dCache News, Status & Roadmap

0xF International dCache workshop

DESY. Fermilab neic NORDIC E-INFRASTRUCTURE COLLABORATION

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Workshop outline

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.

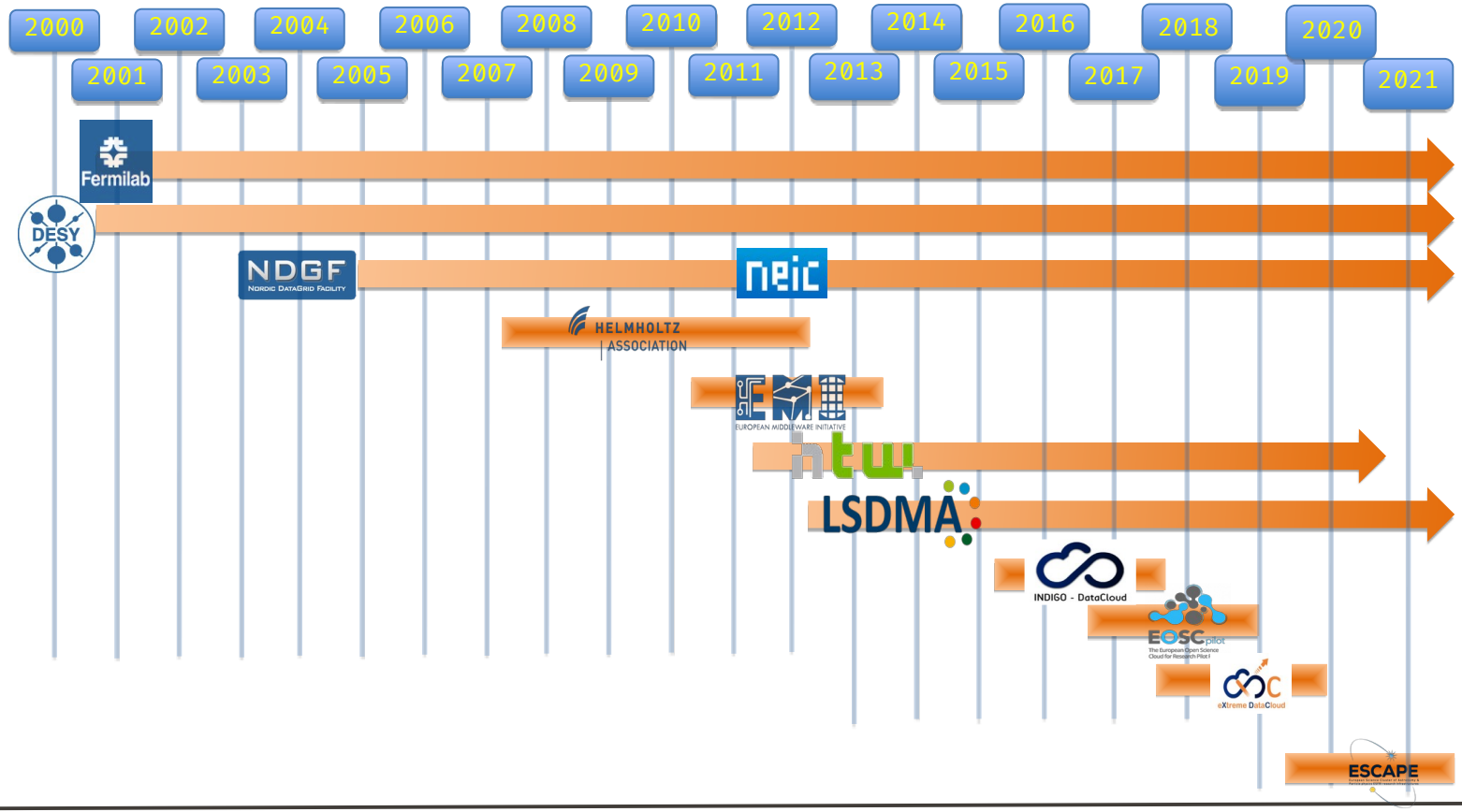
Wednesday, 2 June

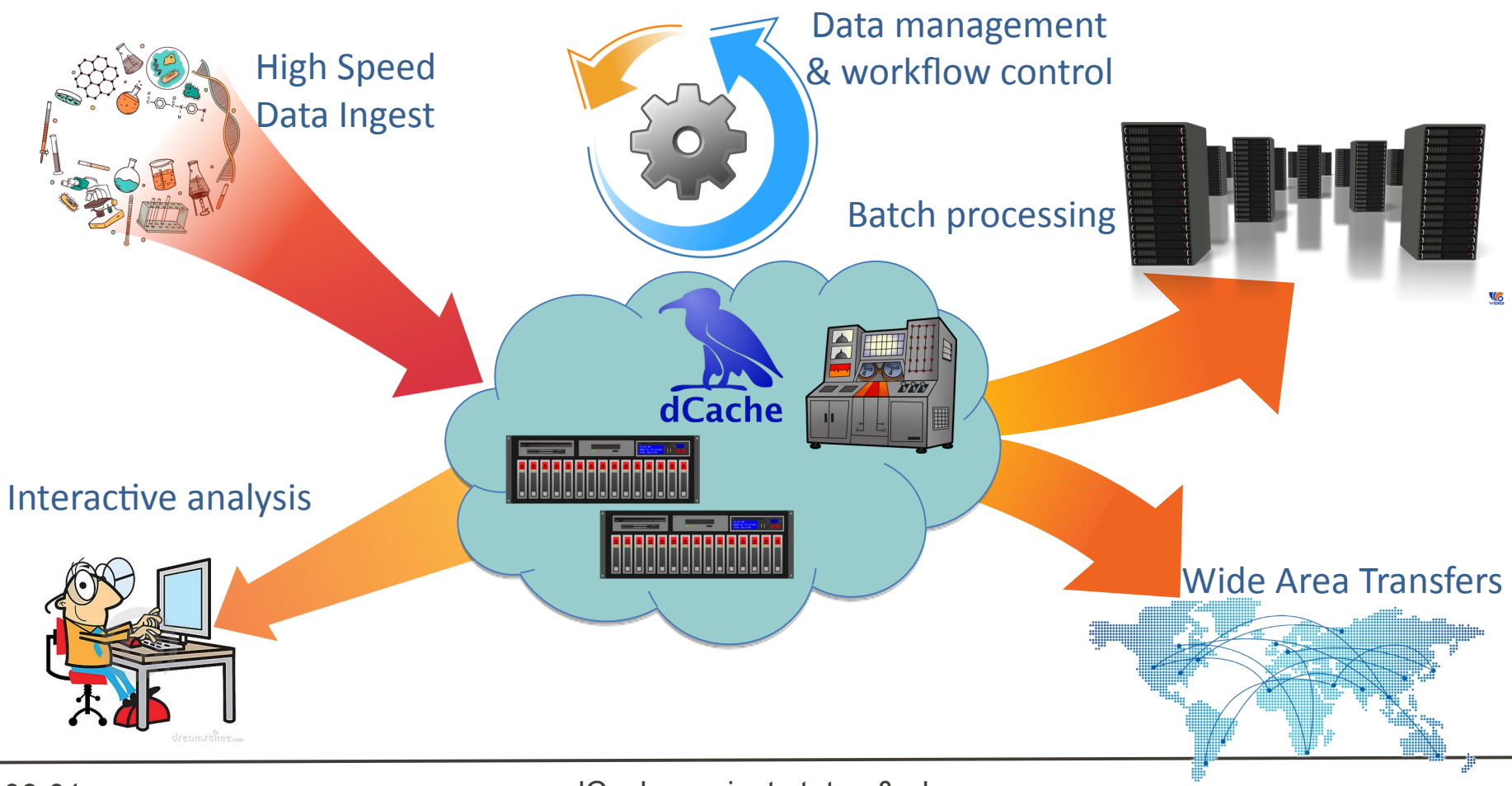
4:00 PM → 6:00 PM **Experience from sites**

