



# 17<sup>th</sup> dCache User Workshop

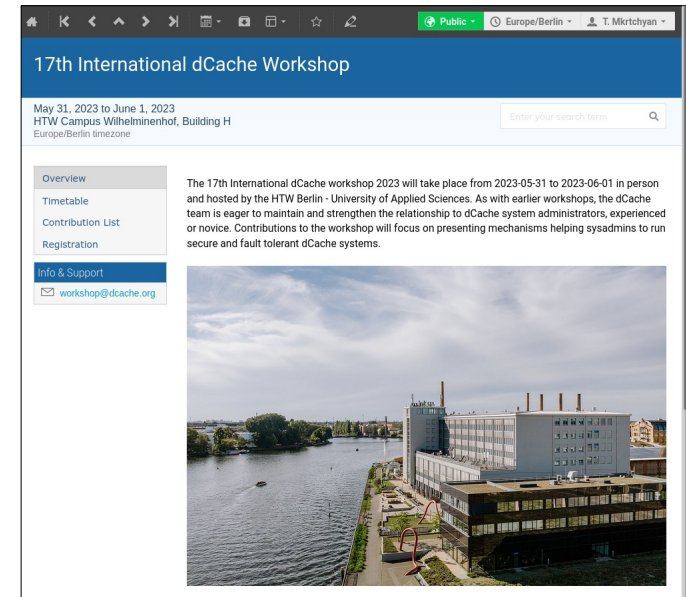


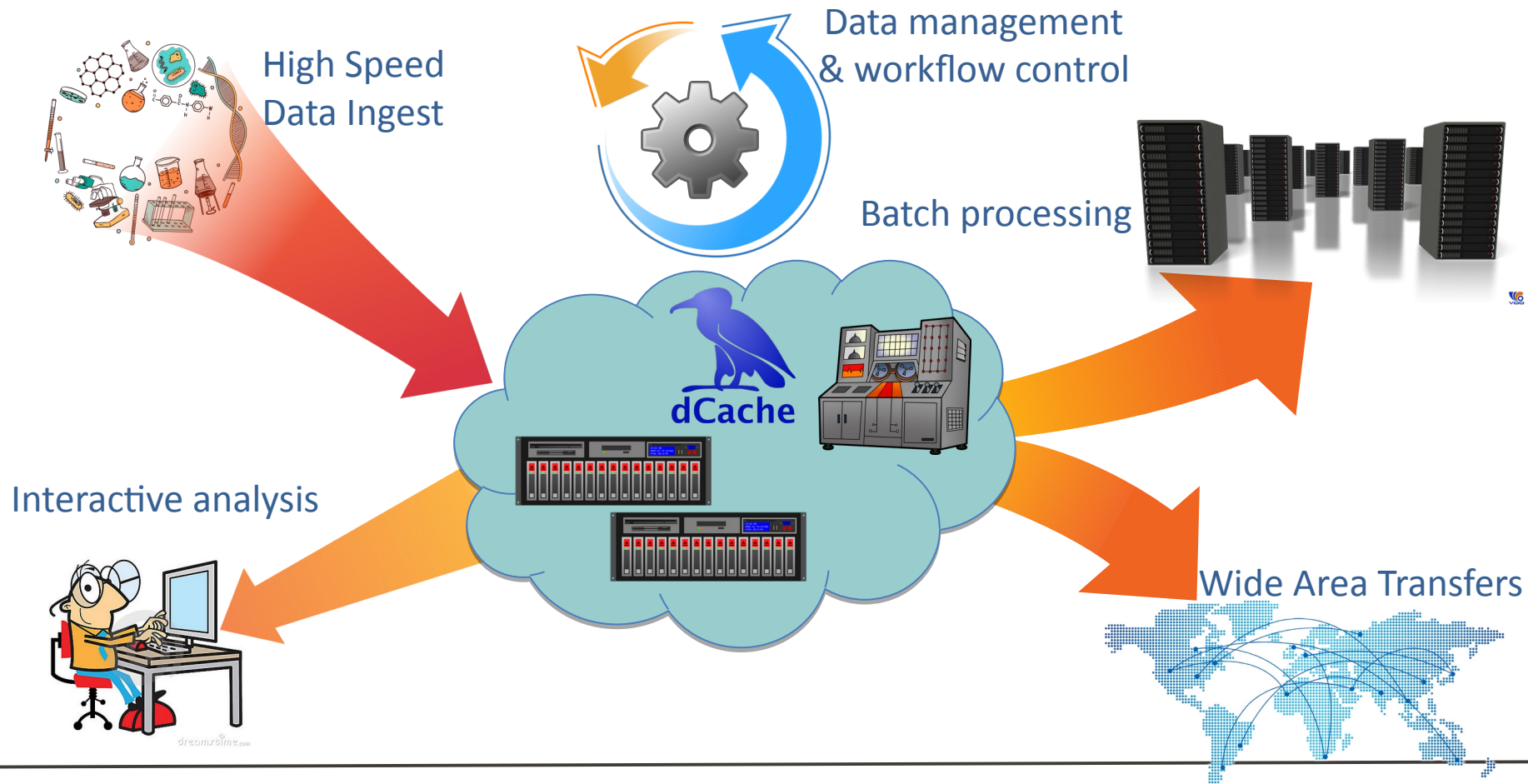
**HELMHOLTZ**

RESEARCH FOR  
GRAND CHALLENGES



- 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)

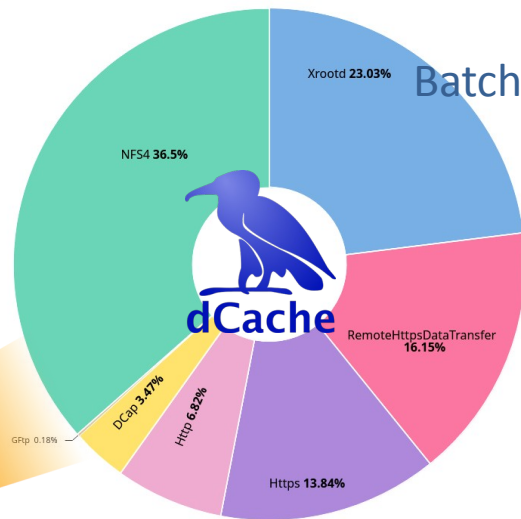






High Speed  
Data Ingest

Interactive analysis



Batch processing



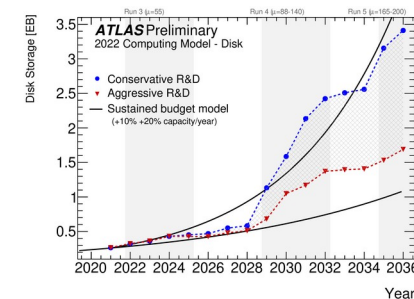
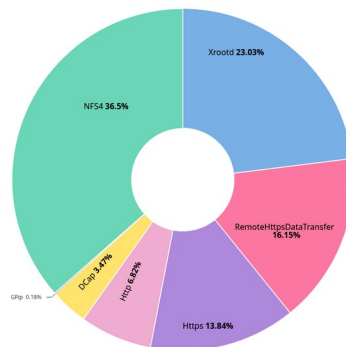
Wide Area Transfers



