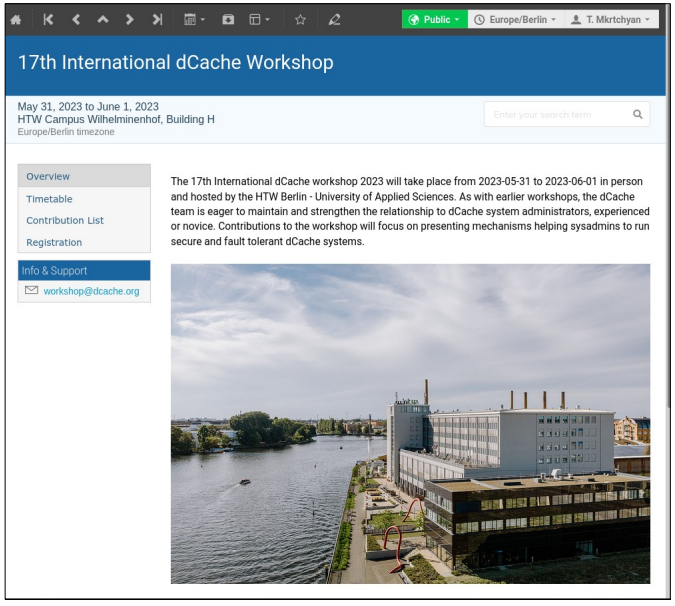
# 17th dCache User Workshop

DESY. Fermilab neic
NORDIC E-INFRASTRUCTURE COLLABORATION
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

# Logistic

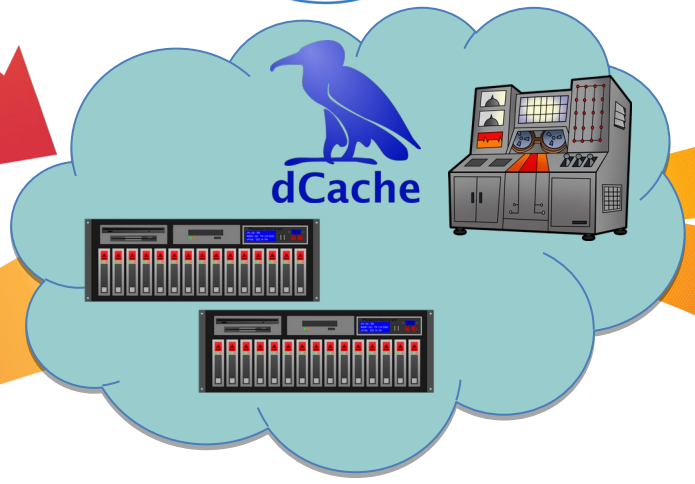
- The first in-person workshop since 2019
  - Hosted by: *HTW Berlin - University of Applied Sciences*
  - May 31 – June 1
  - ½ + 1 day
  - 35 participants (3 remote)
  - 12 contributions ( 9 from sites)



High Speed
Data Ingest

Data management
& workflow control

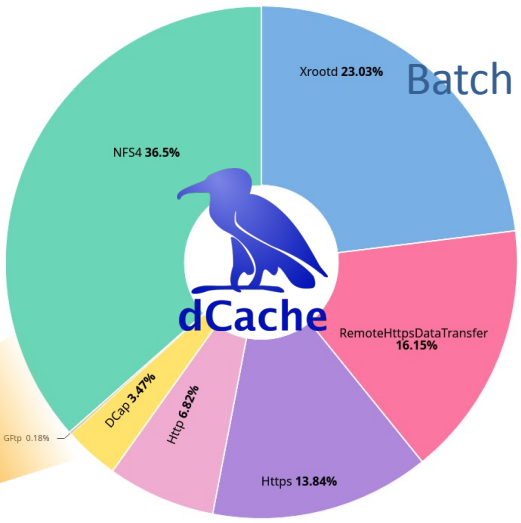
Batch processing




Interactive analysis

Wide Area Transfers

High Speed
Data Ingest



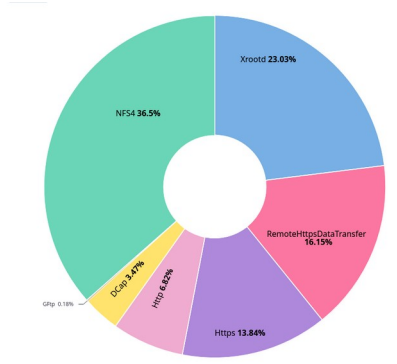

Batch processing

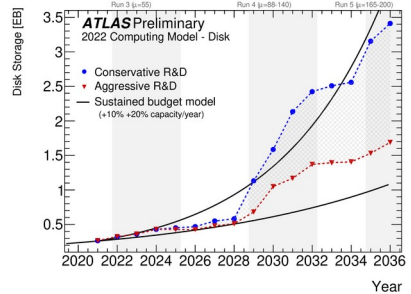
Interactive analysis

Wide Area Transfers

# The Challenges

- Data is going to grow… A lot…
  - High ingest data rates
  - More movements between sites
- Shared Computing Resources
  - Analysis Facilities
  - Grid Farms
  - HPC
  - Cloud resources (CPU&Storage)
- Standard analysis tools
  - ROOT
  - Jupyter Notebooks, non-ROOT analysis
- Competing Tape Operations



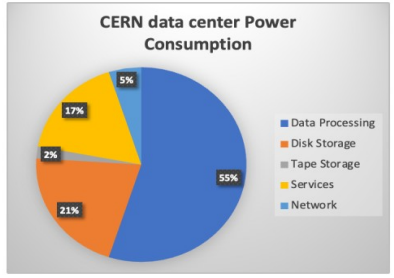




## WLCG data centers power consumption

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

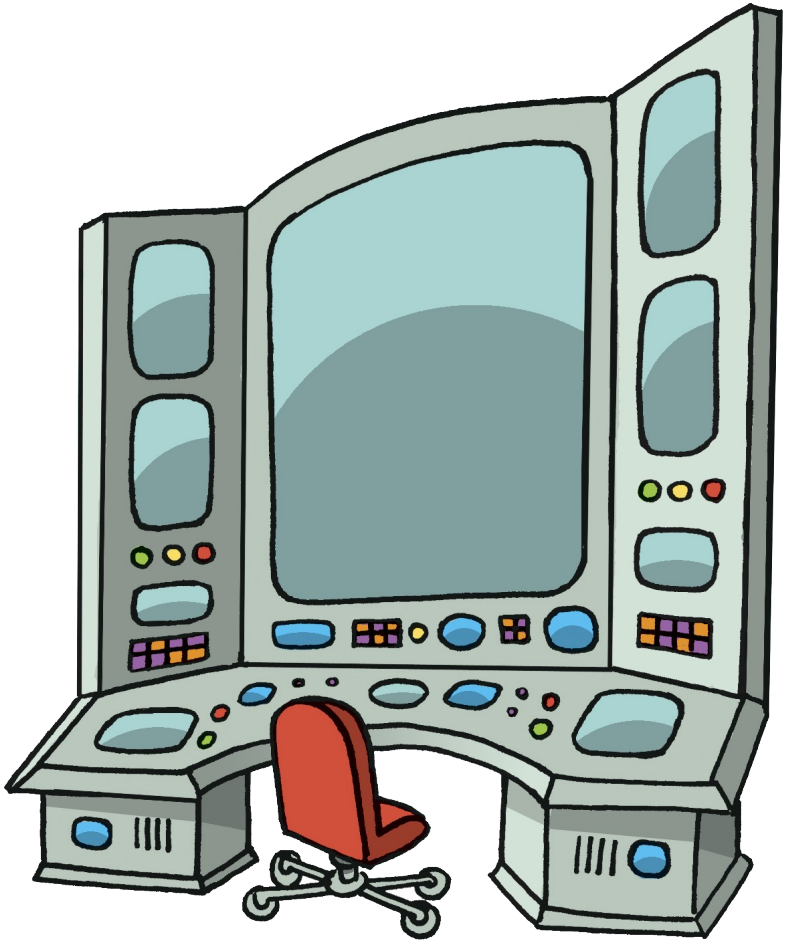


# Technical Directions

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

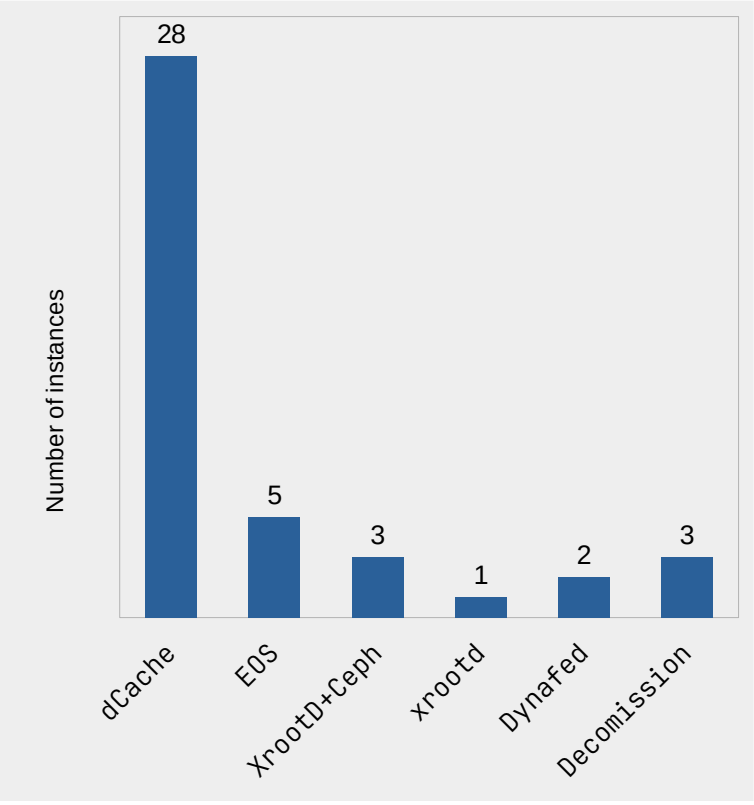
# Some Numbers

- 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
- NextCloud
  - File lifetime < 1s