- 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.

Project Funding & Team

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




High Speed
Data Ingest
Data management
& workflow control
Batch processing
dCache
Interactive analysis
Wide Area Transfers

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

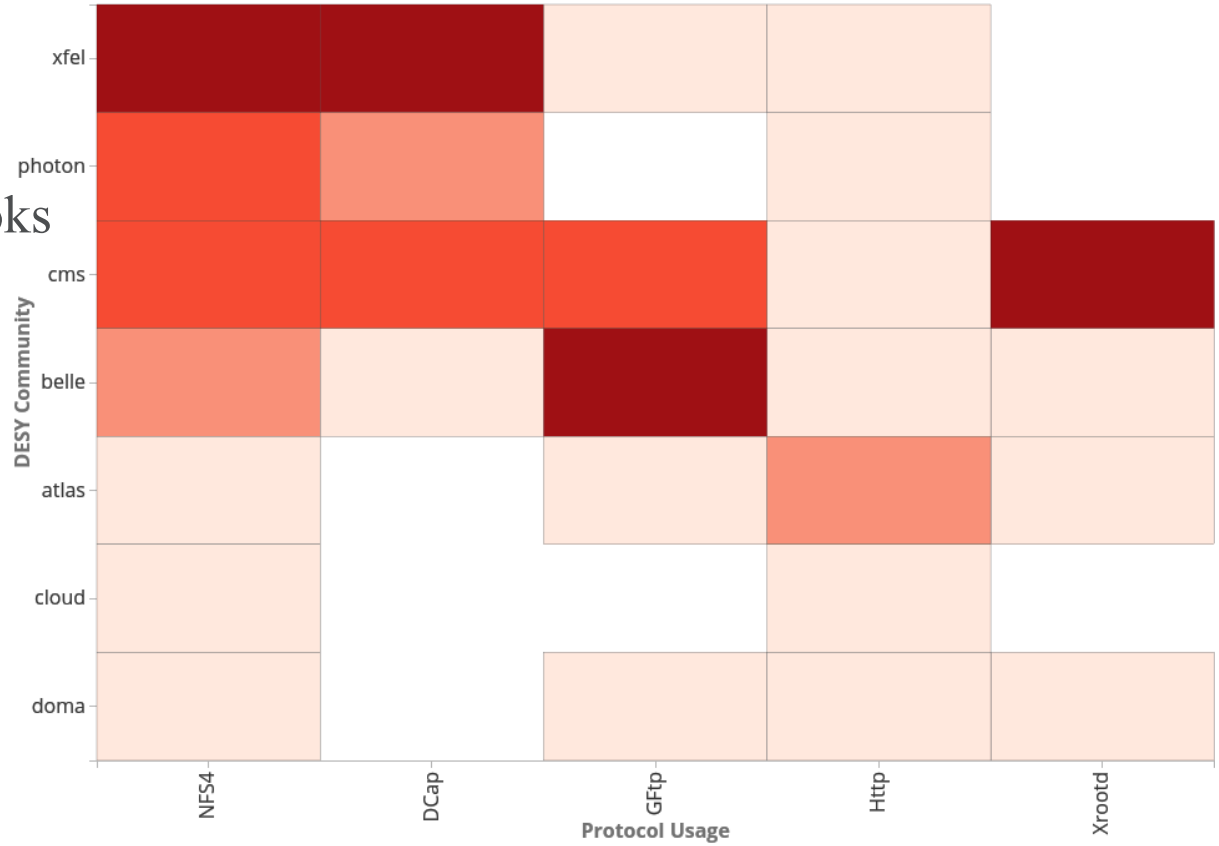
- High Reliability
- Self-healing
- Automatic technology migration
- Persistent identifier

Strategic Communities

- WLCG/HEP
  - Almost all sites run dCache for WLCG
- EuXFEL/Neutrino Physics/Photon Science
  - On-site stakeholders
- Astro- Bio- Life-science
  - Almost all sites support non-WLCG communities

Data Access Variety

- 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



Deployment status

- 80% of sites run dcache-5.2.x
  - **Note: the support ends by the end of THIS summer!**
- 6.2.x get slowly adopted
- 7.1 is on pre-prod instances!

Breaking Changes …

Java-11

- Recommended version to use
  - We test with OpenJDK, other should work as well
- Required starting 6.2
- Some errors to ignore:

  WARNING: An illegal reflective access operation ...

  WARNING: Illegal reflective access by ...

  WARNING: Please consider reporting this ...

  WARNING: Use --illegal-access=warn to ...