# 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

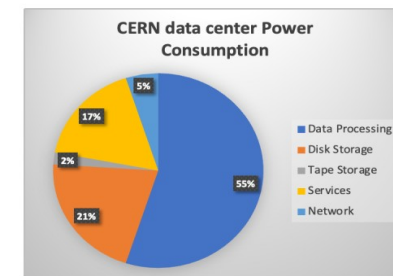


## WLCG data centers power consumption

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



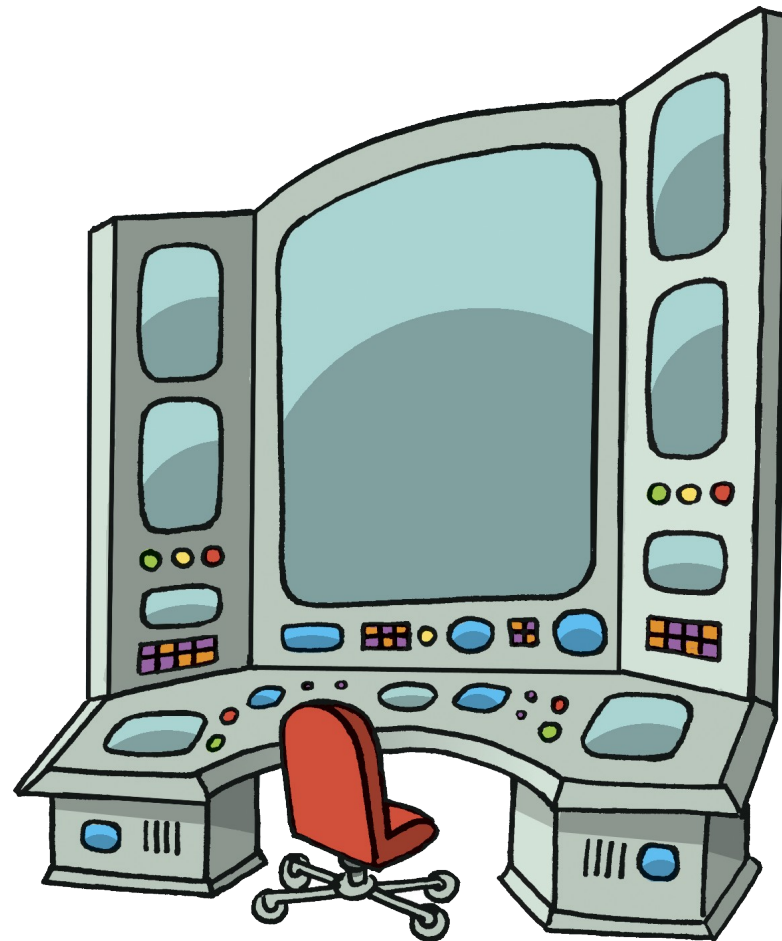
- Scaleout
  - Namespace
  - Number of pools (cells)
- Token-based Authentication
- Better *Analysis Facility* support
  - POSIX access and compliance
  - HPC workload support
- QoS
- Tape integration