# DPM Migration

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



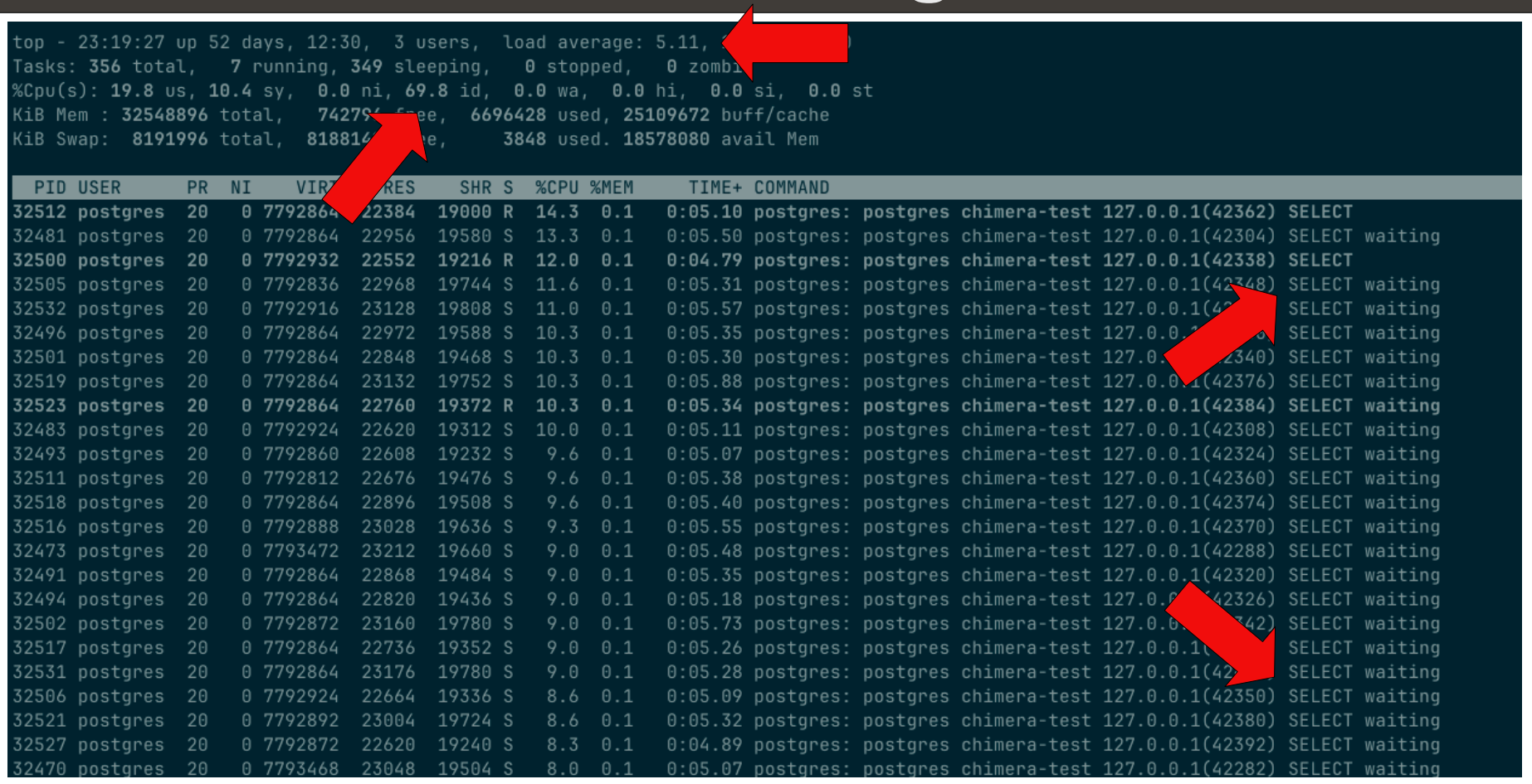

https://docs.google.com/spreadsheets/d/1KDVAJ9JzlycA3Wrz1iY2fQxZndWdAezFnLaDAxXIpUs/edit