Java Flight Recorder

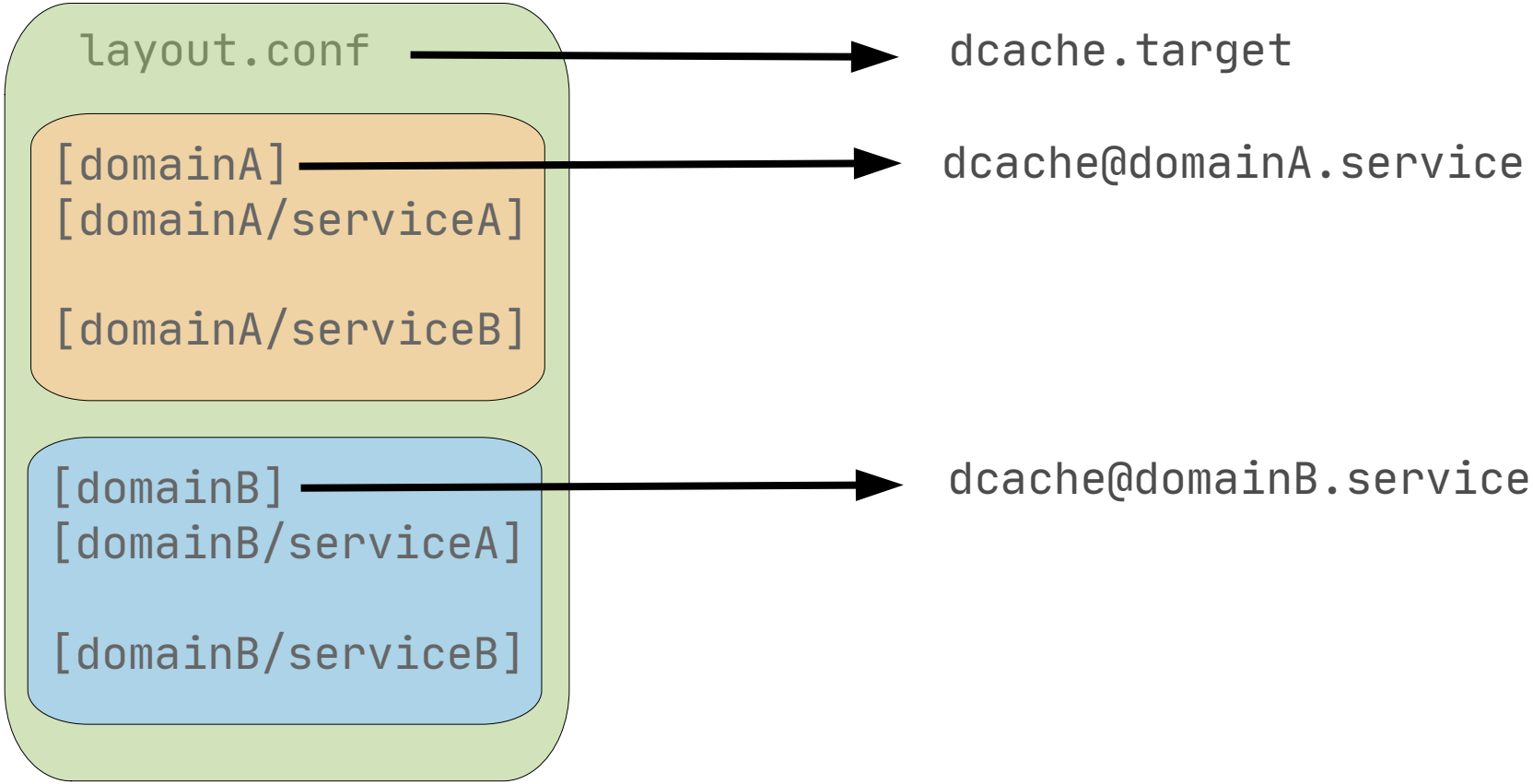
- Available in Java-11
- Allows to collect information from running JVM
  - Requires JMX or extra *enable* option (< 7.2):
    (-XX:+StartAttachListener)

```
$ jcmd <pid> JFR.start duration=60s \
                    filename=/tmp/dcache.jfr
```

Dropped ...

- Dropped pcells GUI support
  - Please use the admin interface in dcache-view
- Sys-V -like *service* files
  - Please consider to switch to *systemd*

systemd Integration



systemd Service Unit Generator

- Generates service units from *layout.conf*
- **MUST** be executed when:
  - Domains added/removed
  - Java options have changed
  - User is change
- Runs automatically on reboot (since Apr 28)

