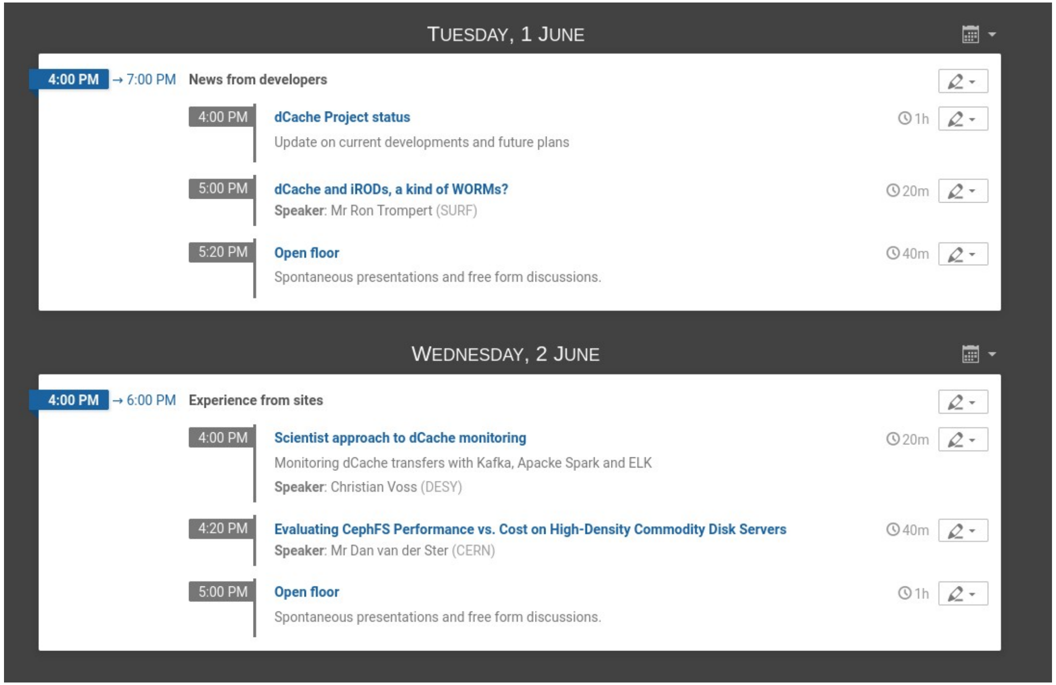
dCache Workshop report

0xF International dCache workshop

The Workshop Format

- Two sessions at 16:00 CEST, 2h each (planned)
  - 3 ½h First day
  - 2 ½h Second day
- Well attended
  - ~60 participants



TUESDAY, 1 JUNE

4:00 PM → 7:00 PM News from developers

4:00 PM **dCache Project status** 1h
Update on current developments and future plans

5:00 PM **dCache and iRODs, a kind of WORMs?** 20m
Speaker: Mr Ron Trompert (SURF)

5:20 PM **Open floor** 40m
Spontaneous presentations and free form discussions.

WEDNESDAY, 2 JUNE

4:00 PM → 6:00 PM Experience from sites

4:00 PM **Scientist approach to dCache monitoring** 20m
Monitoring dCache transfers with Kafka, Apacke Spark and ELK
Speaker: Christian Voss (DESY)

4:20 PM **Evaluating CephFS Performance vs. Cost on High-Density Commodity Disk Servers** 40m
Speaker: Mr Dan van der Ster (CERN)

5:00 PM **Open floor** 1h
Spontaneous presentations and free form discussions.

The Main Topics

- Breaking changes in the new versions
- Developments in metadata handling
- QoS and HSM integration
- Token based AuthN in Xroot protocol
- Changes in Storage Resource Reporting
- Monitoring of large installations

dCache News, Status & Roadmap

0xF International dCache workshop

Breaking Changes ...

2021-06-01 dCache project status & plans 9/46

Roadmap

2021-06-01 dCache project status & plans

Metadata

dCache project status & plans 15/46

HSM & QoS

2021-06-01 22/46

Tokens & IdP Feder

Community effort

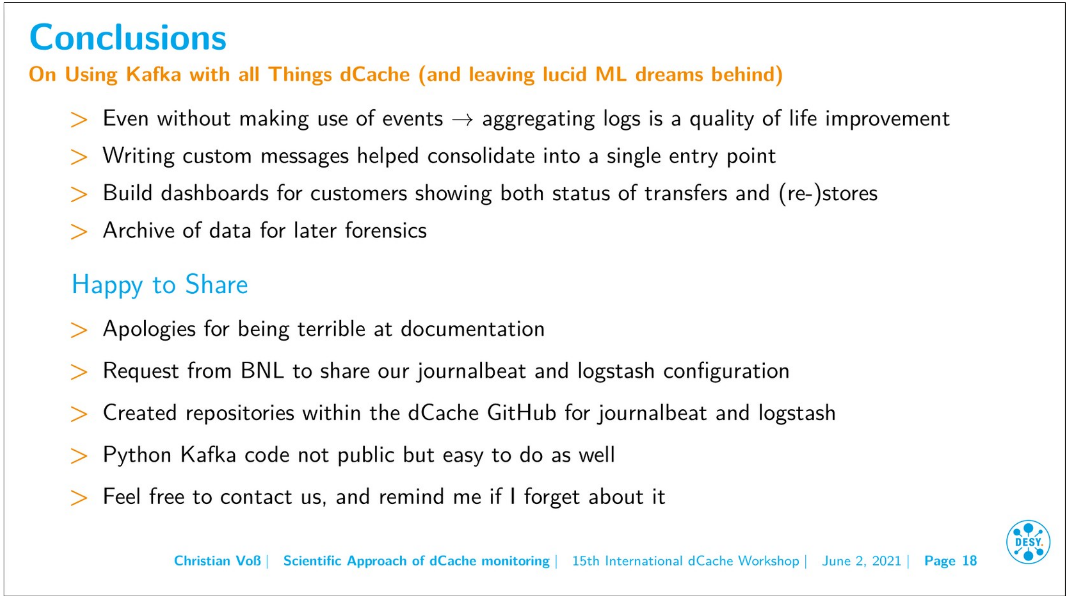
2021-06-01 dCache project status & plans 45/46

Storage Resource Reporting (aka SRR)