# Re-cap

# Prominent Changes

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

# dCache Quiz: What Going On?!



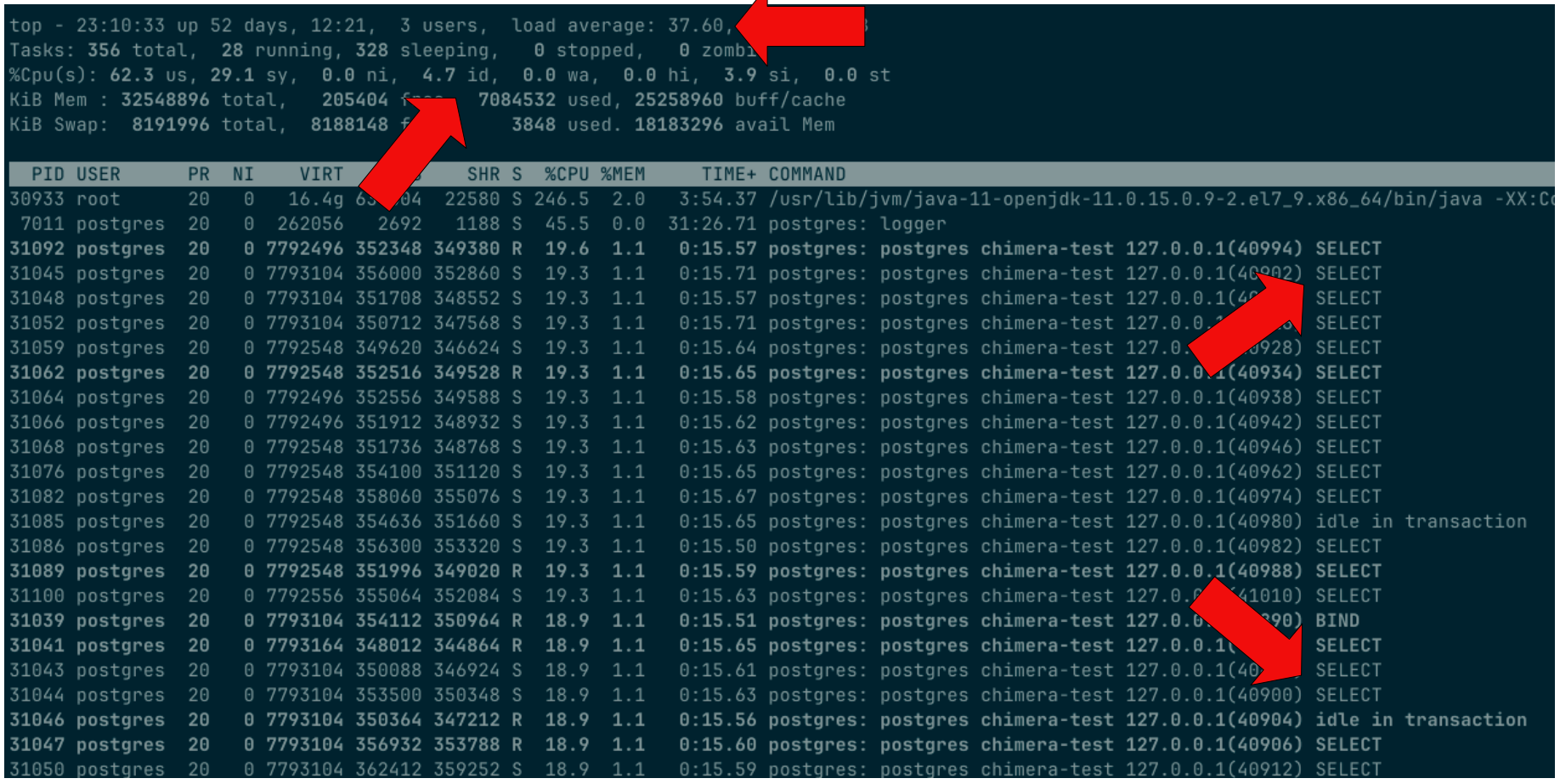
```
top - 23:19:27 up 52 days, 12:30,  3 users,  load average: 5.11,
Tasks: 356 total,   7 running, 349 sleeping,   0 stopped,   0 zombi
%Cpu(s): 19.8 us, 10.4 sy,  0.0 ni, 69.8 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 32548896 total,   74279  free,  6696428 used, 25109672 buff/cache
KiB Swap:  8191996 total,  818814  e,     3848 used. 18578080 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
32512 postgres  20   0 7792864  22384  19000 R  14.3  0.1   0:05.10 postgres: postgres chimera-test 127.0.0.1(42362) SELECT
32481 postgres  20   0 7792864  22956  19580 S  13.3  0.1   0:05.50 postgres: postgres chimera-test 127.0.0.1(42304) SELECT waiting
32500 postgres  20   0 7792932  22552  19216 R  12.0  0.1   0:04.79 postgres: postgres chimera-test 127.0.0.1(42338) SELECT
32505 postgres  20   0 7792836  22968  19744 S  11.6  0.1   0:05.31 postgres: postgres chimera-test 127.0.0.1(42348) SELECT waiting
32532 postgres  20   0 7792916  23128  19808 S  11.0  0.1   0:05.57 postgres: postgres chimera-test 127.0.0.1(42   SELECT waiting
32496 postgres  20   0 7792864  22972  19588 S  10.3  0.1   0:05.35 postgres: postgres chimera-test 127.0.0.1      SELECT waiting
32501 postgres  20   0 7792864  22848  19468 S  10.3  0.1   0:05.30 postgres: postgres chimera-test 127.0.   2340) SELECT waiting
32519 postgres  20   0 7792864  23132  19752 S  10.3  0.1   0:05.88 postgres: postgres chimera-test 127.0.0.1(42376) SELECT waiting
32523 postgres  20   0 7792864  22760  19372 R  10.3  0.1   0:05.34 postgres: postgres chimera-test 127.0.0.1(42384) SELECT waiting
32483 postgres  20   0 7792924  22620  19312 S  10.0  0.1   0:05.11 postgres: postgres chimera-test 127.0.0.1(42308) SELECT waiting
32493 postgres  20   0 7792860  22608  19232 S   9.6  0.1   0:05.07 postgres: postgres chimera-test 127.0.0.1(42324) SELECT waiting
32511 postgres  20   0 7792812  22676  19476 S   9.6  0.1   0:05.38 postgres: postgres chimera-test 127.0.0.1(42360) SELECT waiting
32518 postgres  20   0 7792864  22896  19508 S   9.6  0.1   0:05.40 postgres: postgres chimera-test 127.0.0.1(42374) SELECT waiting
32516 postgres  20   0 7792888  23028  19636 S   9.3  0.1   0:05.55 postgres: postgres chimera-test 127.0.0.1(42370) SELECT waiting
32473 postgres  20   0 7793472  23212  19660 S   9.0  0.1   0:05.48 postgres: postgres chimera-test 127.0.0.1(42288) SELECT waiting
32491 postgres  20   0 7792864  22868  19484 S   9.0  0.1   0:05.35 postgres: postgres chimera-test 127.0.0.1(42320) SELECT waiting
32494 postgres  20   0 7792864  22820  19436 S   9.0  0.1   0:05.18 postgres: postgres chimera-test 127.0.0   42326) SELECT waiting
32502 postgres  20   0 7792872  23160  19780 S   9.0  0.1   0:05.73 postgres: postgres chimera-test 127.0.0   342) SELECT waiting
32517 postgres  20   0 7792864  22736  19352 S   9.0  0.1   0:05.26 postgres: postgres chimera-test 127.0.0.1(   SELECT waiting
32531 postgres  20   0 7792864  23176  19780 S   9.0  0.1   0:05.28 postgres: postgres chimera-test 127.0.0.1(42   SELECT waiting
32506 postgres  20   0 7792924  22664  19336 S   8.6  0.1   0:05.09 postgres: postgres chimera-test 127.0.0.1(42350) SELECT waiting
32521 postgres  20   0 7792892  23004  19724 S   8.6  0.1   0:05.32 postgres: postgres chimera-test 127.0.0.1(42380) SELECT waiting
32527 postgres  20   0 7792872  22620  19240 S   8.3  0.1   0:04.89 postgres: postgres chimera-test 127.0.0.1(42392) SELECT waiting
32470 postgres  20   0 7793468  23048  19504 S   8.0  0.1   0:05.07 postgres: postgres chimera-test 127.0.0.1(42282) SELECT waiting
```

# POSIX Constraints

- According to POSIX standard, on new file system object creation the parent directories *modification time* should be updated.
- To track the directory changes that happen at a higher rate than the precision of mtime attribute Linux kernel has an additional attribute *iversion* that is incremented whenever the inode's data is changed.
- To reduce unnecessary directory listing requests to the servers, the NFSv4 clients utilize the *iversion* attribute to identify the directory content changes and use the locally cached copy of the directory entry list as long as last known *iversion* attribute value matches the remote one.

# Near-POSIX Behavior



# 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

# 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

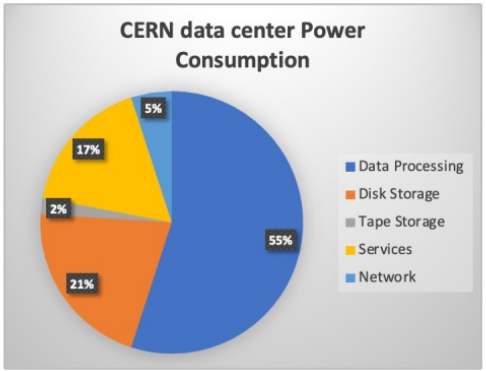
# The Renaissance of Tape?

## WLCG data centers power consumption

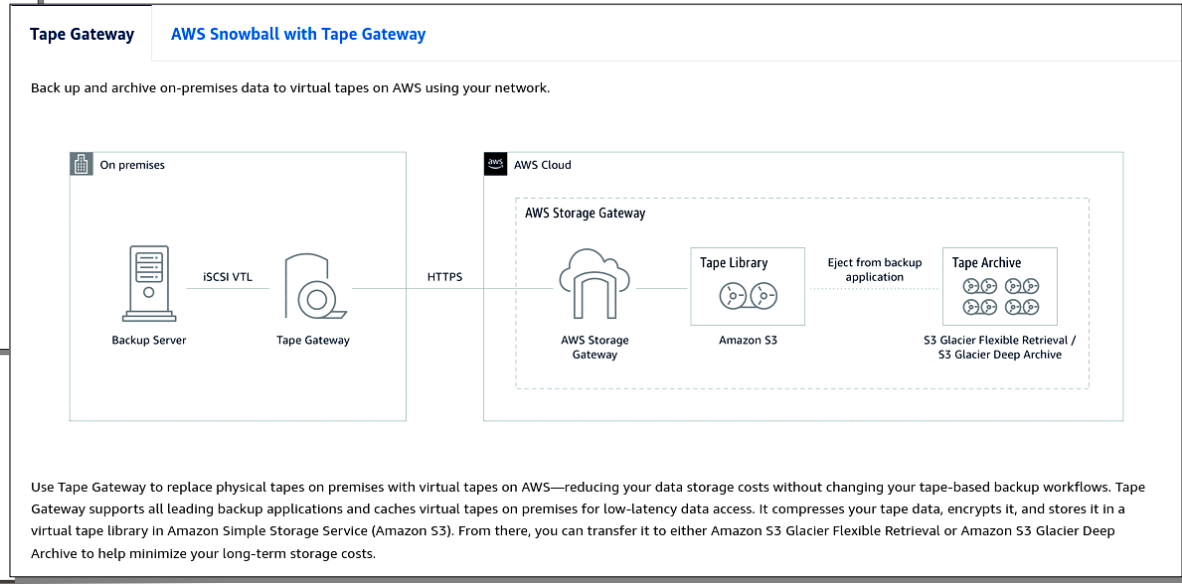
The pie chart shows the breakdown of the power consumption at the CERN data center