More details: https://indico.desy.de/event/28064/

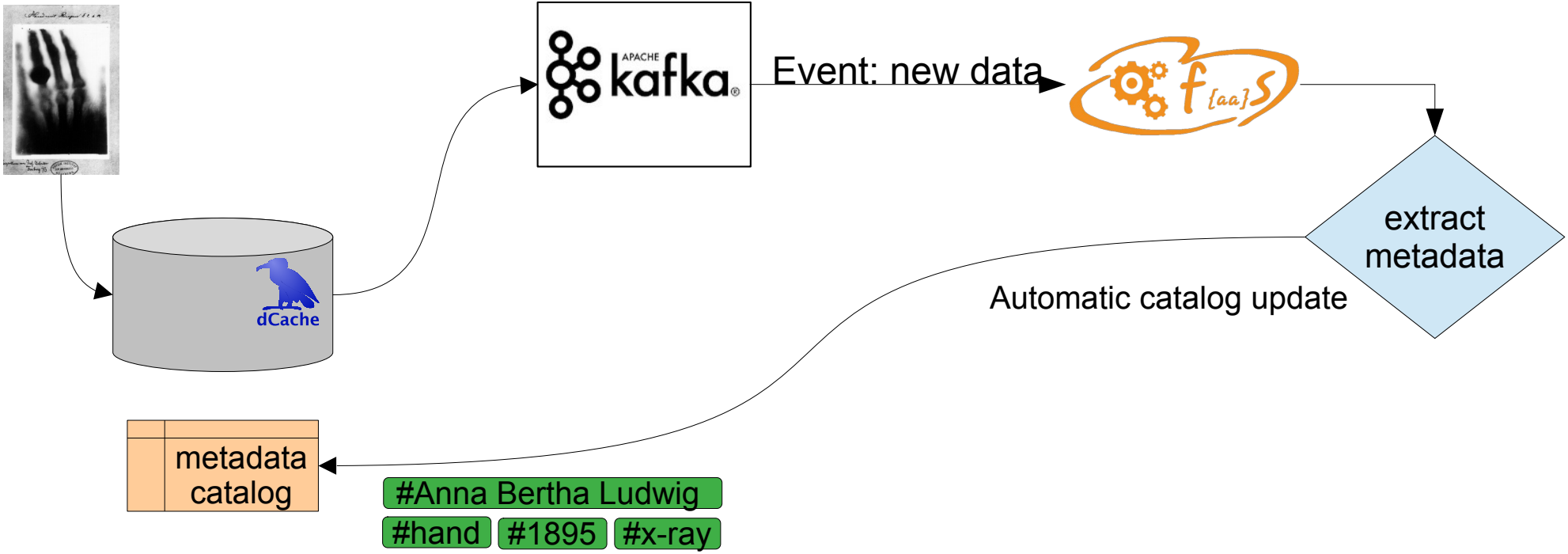
Metadata

User Metadata Handling

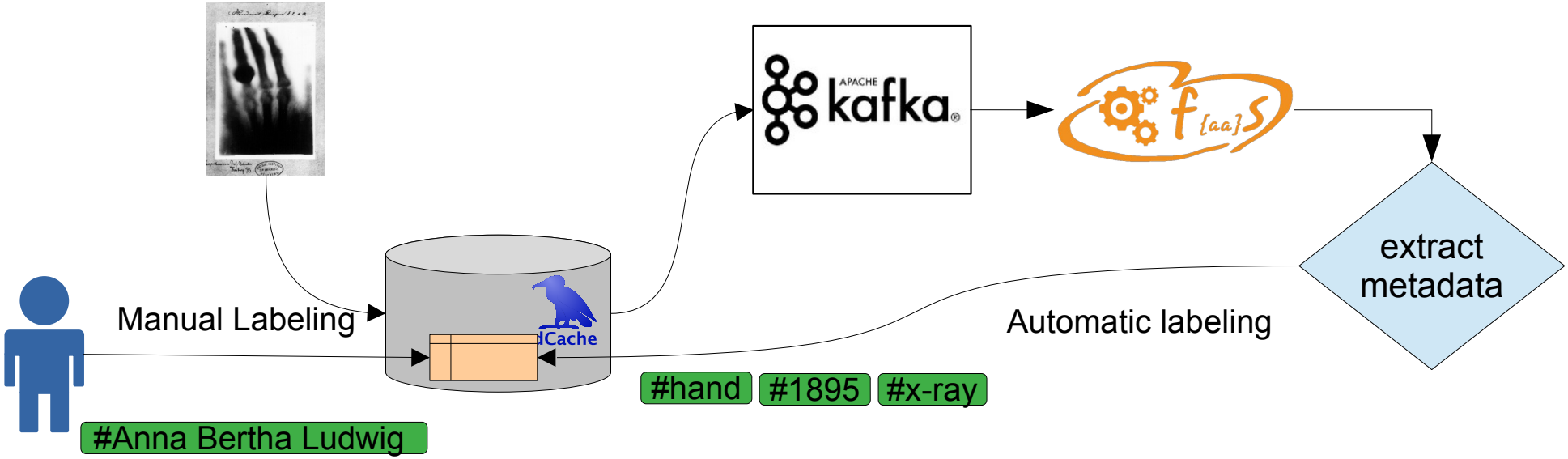
- User metadata important (again)
  - Data labeling/classification
- Can be populated by storage events
  - Some automation is required

#Anna Bertha Ludwig
#hand #1895 #x-ray

Automatic Metadata Population



Metadata Population



User Attribute Support

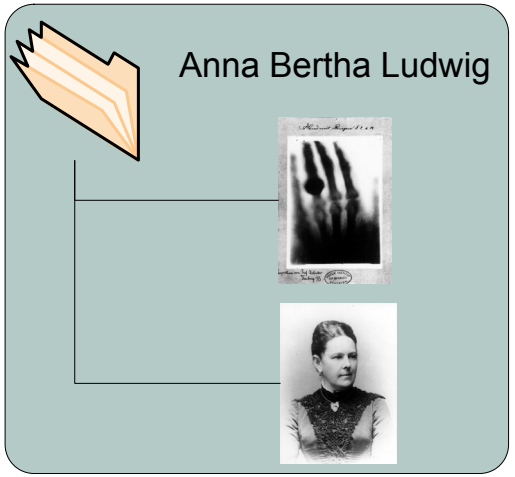
- HTTP(s)
  - As query option on upload
  - Those attributes are available to the flush process!
- ~~POSIX~~ xattrs
  - {get/set}fattr over NFS
  - Exposes directory tags
- File *tagging/labeling*

Example:

```
$ curl –upload-file file.txt \
“https://localhost:2881/data/file.txt?xattr.key1=value1”
$ getfattr -L file.txt
# file: file.txt
user.key1
$ getfattr -n user.key1 file.txt
# file: file.txt
user.key1="value1"
```

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



HSM & QoS

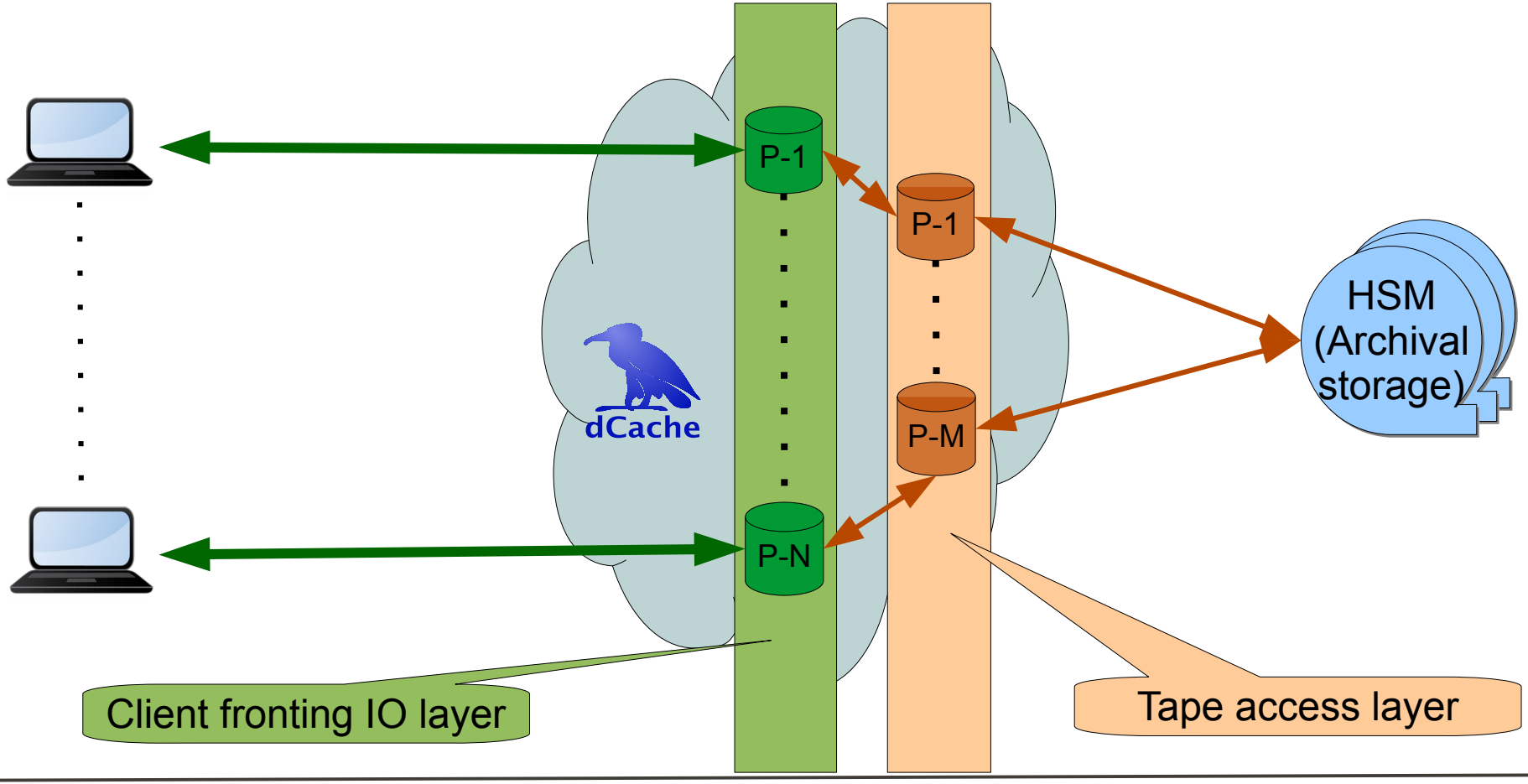
HSM, Tape, QoS

- ATLAS “Tape carousel” ⇒ WLCG “Data carousel”
  - ~~Bad habits die hard~~ Share the best practices
- High data volumes by EuXFEL
  - ~1PB/week
- High number of small files by Photon Science
  - ~4MB, 10^{6} files per directory
- Multi-media copy guarantees