2021-06-01 dCache project status & plans 37/46

Scientific Approach to monitoring

Using popular Big-data tools to analyze dCache access information



Conclusions

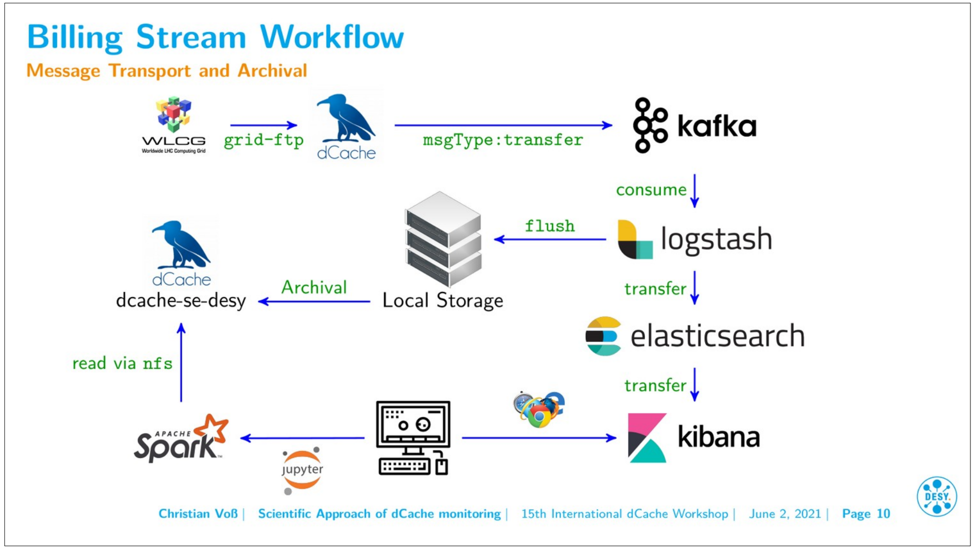
On Using Kafka with all Things dCache (and leaving lucid ML dreams behind)

- Even without making use of events → aggregating logs is a quality of life improvement
- Writing custom messages helped consolidate into a single entry point
- Build dashboards for customers showing both status of transfers and (re-)stores
- Archive of data for later forensics

Happy to Share

- Apologies for being terrible at documentation
- Request from BNL to share our journalbeat and logstash configuration
- Created repositories within the dCache GitHub for journalbeat and logstash
- Python Kafka code not public but easy to do as well
- Feel free to contact us, and remind me if I forget about it

Christian Voß | Scientific Approach of dCache monitoring | 15th International dCache Workshop | June 2, 2021 | Page 18



Billing Stream Workflow

Message Transport and Archival

WLCG

grid-ftp

dCache

msgType:transfer

kafka

consume

logstash

flush

Local Storage

Archival

dCache

dcache-se-desy

transfer

elasticsearch

read via nfs

transfer

kibana

Spark

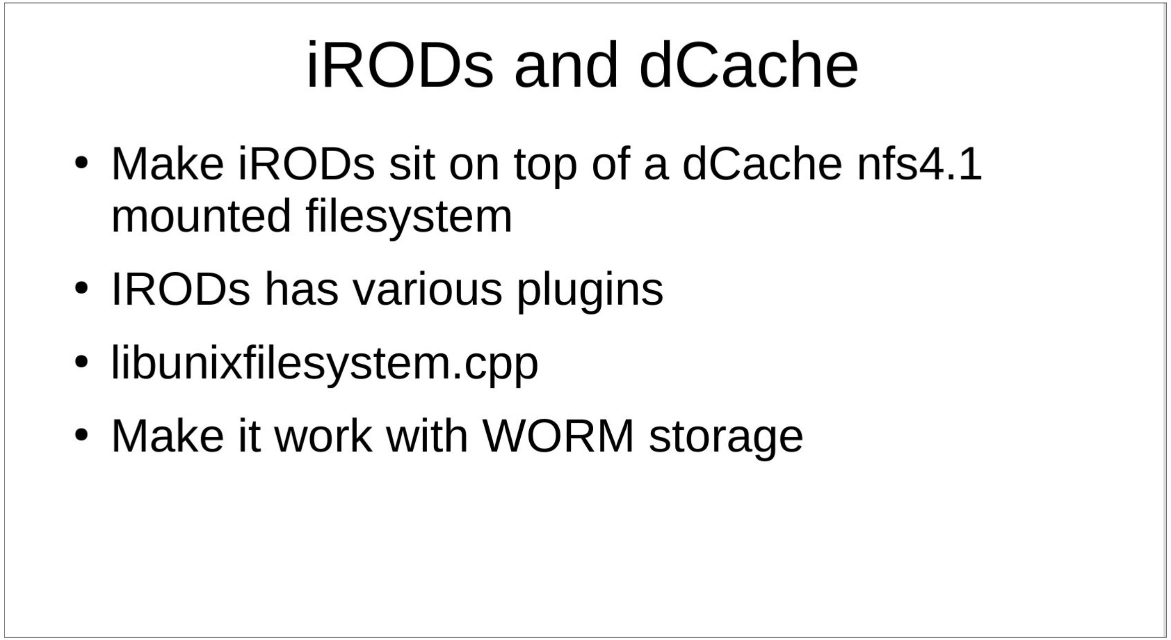
jupyter

Christian Voß | Scientific Approach of dCache monitoring | 15th International dCache Workshop | June 2, 2021 | Page 10

By Christian Voss, DESY

dCache and iRODS

- Some sites do have iRODS & dCache in parallel
- A deeper integration is required
- dCache developers have a direct contact with iRODS team to address issues