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

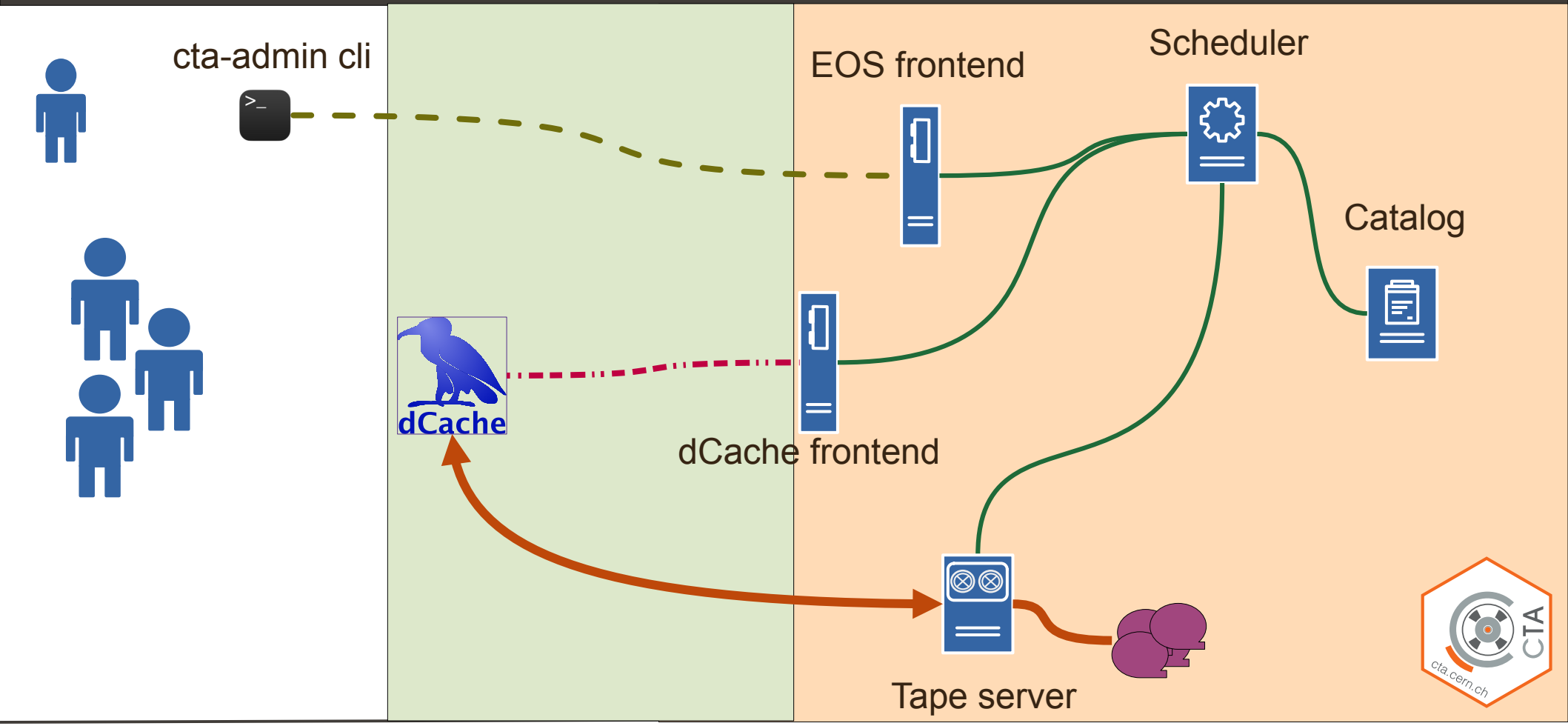
In this study we will focus on the energy needs for CPUs



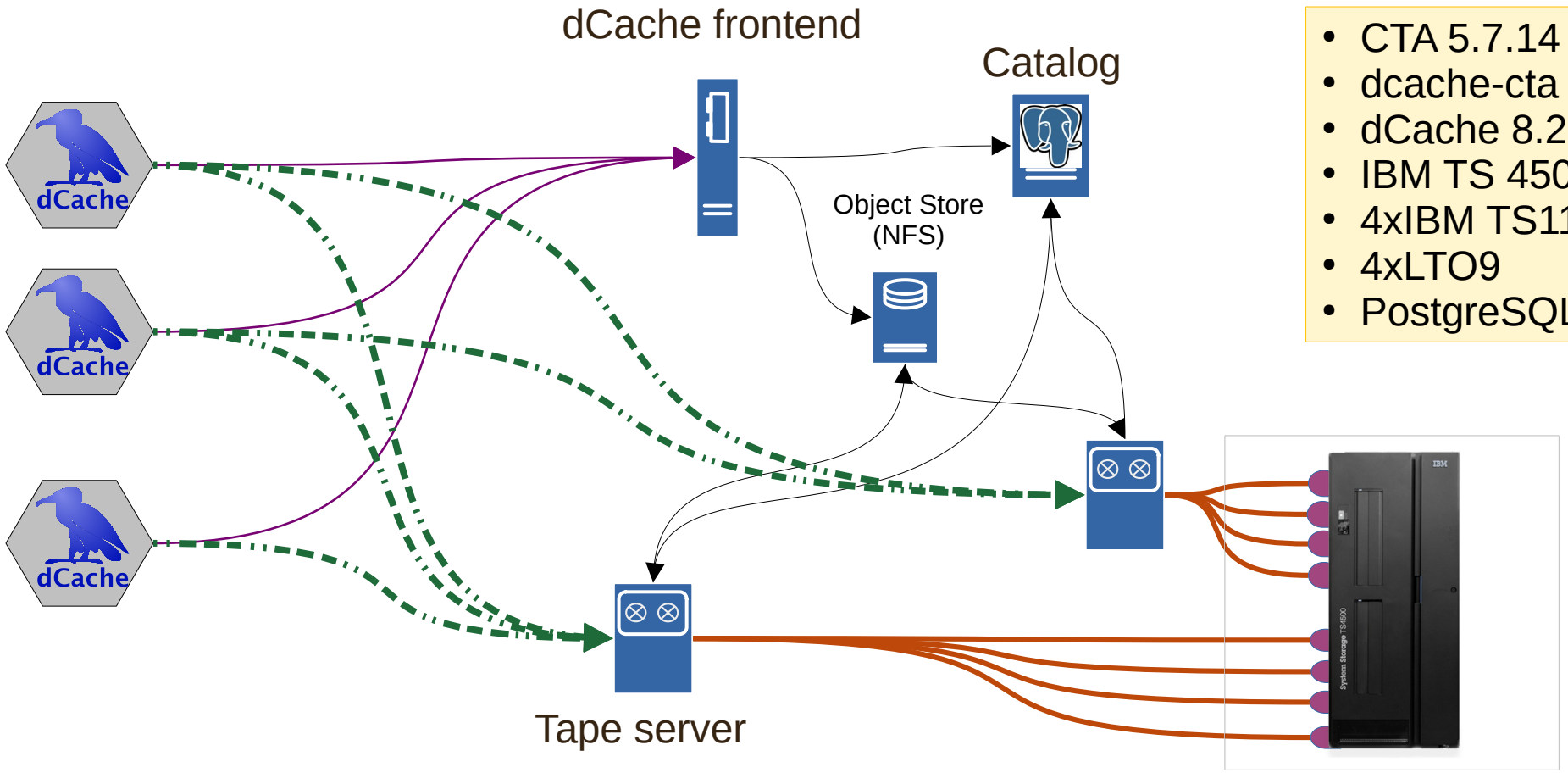

Shameless stolen from Simone Campana



# 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 starting {4,5}.7.12 provide dCache required functionality
  - The proposed dCache interface is under adoption by EOS.
- 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 are migrated to CTA (5.4 PB)
  - EuXFEL migration will take place next week (Jul 17-21) (99 PB)
- dCache+CTA deployment replicate to by other HEP sites
  - Fermilab and PIC Barcelona have successfully replicated our setup (currently dCache + ENSTORE).
  - RAL in UK plans to migrate to PostgreSQL from ORACLE based on our experience

# dCache Bulk Service and WLCG TAPE API: The Demo, Redux

Albert L. Rossi (FNAL)