Three Directions to Address

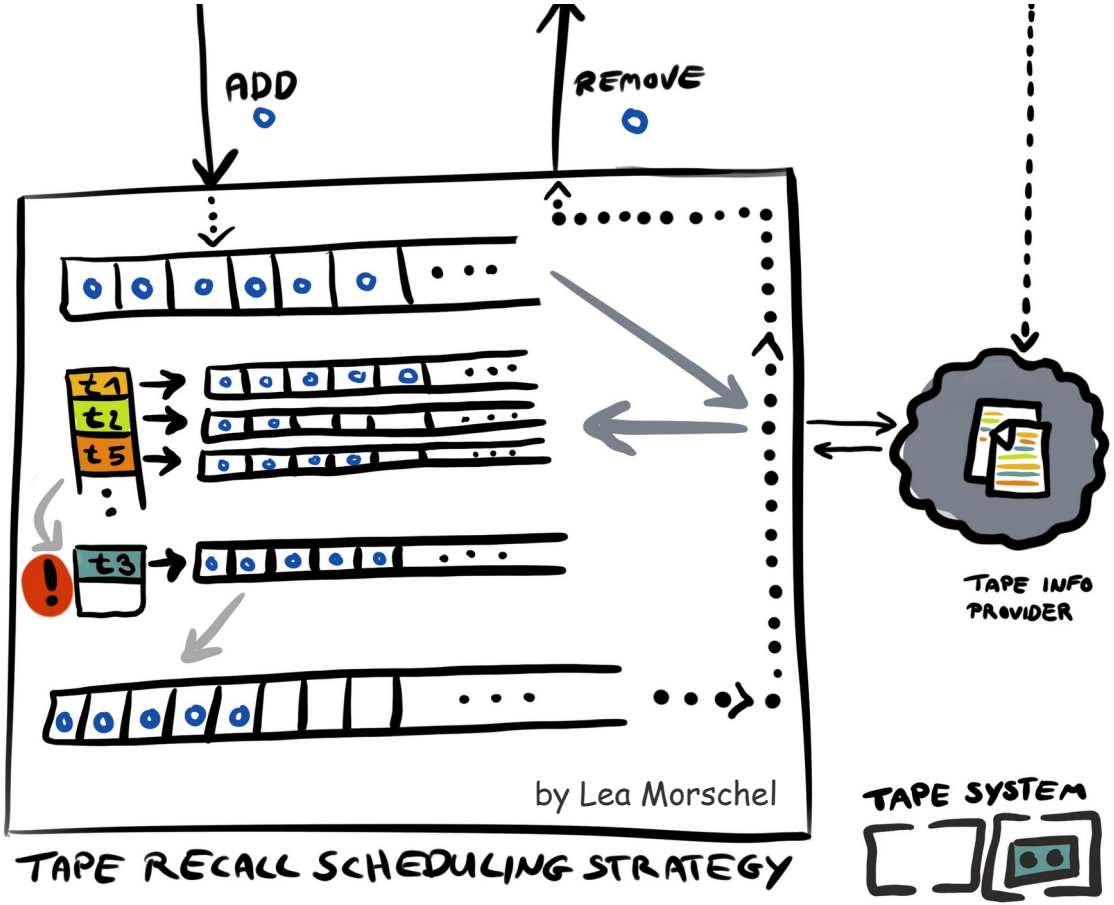
- Better HW split on tape/disk pools
  - Some nodes can be optimized for tape access only
  - A-la QoS for hardware
- Tape recall grouping by tape
  - Collect request for a single tape
  - Prototype in SRM
- **Sapphire** - small file aggregation
  - dCache native HSM driver