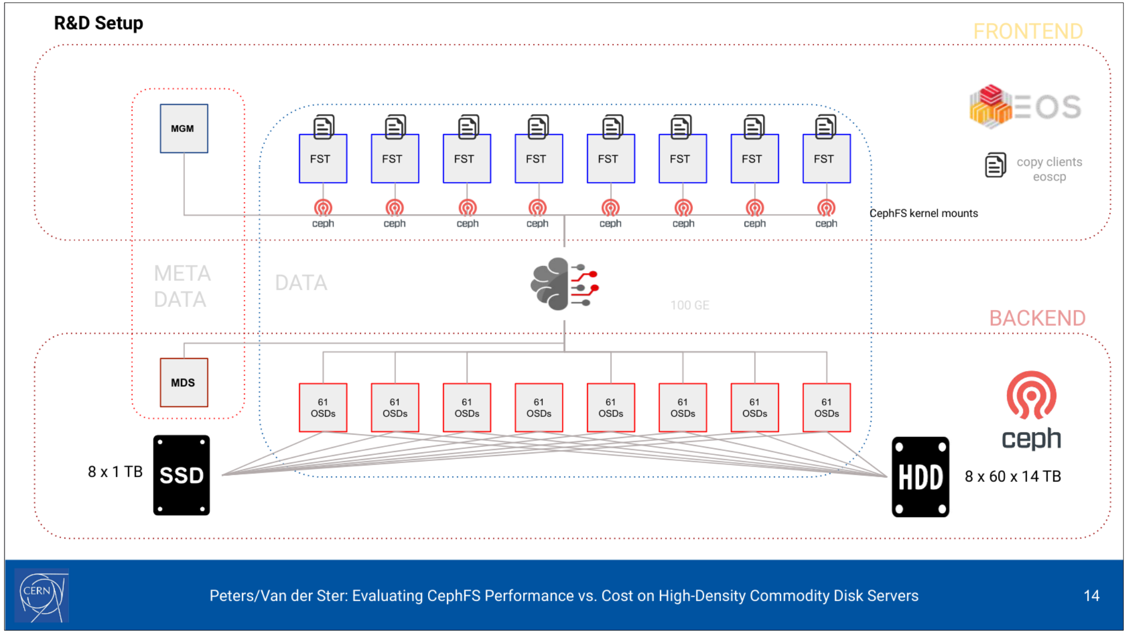


iRODs and dCache

- Make iRODs sit on top of a dCache nfs4.1 mounted filesystem
- IRODs has various plugins
- libunixfilesystem.cpp
- Make it work with WORM storage

By Ron Trompert, SURF

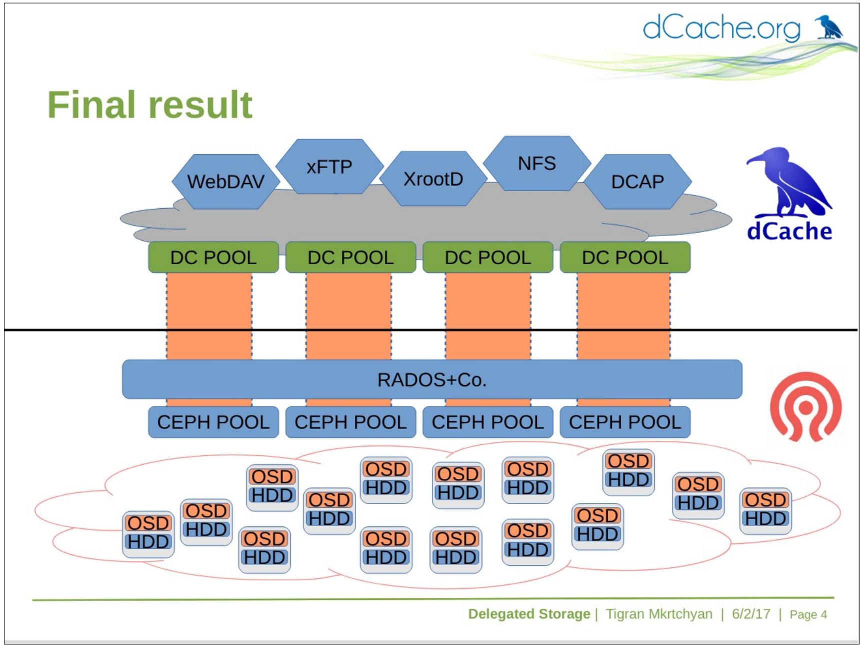
EOS Presentation?!



R&D Setup

Peters/Van der Ster: Evaluating CephFS Performance vs. Cost on High-Density Commodity Disk Servers 14

EOS and dCache trying to solve the same problem, have similar architecture and back-end storage requirements



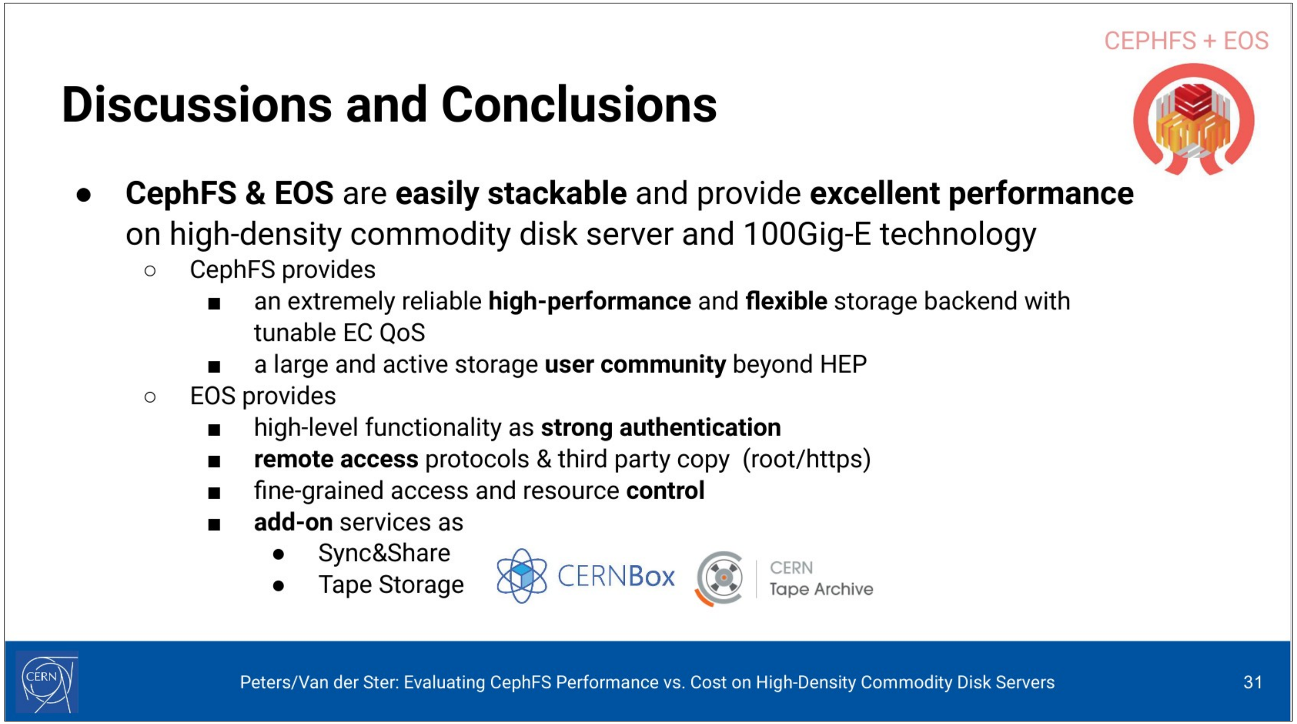
Final result

Delegated Storage | Tigran Mkrtchyan | 6/2/17 | Page 4

By Dan Van Der Ster, CERN

CEPH as Storage Building Block?