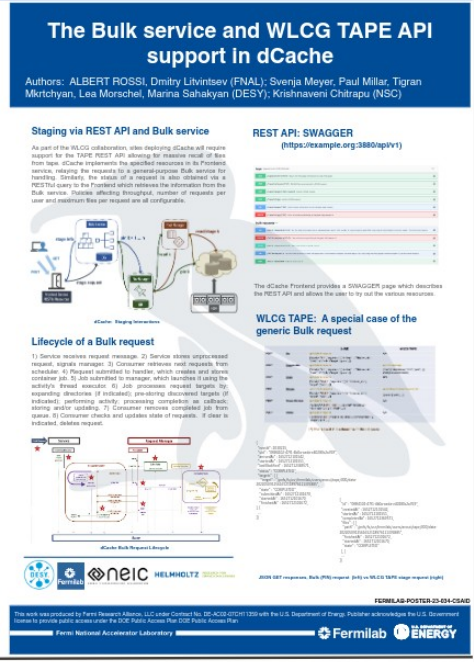
dCache
and WLC
The D

Albert

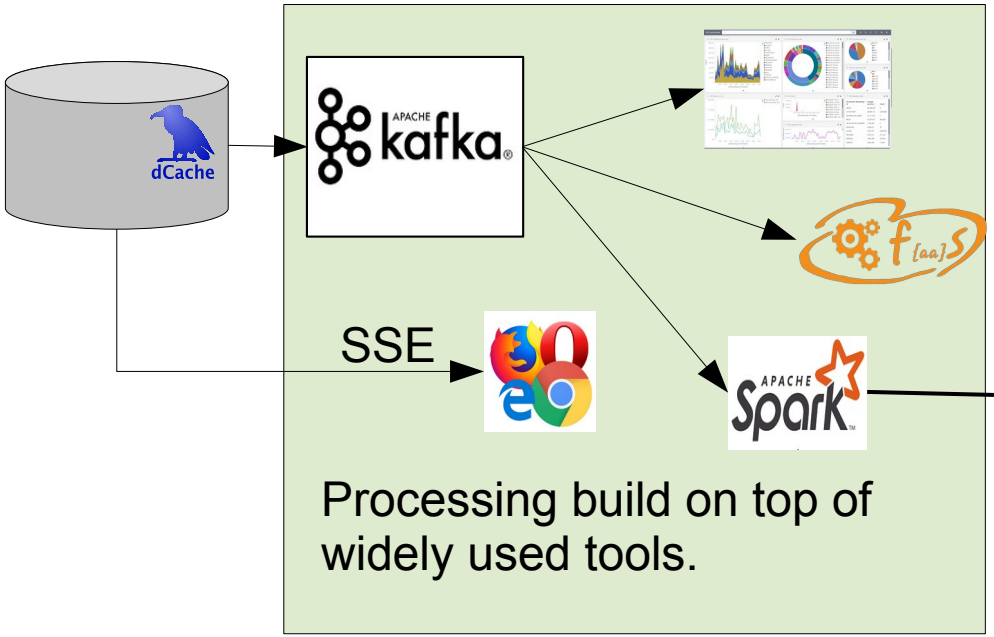
Thursday, June 1, 2023 Bulk v2

# The dCache Bulk Service

- Introduced last year.
- Since then, many improvements, especially a substantial reworking of the data storage layer.
- Will not describe, but simply demo the capabilities.



# Big-Data Tools for Log Processing



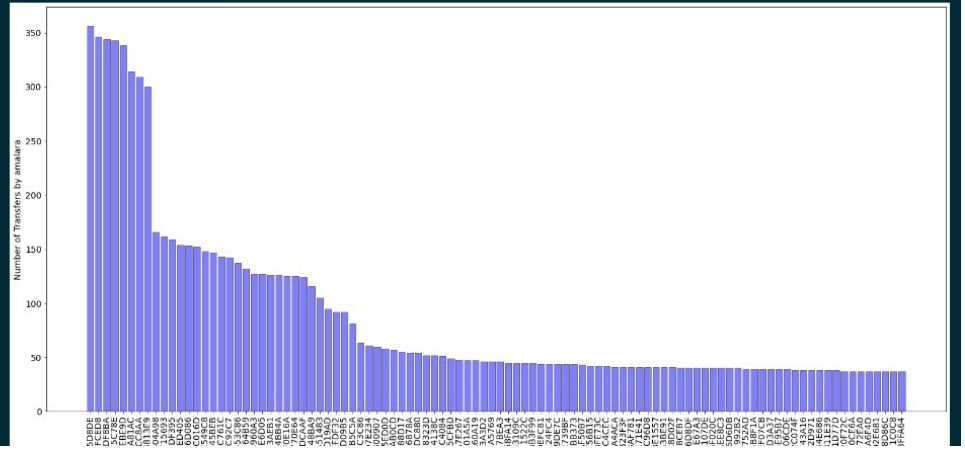

```
files_array  = user_pool.rdd.map(lambda row: row[0]).collect()
counts_array = user_pool.rdd.map(lambda row: row[1]).collect()

plt.rcParams.update({'font.size': 14})
fig = plt.figure(figsize=(26, 12), dpi=72, facecolor='w')
plt.xticks(rotation=90)
plot = fig.add_subplot(111)
plot.bar(files_array, counts_array, color='blue',edgecolor = 'black', alpha=0.5)

plt.ylabel('Number of Transfers by amalara')
plt.xlabel('CMS dCache PNFSID')
plt.show()
```

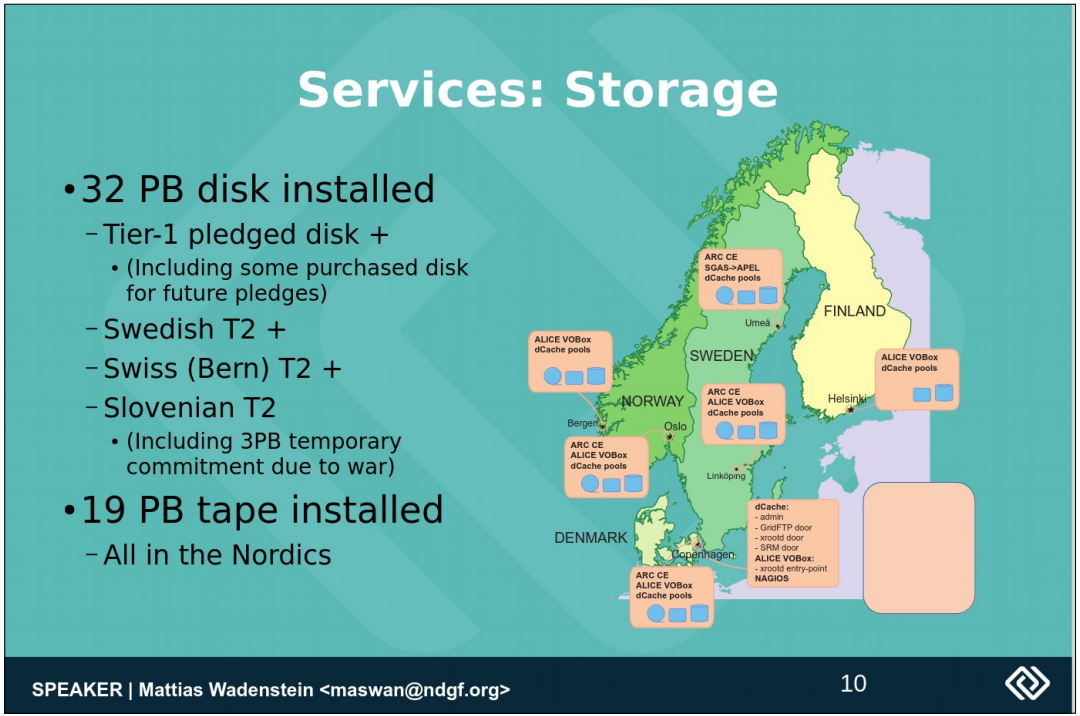
[83] Python

```
/usr/local/lib/python3.6/site-packages/ipykernel_launcher.py:7: MatplotlibDeprecationWarning:
  import sys
```




*By Christian Voß*

# Distributed dCache (datalake) Operation




*By: Mattias Wadenstein*

# Distributed dCache operation

## Services: Storage

- 32 PB disk installed
  - Tier-1 pledged disk +
    - (Including some purchased disk for future pledges)
  - Swedish T2 +
  - Swiss (Bern) T2 +
  - Slovenian T2
    - (Including 3PB temporary commitment due to war)
- 19 PB tape installed
  - All in the Nordics

SPEAKER | Mattias Wadenstein <maswan@ndgf.org>

## Challenges in federated storage

- The local funding agencies would like to see their contributions in WLCG storage accounting
- dCache supports SRR to publish partitioned storage accounting
- Apparently there is development needed in WLCG accounting in order to make this visible
  - Somewhat surprising to us

SPEAKER | Mattias Wadenstein <maswan@ndgf.org> 11

# Distributed dCache operation

## Services: Storage

- 32 PB disk installed
  - Tier-1 pledged disk +
    - (Including some purchased disk for future pledges)
  - Swedish T2 +
  - Swiss (Bern) T2 +
  - Slovenian T2
    - (Including 3PB temporary commitment due to war)
- 19 PB tape installed
  - All in the Nordics

SPEAKER | Mattias Wadenstein <maswan@ndgf.org>

## Challenges in federated storage

- The local funding ager
  contributions in WLCG
- dCache supports SRR
  storage accounting
- Apparently there is de
  accounting in order to
  - Somewhat surprising to us

SPEAKER | Mattias Wadenstein <maswan@ndgf.or

## Storage operations

- Local site admins maintain hardware, filesystem, operating system, networking, kernel tuning
  - Provides one unpriviledged account with lots of storage to the central ops team
- Central ops team runs dCache pools
  - Install java + dCache
  - Configure, upgrade, restart dCache
- Investigating issues sometimes takes cooperation
  - Pool shutdown (central ops notice) due to IO error (investigation by both) because of raid controller issue (local ops fix)

SPEAKER | Mattias Wadenstein <maswan@ndgf.org> 12

# Distributed dCache operation

## Serv

- 32 PB disk installe
  - Tier-1 pledged disk +
    - (Including some purchase for future pledges)
  - Swedish T2 +
  - Swiss (Bern) T2 +
  - Slovenian T2
    - (Including 3PB temporary commitment due to war)
- 19 PB tape installe
  - All in the Nordics

SPEAKER | Mattias Wadenstein <maswan(

## Collaboration

- A successful data lake is a successful collaboration between:
  - Funding agencies – usually one in each participating country
  - Sysadmins – NeIC central team and site admins at each site
  - Physics projects and their PIs – one to two per country for us
  - Networking providers – NORDUNet, GEANT, CERN, plus all NRENs
  - Researchers – the entire purpose of research infrastructure
  - Experiment coordinators – ALICE and ATLAS currently
  - Scientific computing centers – Nine currently participating
  - Coordinating body – Nordic e-Infrastructure Collaboration, NeIC
  - etc
  - etc