Layered Pools Model



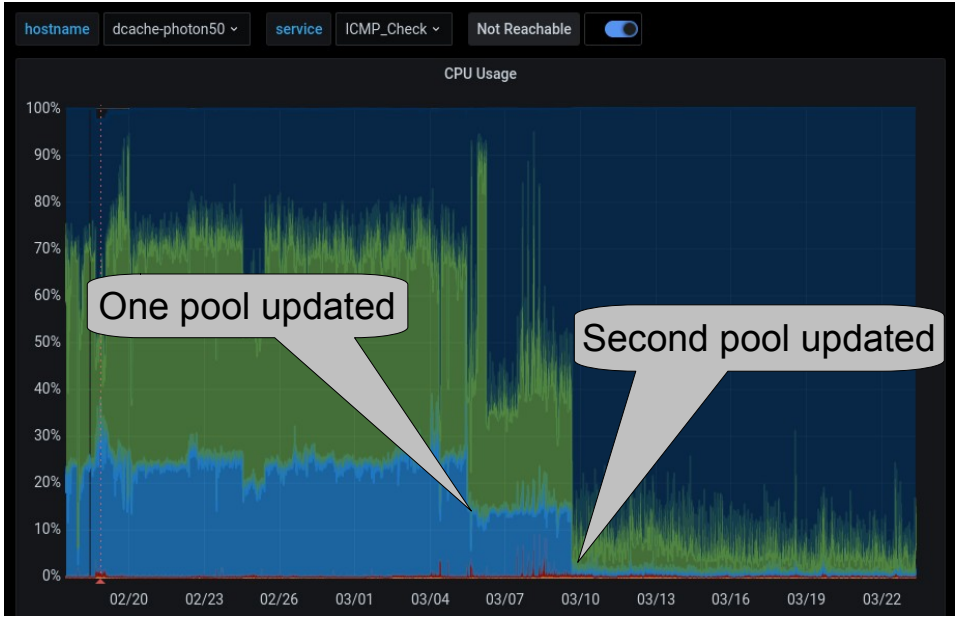
Tape recall grouping

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

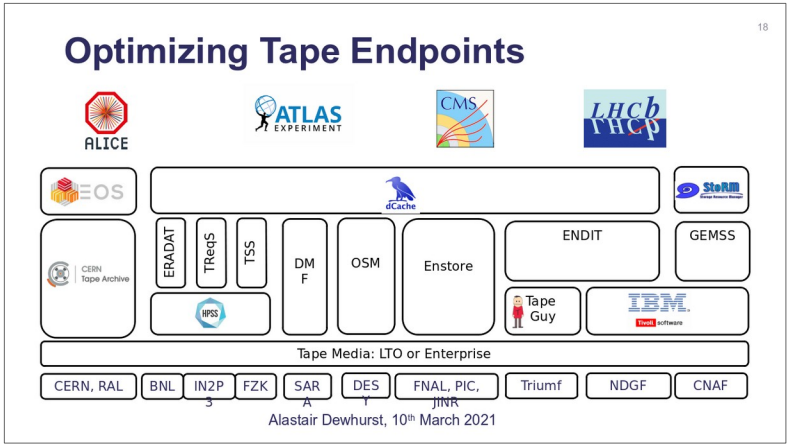


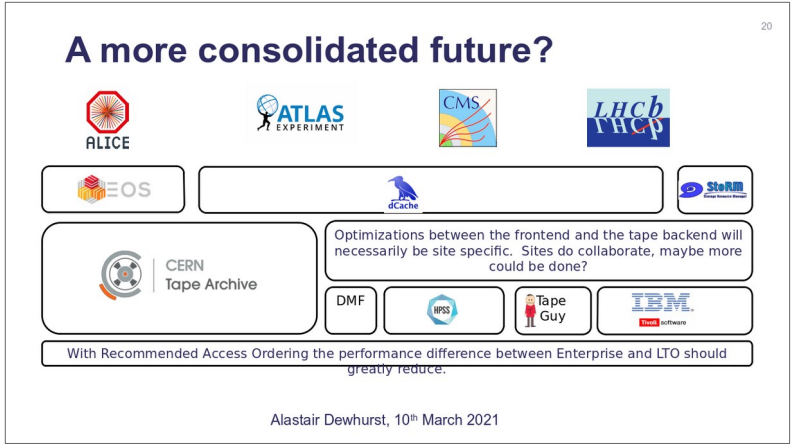
Sapphire (small file plugin)

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



dCache ⟺ CTA Integration





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

v1 Bulk REST-API (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

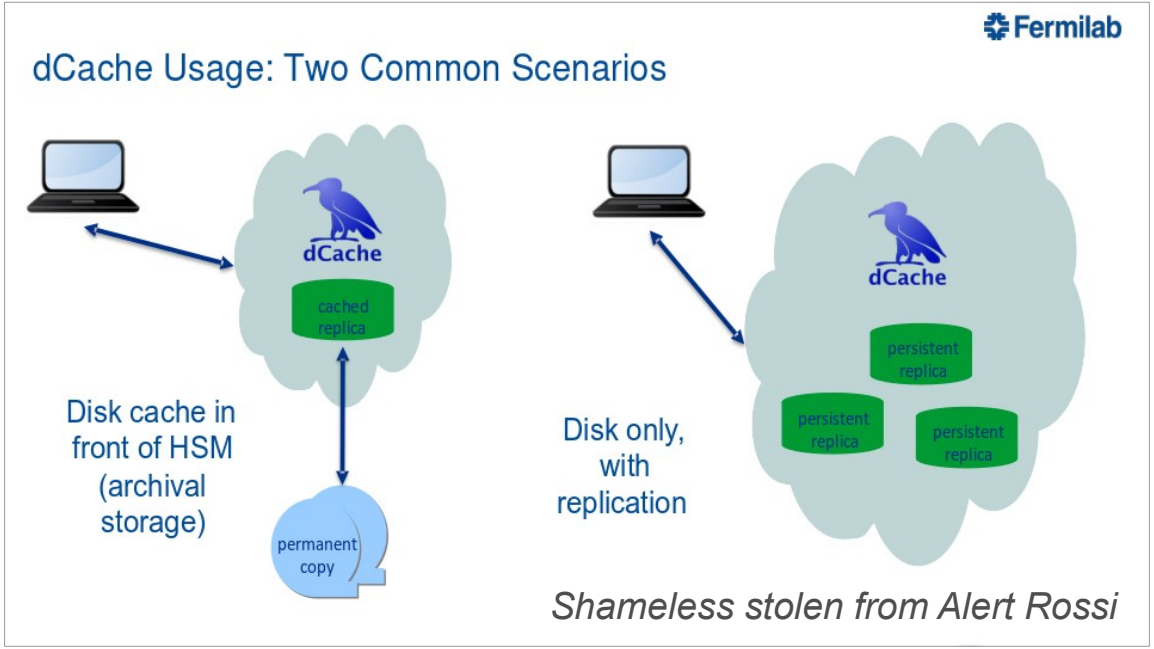
- Pin cached copies with a lifetime

FILEINFO

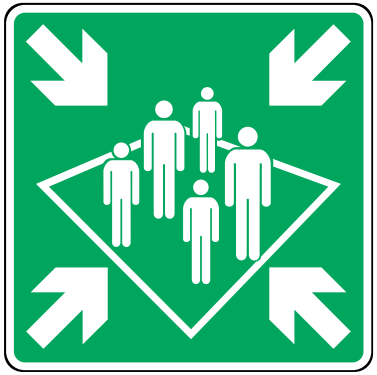
- Request status many files at once (locality, checksum)

QoS - Technology & Policy

- Availability
- Durability
- Access latency



Tokens
&
IdP Federation



Rule #1

USE

TLS

for token-based authentication

Keycloak Integration

- Open source SSO & IAM solution
  - Generates *identity* and *access* tokens.
  - Popular at many-sites including DESY
  - Delegated group membership management
- Supports standards-based OpenID-connect, OAuth2 and LDAP
  - ***preferred_username*** to map to LDAP accounts
  - ***eduPersonEntitlement*** mapped to group membership
- dCache's gPlazma config can be combined with X509, VOMS and others
  - Some code changes are needed to improve integration

Keycloak Integration

PaNOSC

Laser-Driven Proton Acceleration from Cryogenic Hydrogen Target

Description
2D particle-in-cell simulation of the interaction of high-intensity laser pulse (parameters are relevant to L4 laser) with a cryogenic hydrogen target. Only protons with energy above 300 MeV at the end of the simulation are tracked and their position and energy are visualized. Two different groups of protons accelerated by different mechanisms can be distinguished from each other in space: Protons originated from the target interior and from the target rear side.