- CephFS looks a promising solution for storage
- Missing functionality covered by EOS and dCache
- On site expertise is required. (How many Dans are around?)
- Can we coordinate the effort?


CEPHFS + EOS

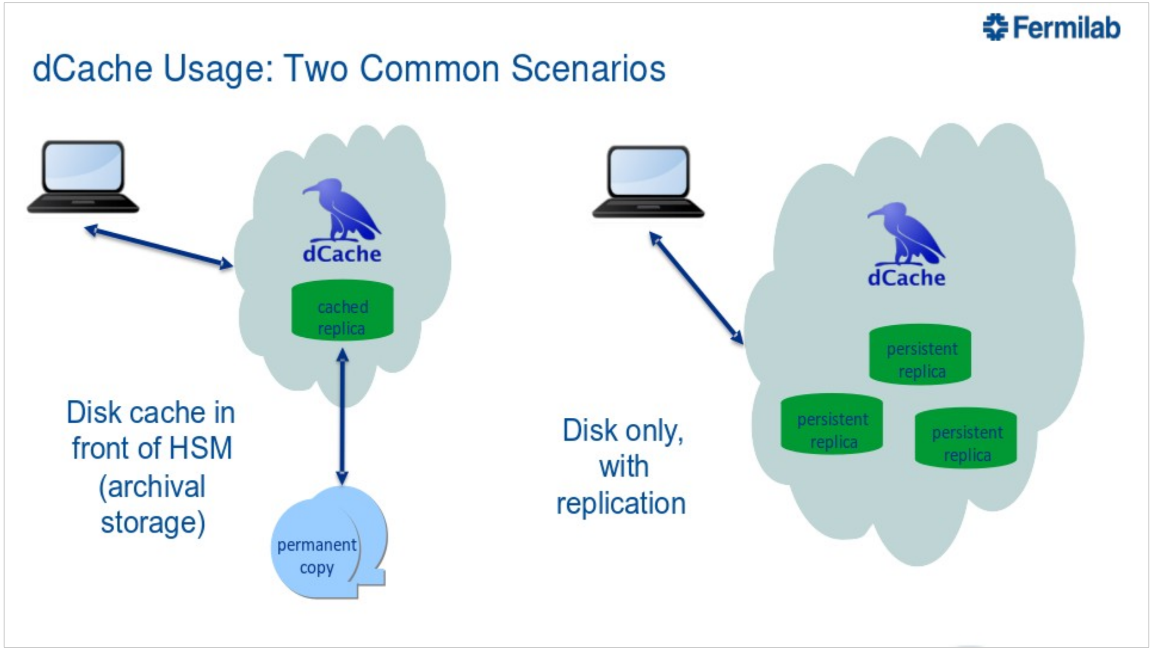
Discussions and Conclusions

- **CephFS & EOS** are **easily stackable** and provide **excellent performance** on high-density commodity disk server and 100Gig-E technology
  - CephFS provides
    - an extremely reliable **high-performance** and **flexible** storage backend with tunable EC QoS
    - a large and active storage **user community** beyond HEP
  - EOS provides
    - high-level functionality as **strong authentication**
    - **remote access** protocols & third party copy (root/https)
    - fine-grained access and resource **control**
    - **add-on** services as
      - Sync&Share
      - Tape Storage

Peters/Van der Ster: Evaluating CephFS Performance vs. Cost on High-Density Commodity Disk Servers 31

QoS - Technology & Policy

- Availability
- Durability
- Access latency



Fermilab

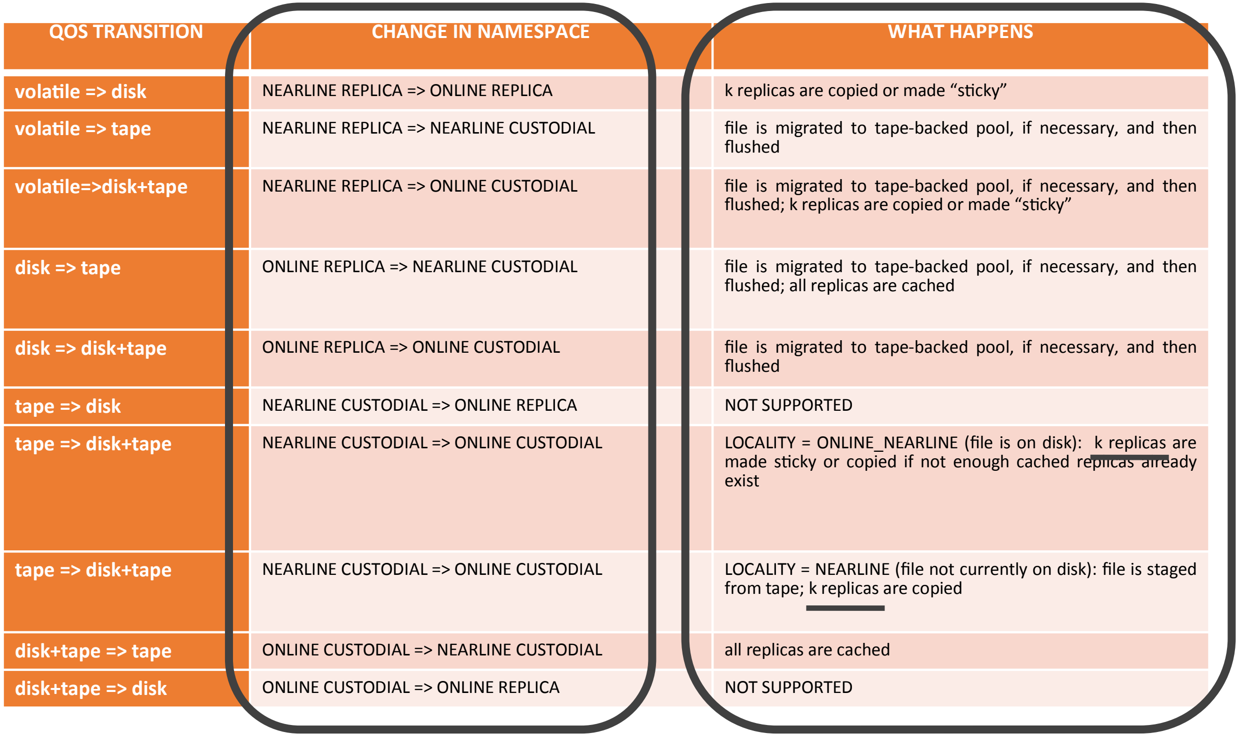
dCache Usage: Two Common Scenarios

Disk cache in front of HSM (archival storage)

Disk only, with replication

QoS Rule Engine Prototype

Uses the current combination (from Resilience) of namespace attributes (**Access Latency** and **Retention Policy**) plus membership in a storage group (**storage unit**) expressing the number and distribution of disk replicas, to define a set of very basic QoS classes.