SPEAKER | Mattias Wadenstein <maswan@ndgf.org> 16

## ons

re, filesystem,
nel tuning
of storage to the

ls

kes cooperation
error (investigation by
ps fix)

SPEAKER | Mattias Wadenstein <maswan@ndgf.org> 12

# dCache for Rubin LSST

## HISTORY OF dCache AT CC-IN2P3

- Started getting familiar with dCache v1.6.5 in 2004
- Currently operating several instances for different projects
- LHC [v8.2.16]
  *shared by ATLAS, CMS, LHCb*
  *38 PB, 157 servers, 155 M objects*
- EGEE [v7.2.27]
  *shared by CTA, Juno, Belle II, Calice, Dune, Xenon*
  *1.4 PB, 7 servers, 50 M objects*
- NESSIE [v8.2.16]
  *for R&D purposes, currently mainly ESCAPE and DOMA*
  *3 servers, 300 TB*
- Rubin LSST
  *the subject of this talk*

georget | hernandez | le boulc'h

CCIN2P3 2

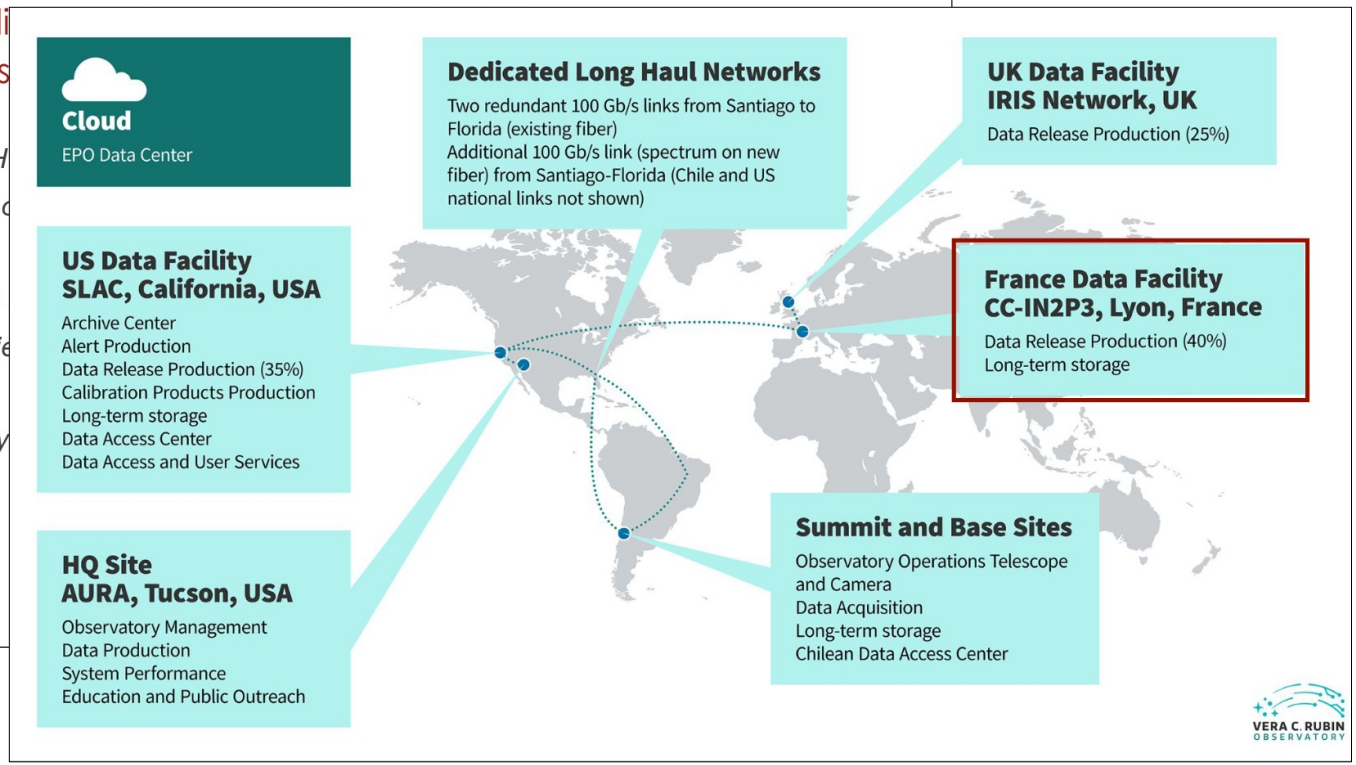
*By: Fabio Hernandez*

# dCache for Rubin LSST

## HISTORY OF dCache AT CC-IN2P3

- Started getting famili
- Currently operating s
- LHC [v8.2.16]
  *shared by ATLAS, CMS, LH*
  *38 PB, 157 servers, 155 M*
- EGEE [v7.2.27]
  *shared by CTA, Juno, Belle*
  *1.4 PB, 7 servers, 50 M obje*
- NESSIE [v8.2.16]
  *for R&D purposes, currently*
  *3 servers, 300 TB*
- Rubin LSST
  *the subject of this talk*

georget | hernandez | le boulc'h



# dCache for Rubin LSST

## HISTORY OF dCache AT CC-IN2P3

- Started getting f
- Currently operat
- LHC [v8.2.16]
  *shared by ATLAS, CM*
  *38 PB, 157 servers, 15*
- EGEE [v7.2.27]
  *shared by CTA, Juno,*
  *1.4 PB, 7 servers, 50 I*
- NESSIE [v8.2.16]
  *for R&D purposes, cu*
  *3 servers, 300 TB*
- Rubin LSST
  *the subject of this talk*

georget | hernandez | le boulc'h

**Cloud**
EPO Data Center

**US Data Facility**
**SLAC, California, USA**
Archive Center
Alert Production
Data Release Production (35%)
Calibration Products Production
Long-term storage
Data Access Center
Data Access and User Services

**HQ Site**
**AURA, Tucson, USA**
Observatory Management
Data Production
System Performance
Education and Public Outreach

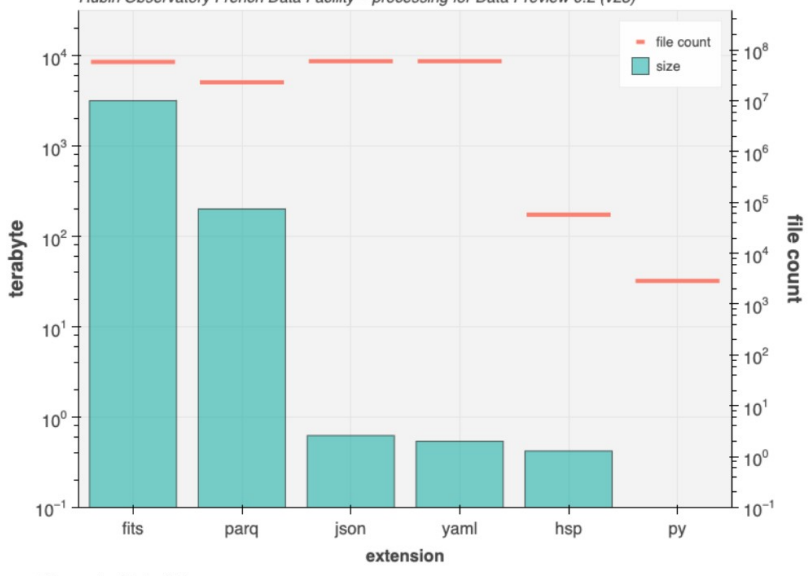
**Dedicated Long Haul Networks**
Two redundant 100 Gb/s links from Santiago to Florida (existing fiber)

**UK Data Facility**
**IRIS Network, UK**
Data Release Production (25%)

## RUBIN DATA PREVIEW: DATA PRODUCTS SIZES

**DP.02 products: file count and aggregated size**

*Rubin Observatory French Data Facility – processing for Data Preview 0.2 (v23)*




High number of (very) small files resulting from processing of raw images

georget | hernandez | le boulc'h

CCIN2P3 16

# dCache for Rubin LSST

## History of dCache at CC-IN2P3

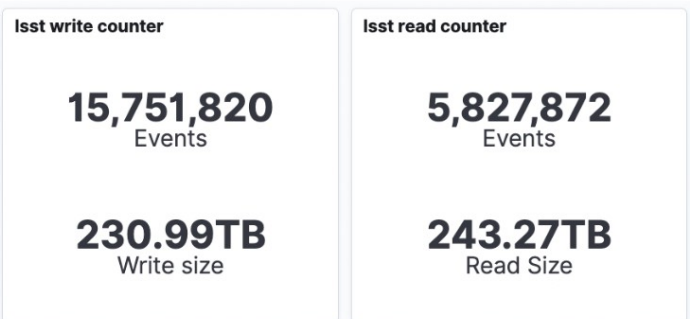
- Started gett
- Currently op
- LHC [v8.2.16]
  *shared by ATL*
  *38 PB, 157 ser*
- EGEE [v7.2.27]
  *shared by CTA,*
  *1.4 PB, 7 serve*
- NESSIE [v8.2.1
  *for R&D purpos*
  *3 servers, 300*
- Rubin LSST
  *the subject of t*

georget | hernandez | le boulc'h

## Activity profile: LCG vs Rubin LSST


LCG instance
5M events
(HTTP + XRootD)