Citation	Dana Scully; (2020), Re-polarization of the aft quantum plasma collector, DOI:10.9563/if.2015.87.012
Keywords	X-ray excited optical luminescence,
Type	Proposal
Author	Laima Reinhold
Other	Stuff

Stolen from Michael Schuh

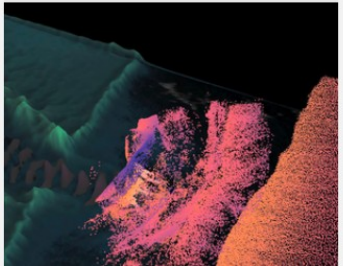
Datasets

PaNOSC Test Dataset 11

HEIMDAL @ ESS

Name
my-environment
Descritption
re-produce research
Flavour
jupyter_small
Spawn

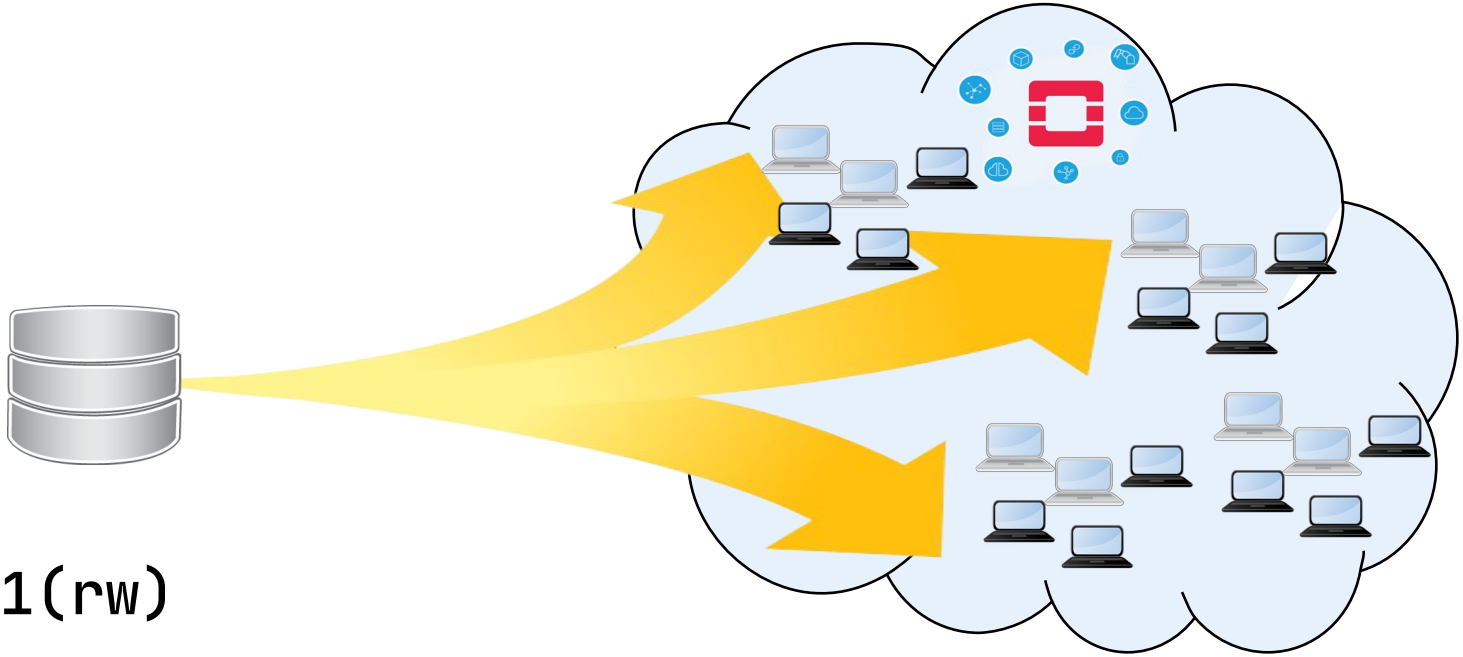
Preview Visualization



The issues with NAS as cloud share

"IP-based access control allows you to control access to Filestore Instances based on the IP address of clients."

https://cloud.google.com/filestore/docs



```
# /etc/export
/data 10.1.0.1(rw)
```

Interface to Manage Shares

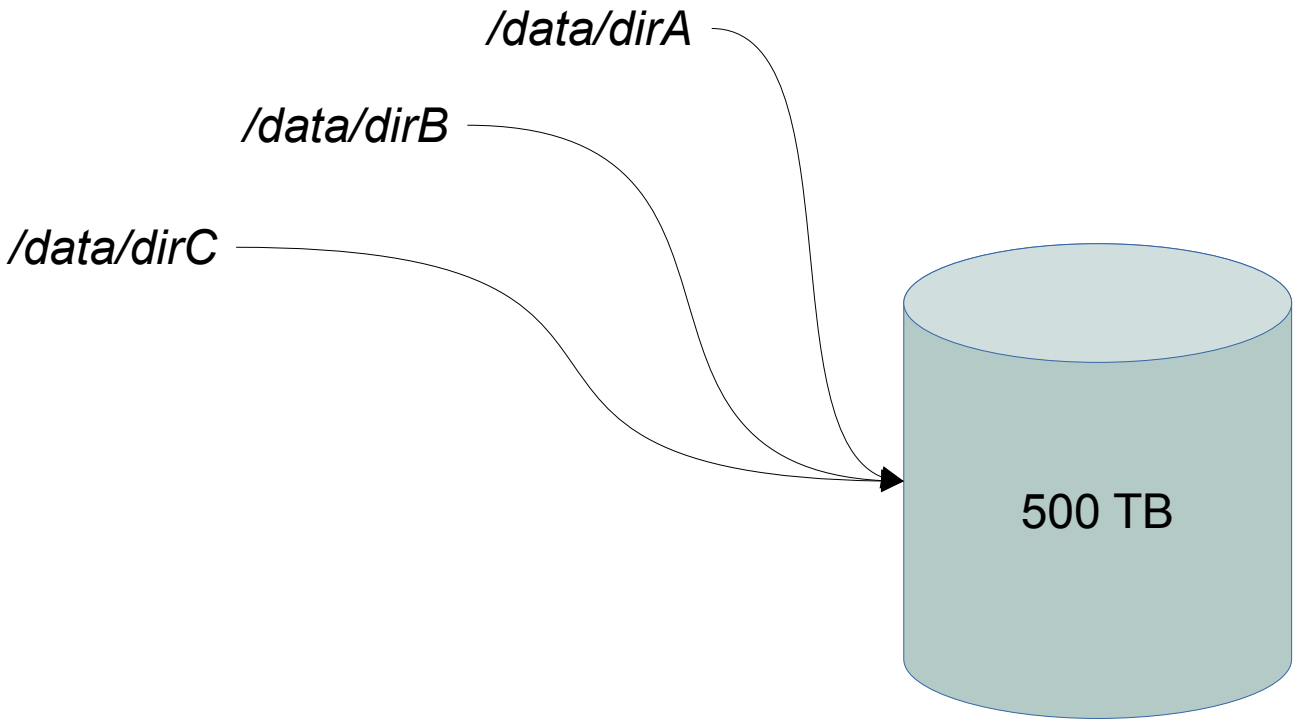
- REST-API for share management
  - Integration into projects workflow or portal
  - Simple API to manage export table
  - Compatible with OpenStack Manila
  - Same building blocks as dCache REST interface
  - All dCache supported authN & authZ *for free*

Storage Resource Reporting (aka SRR)