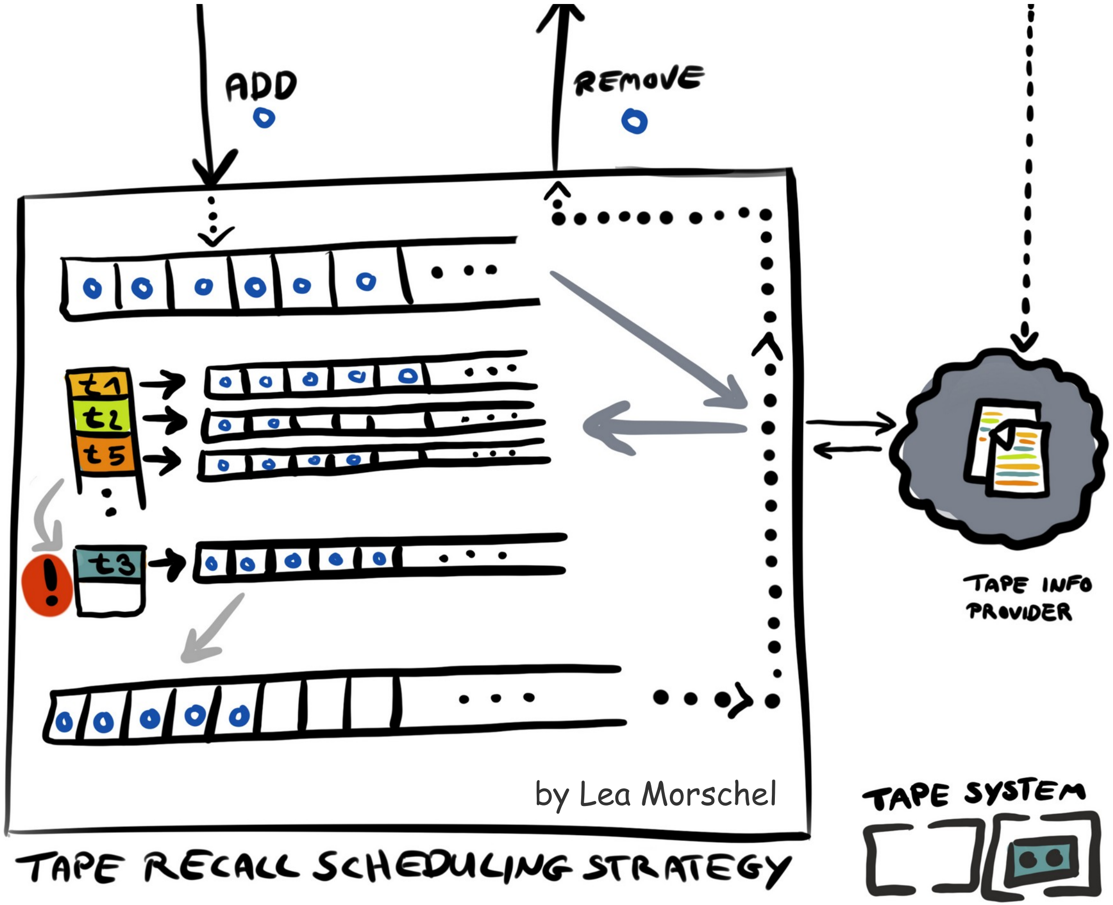
QOS TRANSITION	CHANGE IN NAMESPACE	WHAT HAPPENS
volatile => disk	NEARLINE REPLICA => ONLINE REPLICA	k replicas are copied or made "sticky"
volatile => tape	NEARLINE REPLICA => NEARLINE CUSTODIAL	file is migrated to tape-backed pool, if necessary, and then flushed
volatile=>disk+tape	NEARLINE REPLICA => ONLINE CUSTODIAL	file is migrated to tape-backed pool, if necessary, and then flushed; k replicas are copied or made "sticky"
disk => tape	ONLINE REPLICA => NEARLINE CUSTODIAL	file is migrated to tape-backed pool, if necessary, and then flushed; all replicas are cached
disk => disk+tape	ONLINE REPLICA => ONLINE CUSTODIAL	file is migrated to tape-backed pool, if necessary, and then flushed
tape => disk	NEARLINE CUSTODIAL => ONLINE REPLICA	NOT SUPPORTED
tape => disk+tape	NEARLINE CUSTODIAL => ONLINE CUSTODIAL	LOCALITY = ONLINE_NEARLINE (file is on disk): k replicas are made sticky or copied if not enough cached replicas already exist
tape => disk+tape	NEARLINE CUSTODIAL => ONLINE CUSTODIAL	LOCALITY = NEARLINE (file not currently on disk): file is staged from tape; k replicas are copied
disk+tape => tape	ONLINE CUSTODIAL => NEARLINE CUSTODIAL	all replicas are cached
disk+tape => disk	ONLINE CUSTODIAL => ONLINE REPLICA	NOT SUPPORTED

By Albert Rossi, Fermilab

Tape Recall Grouping

- Group requests by tape
- Recall triggered by
  - Size
  - Max idle time
- Number of parallel recall based on number of tape drives