15k jobs
157 pools




Rubin LSST instance
21M events (webDAV)

5k jobs in execution
19 pools

☝ *observed activity over a representative period of 48 h*

georget | hernandez | le boulc'h

CCIN2P3 22

## ...ZES

...ery) small files
...essing of raw

georget | hernandez | le boulc'h

CCIN2P3 16

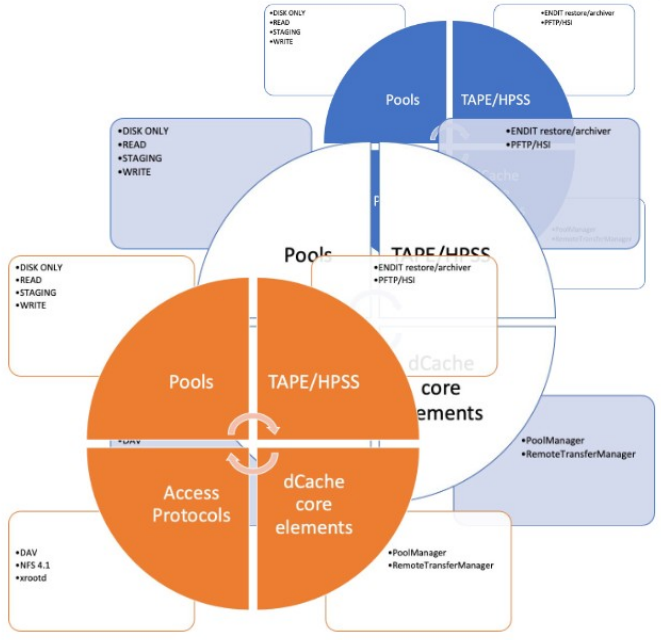
# dCache at BNL

dCache.org

distributed storage for scientific data

dCache instances are isolated per SC

- SC diverge in their requirements
- Procurement and resource control
- Infrastructure supported on physical and virtual Machines

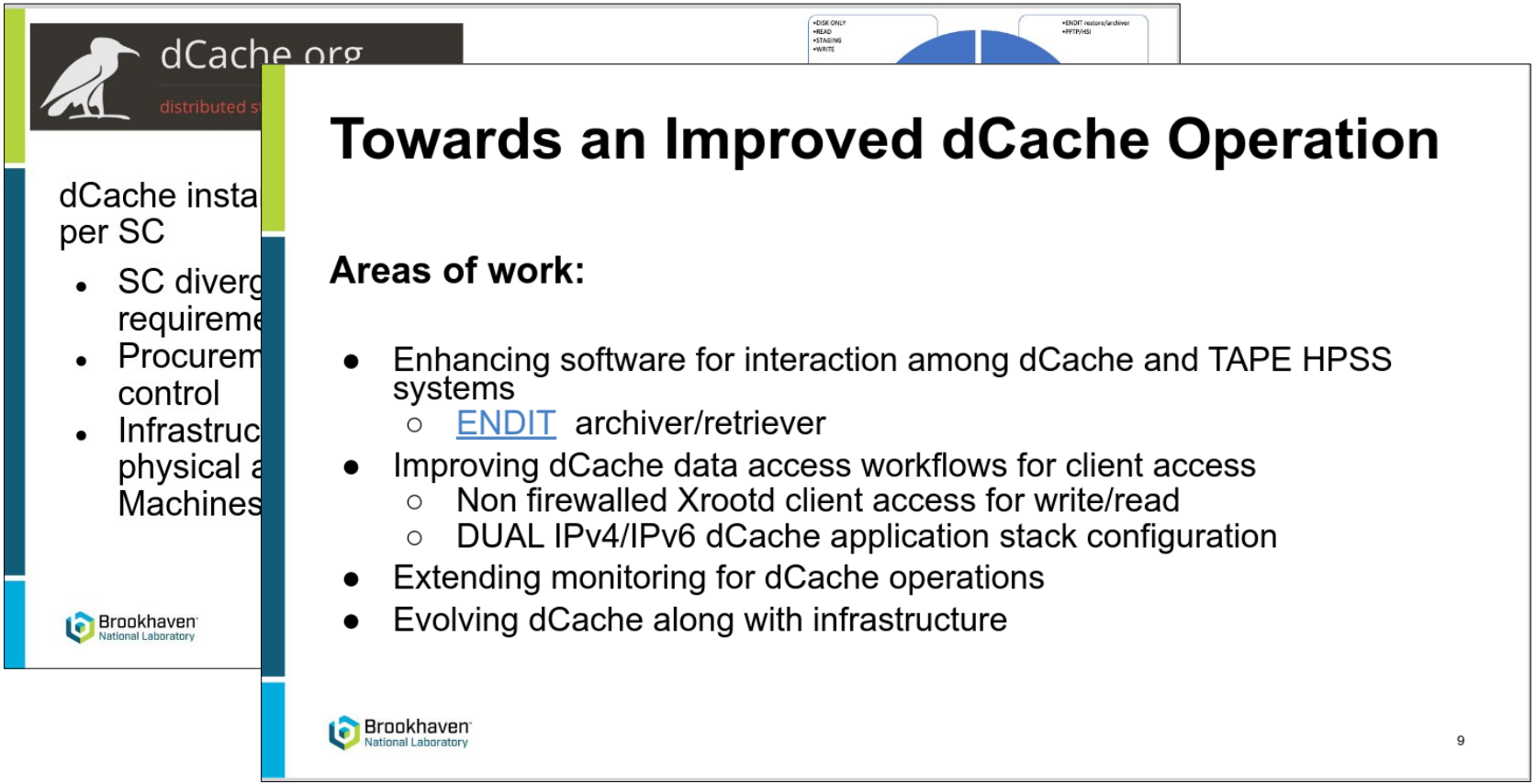



Brookhaven National Laboratory

Storage technology flexible to support SC individual requirements

8

# dCache at BNL



# dCache at BNL

dCache.org

distributed st

dCache insta
per SC

•DISK ONLY
•READ
•STAGING
•WRITE

## Towards an Improved

## dCache and ELK Stack Started to be Used in Operations

Filebeat / Logstash pipelines enabled for domain logs and billing logs

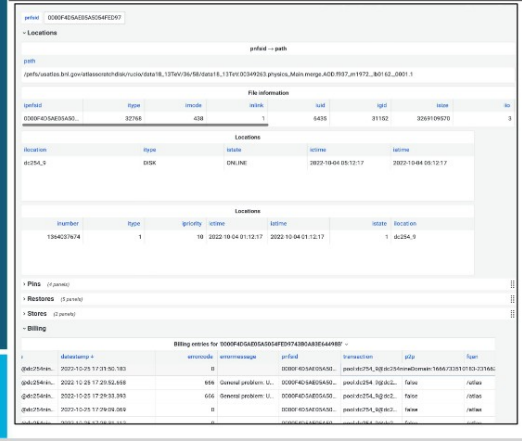

ELK use to mine the billing logs with arbitrary queries

Brookhaven National Laboratory

## Monitoring Enhancement

Grafana based monitor using the dCache billing/chimera/srm databases to provide information use in operations

Allows aggregate information from different dCache events by entering the PNFSID (dCache file ID)



Feature driven dashboards

Performance of dCache

# dCache at BNL

## Monitoring Enhancement

Grafana based monitor using the dCache billing/chimera/srm databases to provide information use in operations

