasever/\$basearch/

enabled=1

[dcache-cta-driver]

name=dCache nearline storage driver for CTA

baseurl=https://download.dcache.org/nexus/repository/dcachecta

enabled=1

Thanks!



FIO write workload

```
[seq-write]
description=sequential write workload
ioengine=sync
filename_format=$jobname-$jobnum-$filenum
create_on_open=1
; one file at the time
openfiles=1
file_service_type=sequential
create_serialize=1
rw=write
refill_buffers=1
;
; a large number to hit the max specified by `size`
;
nrfiles=10000
size=2000g
; file size range
filesize=10m:16g
bs=1m
```

dCache+HSM Deployments



Script

- HPSS (2)
 - TSM (1)
 - OSM (2)
 - ENSTORE (4)
 - DMF (1)

Driver

- HPSS (1)
 - TSM (3)
 - TapeGuy (1)
 - Sapphire (1)
 - dcache-cta (x)