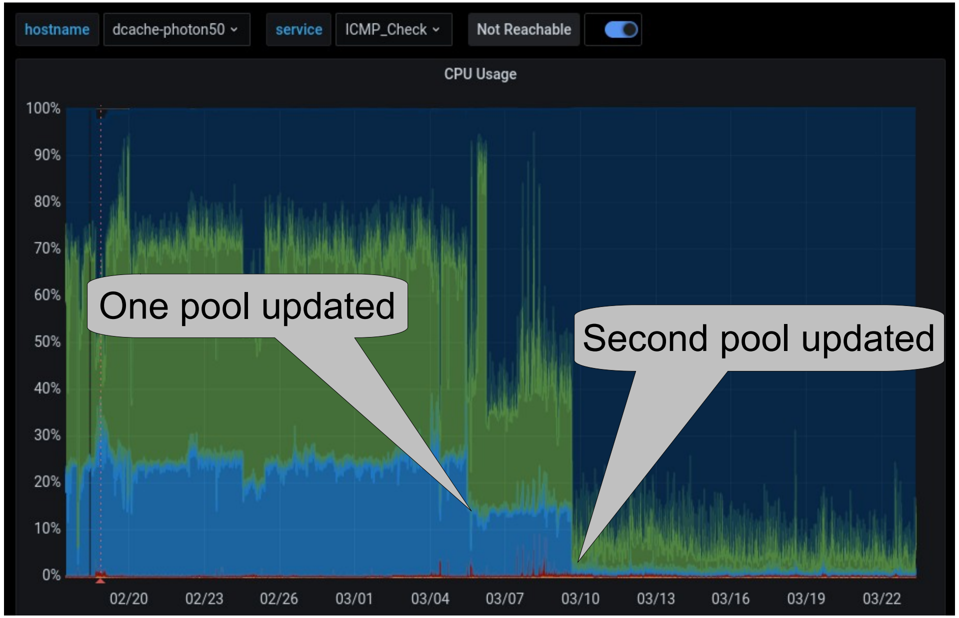
By Lea Morschel, DESY



Sapphire (small file plugin)

By Svenja Meyer, DESY

- Evolution of *Small-file-plugin*
  - Addresses discovered limitations
- In-dCache HSM driver
  - Full access to metadata
  - No external script
  - Stateful
- Better resource utilization



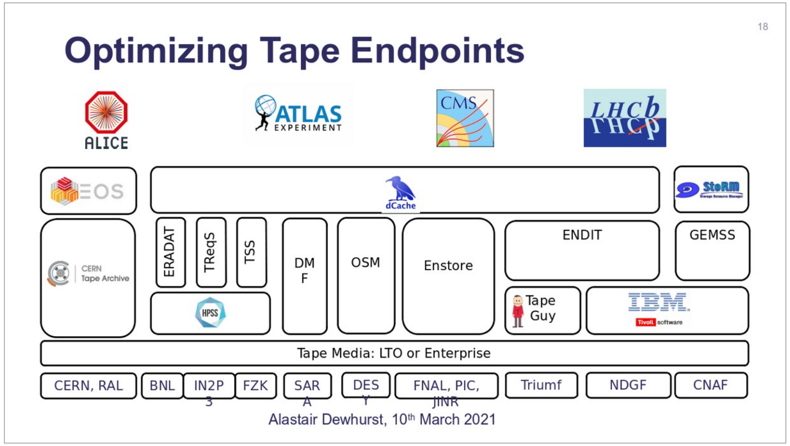
hostname: dcache-photon50

service: ICMP_Check

One pool updated

Second pool updated

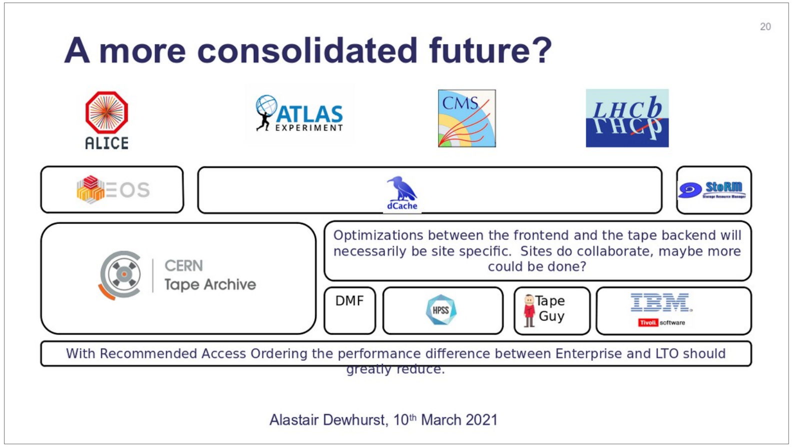
dCache ⟺ CTA Integration



18

Optimizing Tape Endpoints

Alastair Dewhurst, 10th March 2021



20

A more consolidated future?

Optimizations between the frontend and the tape backend will necessarily be site specific. Sites do collaborate, maybe more could be done?

With Recommended Access Ordering the performance difference between Enterprise and LTO should greatly reduce.

Alastair Dewhurst, 10th March 2021

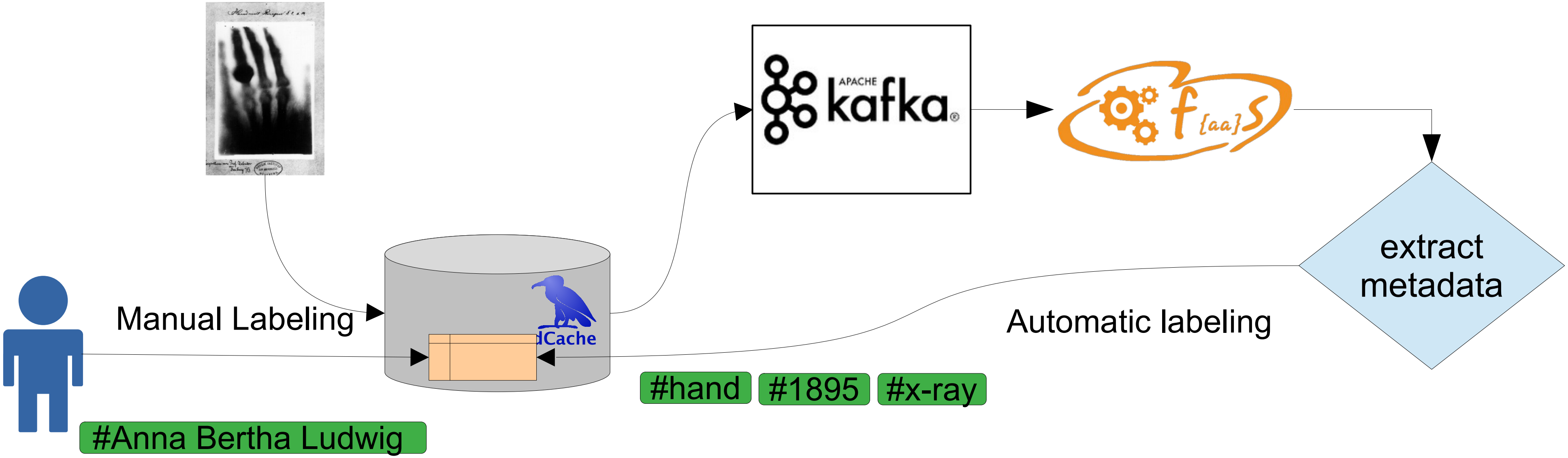
Pros

- CERN Product
- GPL3
- Well defined software development process
  - CI replicated at DESY
- Test setup at DESY with Virtual Tape Library