Feature driven dashboards

Performance of dCache

dCache.org

distributed st

dCache insta
per SC

dCache and ELK S
Started to be Used

Filebeat / Logstash pipelines enabled for dom

usatlas-logs-dcache-billing-log

ID Events Emitted

usatlas-logs-dcache-billing-log

Rows per page: 50

Brookhaven National Laboratory

Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

**Performance evaluation & comparison: Lustre, dCache, Xrootd, zfs, etc**

@BrookhavenLab

1

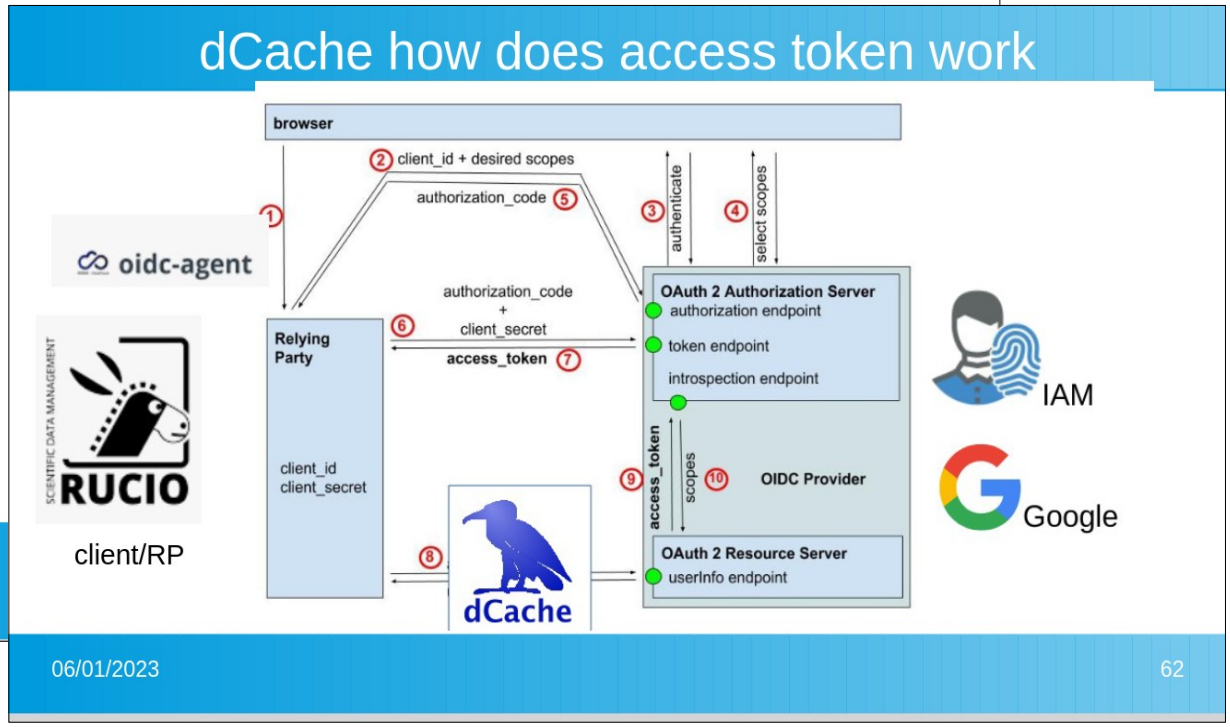
9

13

# OIDC and Token-based Access

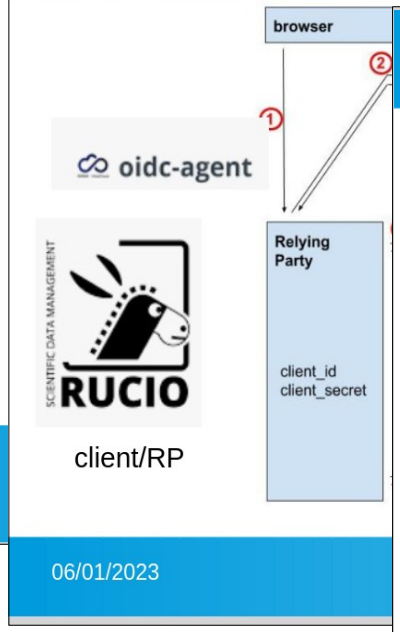

*By: Marina Sahakyan*

# OIDC and Token-based Access



# OIDC and Token-based Access

## dCache how does access token work




06/01/2023

## Identifying Token Types

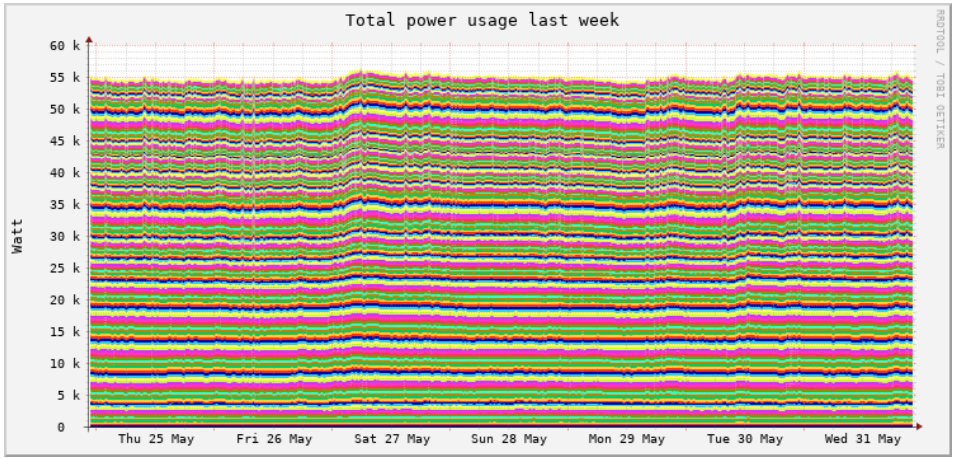
- It can be confusing sometimes to distinguish between the different token types.
  - **ID tokens**
    - carry identity information encoded in the token itself, which must be a JWT
  - **Access tokens**
    - used to gain access to resources by using them as bearer tokens
  - **Refresh tokens**
    - exist solely to get more access tokens

06/01/2023 19

# Sustainability

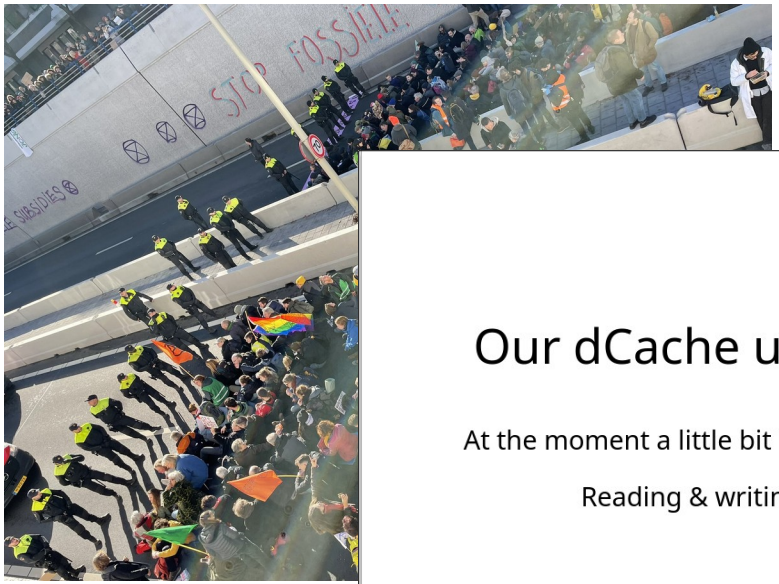

*By: Onno Zweers*

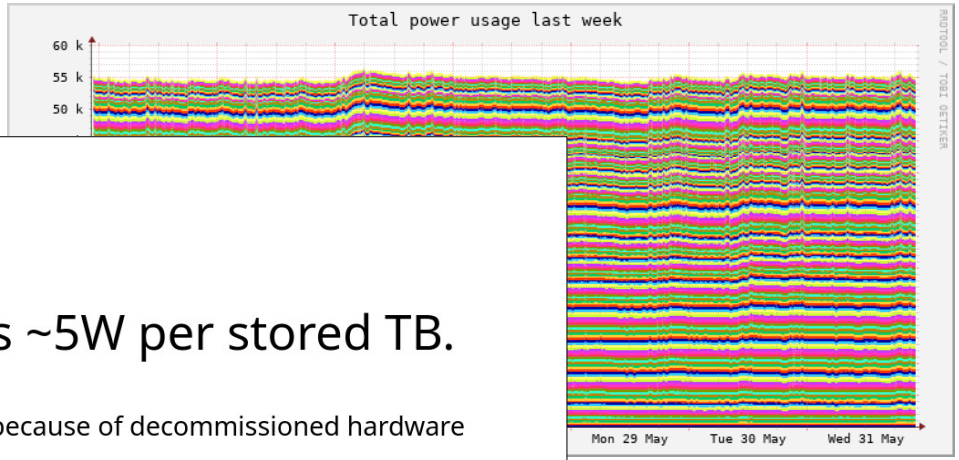
# Sustainability


*By: Onno Zweers*

# Sustainability




## Our dCache uses ~5W per stored TB.

At the moment a little bit less because of decommissioned hardware

Reading & writing does not make a big difference

*By: Onno Zweers*

# Sustainability


Total power usage last week

## Wrap-up

- Moving data from old, small, inefficient servers to new, large, efficient servers and switching them off early is an easy way to reduce energy consumption.
- Still investigating environmental impact, especially production of the hardware. When we have time.
- Haven't looked at hardware energy settings yet.
- "Green" electricity often based on certificates – doesn't help. Buying locally produced green electricity may stimulate energy suppliers to become more sustainable.
- All this pales in comparison to the effect you can have as a climate activist.

*By: Onno Zweers*

# 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*

# Useful Links

- Workshop Indico:
  - https://indico.desy.de/e/dcache-ws17
- dCache documentation:
  - https://www.dcache.org/documentation/
- Mini hands on:
  - https://github.com/dCache/dcache/blob/master/docs/TheBook/src/main/markdown/dcache-minimal-installation.md