SRR Problem Statement

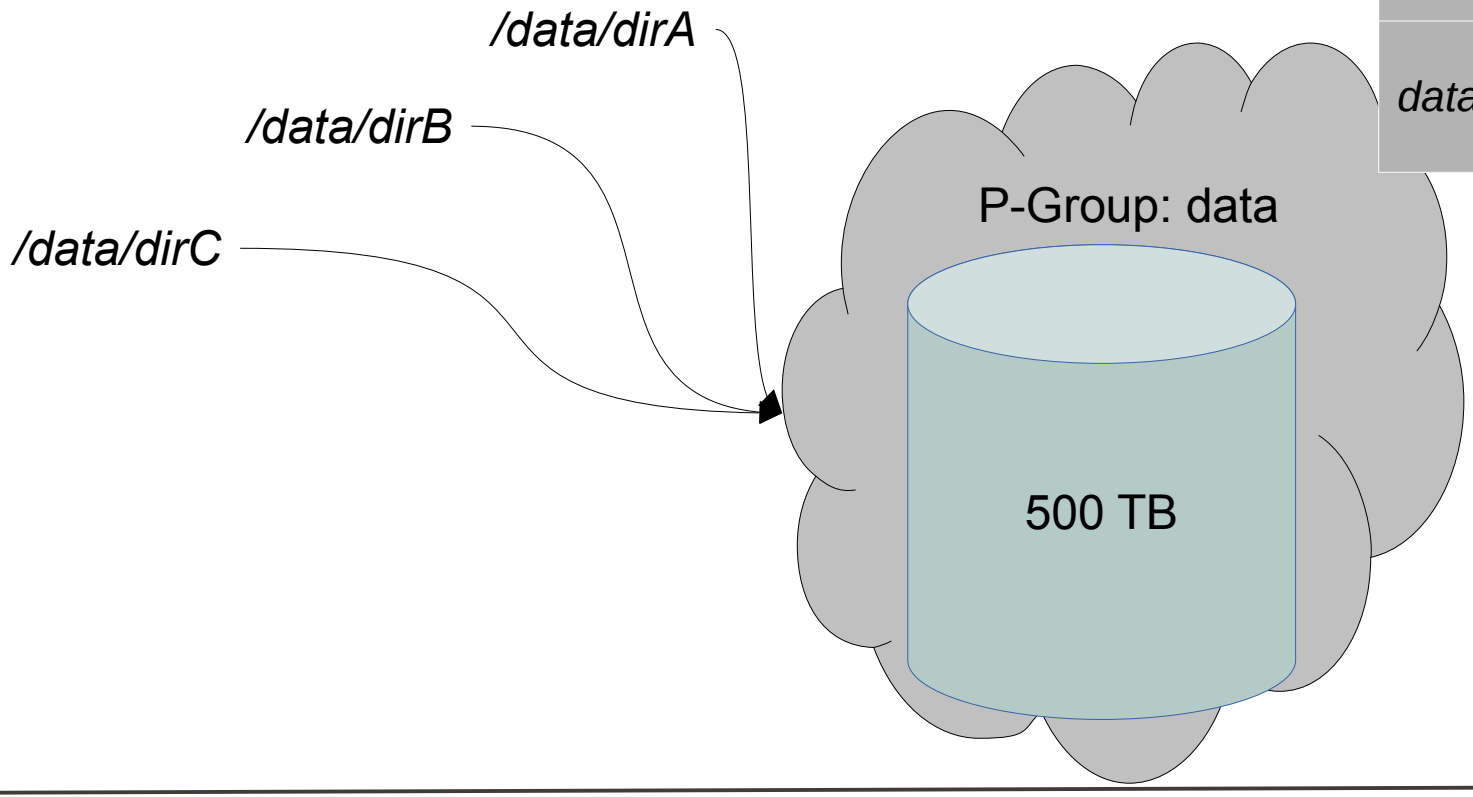
```
"storageshares" : [
  {
    "name" : "dirA",
    "path" : [ "/data/dirA" ],
    "totalsize" : 500TB,
    "usedsize" : 0,
    "vos" : [ "foo" ]
  }
```

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

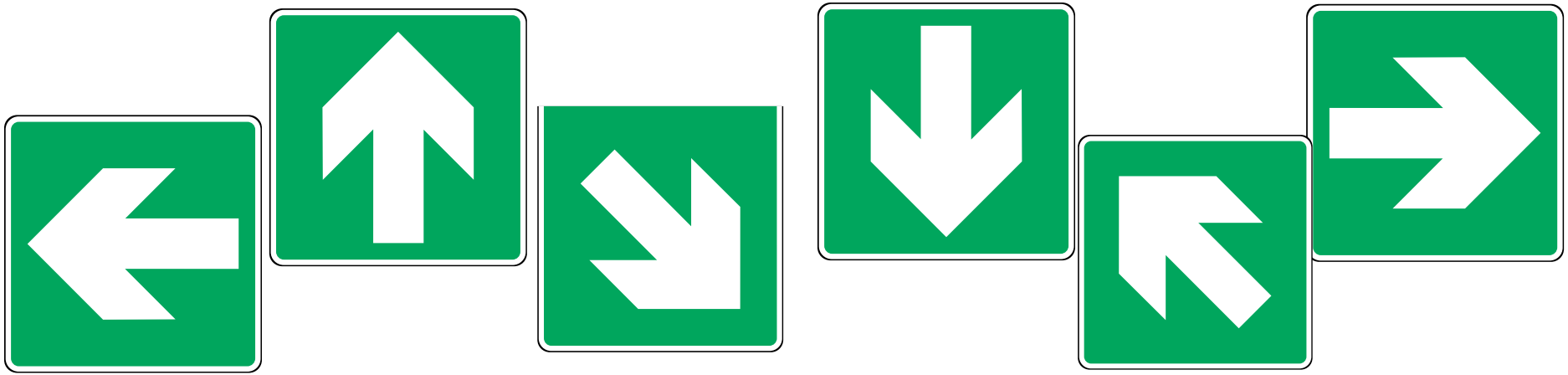
SRR Solution(?)

```
"storageshares" : [
  {
    "name" : "data",
    "totalsize" : 500TB,
    "usedsize" : 0,
    "vos" : [ "foo" ]
  }
```

SRR Solution(?)

```
[srrDomain]
[srr/frontend]
frontend.authn.basic=true
frontend.authn.protocol=http
frontend.authz.anonymous-operations=READONLY
frontend.srr.shares=data:/vo1,data:/vo2,archive:/vo1
```

Roadmap



Pick Your Favorite One

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

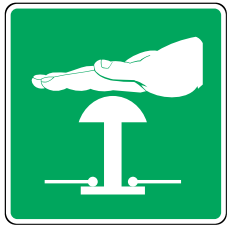
Sharing & Exchange

- 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

Community effort



- You can contribute with ...
  - Code
  - Configuration
  - HW setup
  - Knowledge
- You can make dCache visible with …
  - Sharing your use case
  - Demonstrate dCache use in various projects (DOMA, ESCAPE, ...)