Cons

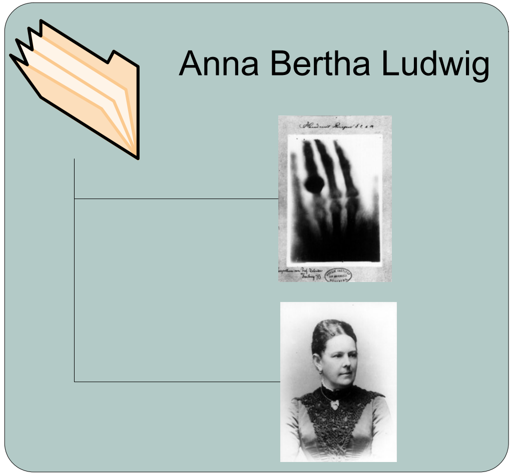
- CERN Product
- In *early production* stage
- Orthogonal to dCache *tape awareness*
- Non-standard access protocol
- Non-standard on tape format

Metadata Population

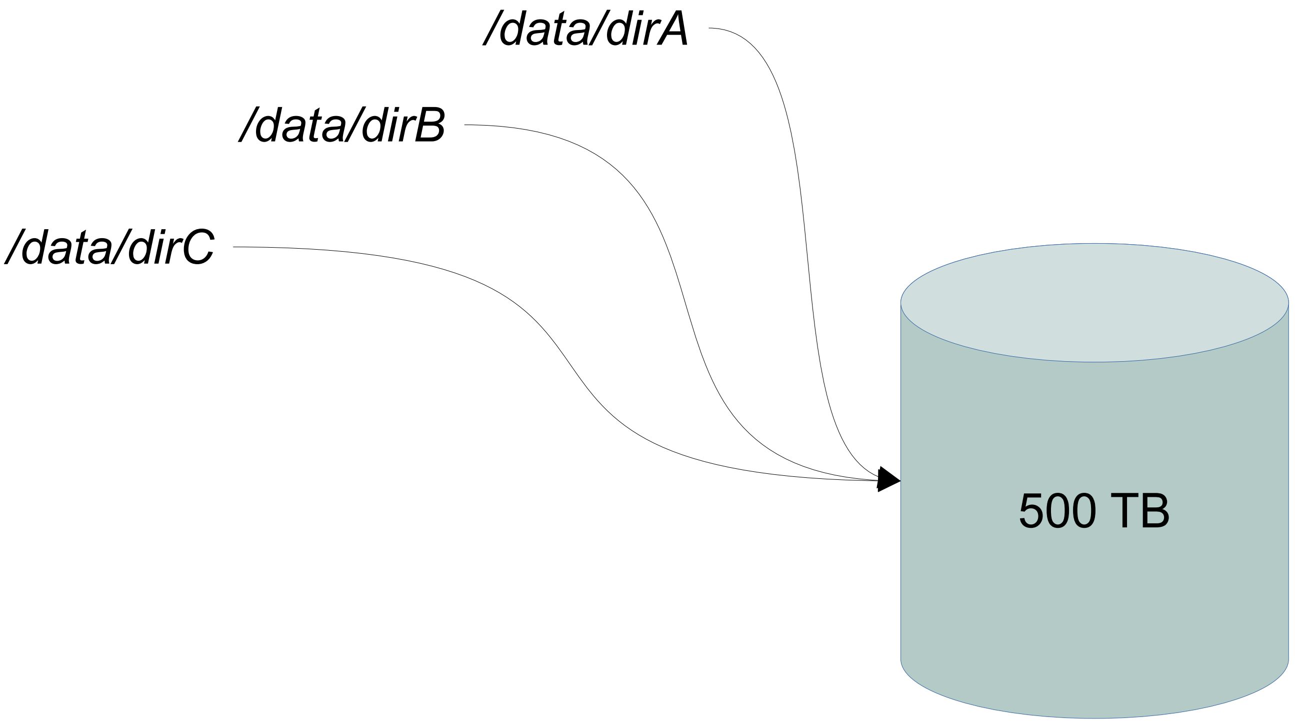


User Metadata/Labeling in dCache

- Extended attributes
  - Exposed via NFS, WebDAV, REST
- Label-based virtual **read-only** directories (WIP)
  - List all files with a given label
- dCache rules applies
  - Visible through all protocols
  - Respect file/dir permissions



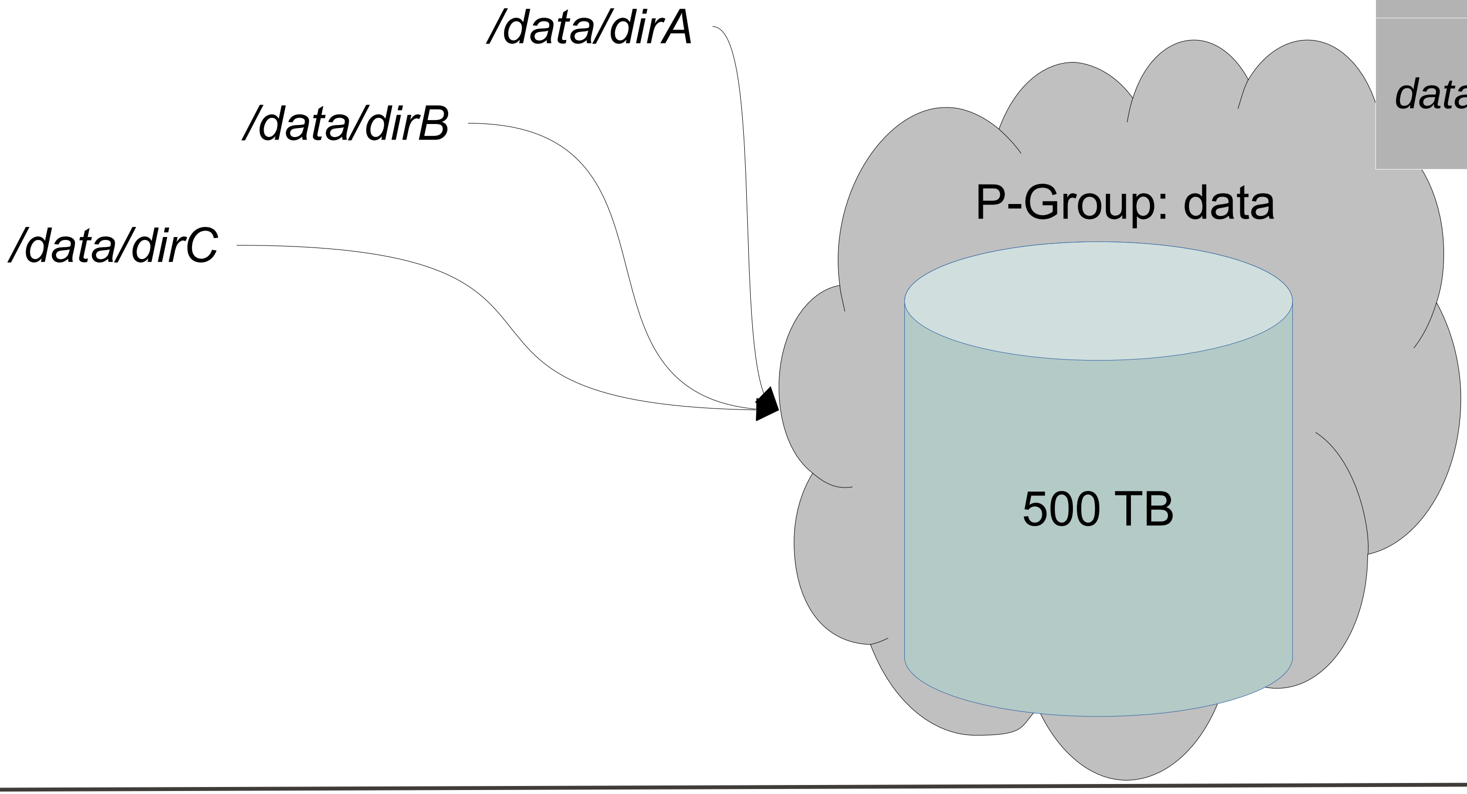
SRR Problem Statement



Directory	Available space
/data/dirA	500 TB
/data/dirB	500 TB
/data/dirC	500 TB
Total:	1.5 PB

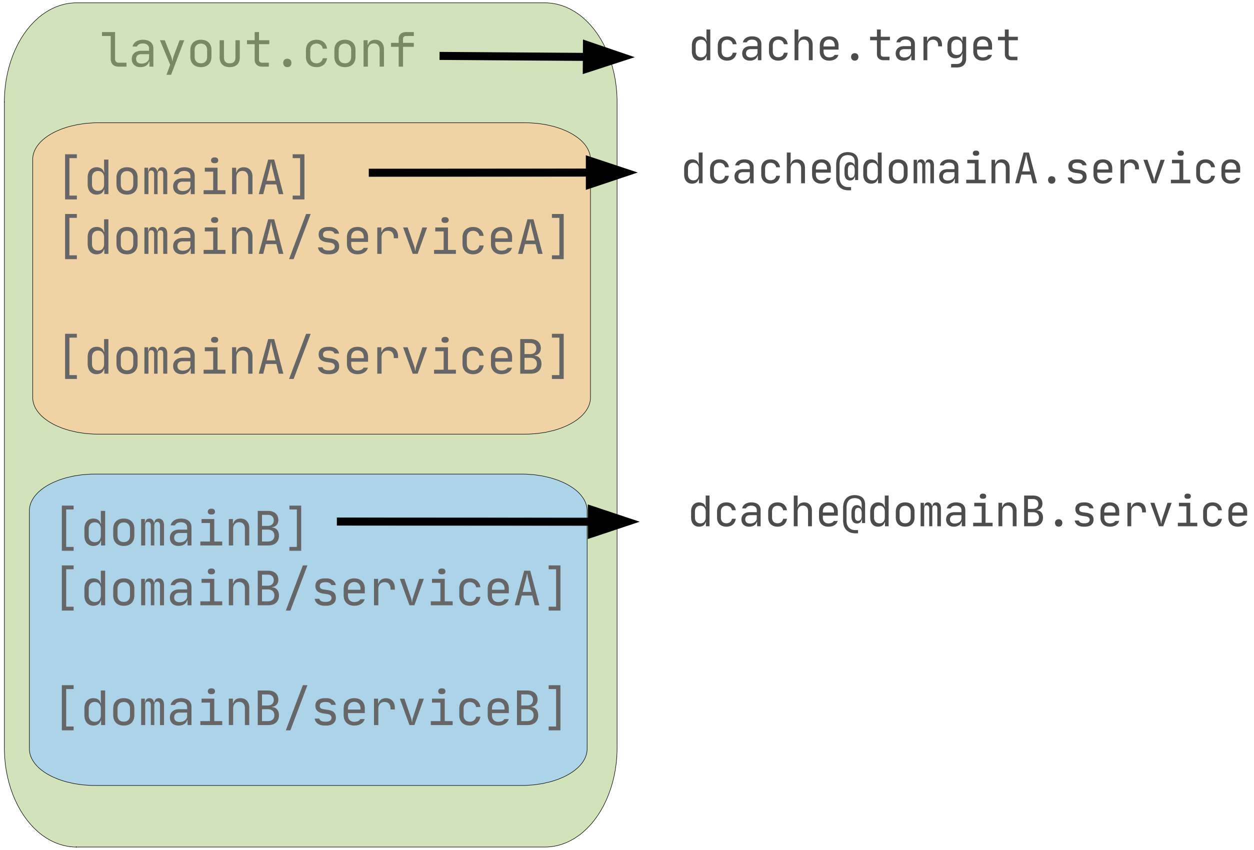
SRR Solution(?)

Share	Available space
data	500 TB



systemd – The Breaking change!

- dCache-6.2 is systemd only
  - Some like it, others hate it
- Additional mini workshop back in November 2020



layout.conf → dcache.target

[domainA] → dcache@domainA.service

[domainA/serviceA]

[domainA/serviceB]

[domainB] → dcache@domainB.service

[domainB/serviceA]

[domainB/serviceB]

Summaries

Workshop Topics

- Site operation
  - Ease of installation
  - Monitoring
  - HW utilization/efficiency
- Integration with other services
  - iRODS
  - Globus-Online
- Long term data archival
  - Tape access optimization
  - Small file aggregation
  - CTA

Workshop organization

- Video workshops are well received
  - Positive feedback
  - Larger audience
  - Lots of spontaneous discussions
- Positive experience with mini-workshops on selected topics
- Hands on sessions still required
  - Do we have such experience?

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