# Some Numbers



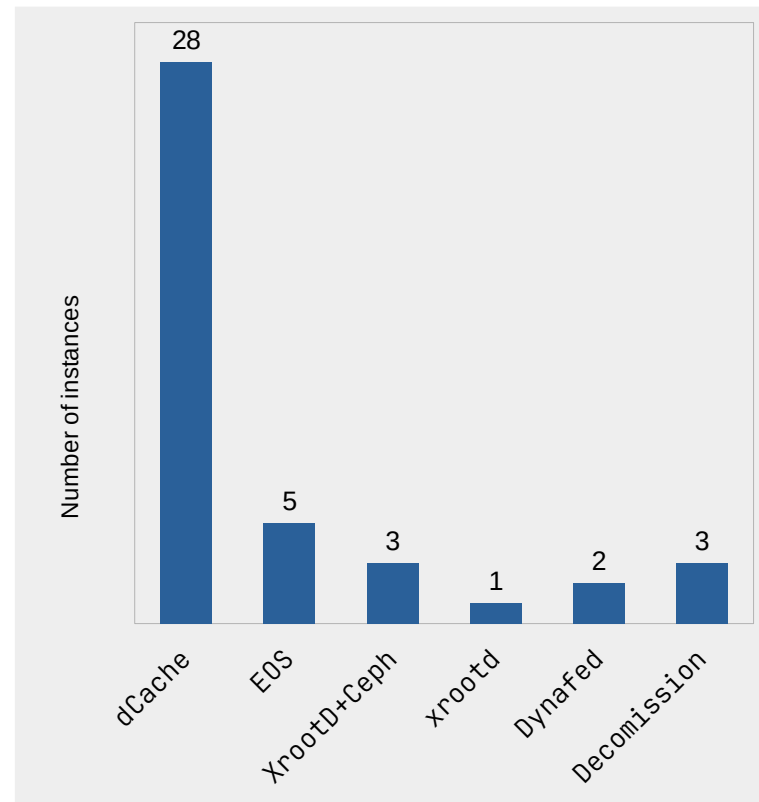
- XFEL
  - Total capacity ~120 PB
  - ~400 physical hosts (~4000 dCache pools)
  - 20-40 GB/s ingest
- Photon
  - DB size – 2.5TB
  - ACL table 600GB
  - Directories with  $3 \cdot 10^6$  files
  - $1.2 \cdot 10^9$  file system objects
  - 100K files in the flush queue
  - Two tape copies, different 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, 742796 free, 6696428 used, 25109672 buff/cache  
KiB Swap: 8191996 total, 818814 free, 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(42340) 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(42340) SELECT waiting
32501	postgres	20	0	7792864	22848	19468	S	10.3	0.1	0:05.30	postgres: postgres chimera-test 127.0.0.1(42340) 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 127.0.0.1(42326) SELECT waiting
32502	postgres	20	0	7792872	23160	19780	S	9.0	0.1	0:05.73	postgres: postgres chimera-test 127.0.0.1(42342) 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(42342) 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(42342) 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



- 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



```
top - 23:10:33 up 52 days, 12:21, 3 users, load average: 37.60,
Tasks: 356 total, 28 running, 328 sleeping, 0 stopped, 0 zombie
%Cpu(s): 62.3 us, 29.1 sy, 0.0 ni, 4.7 id, 0.0 wa, 0.0 hi, 3.9 si, 0.0 st
KiB Mem : 32548896 total, 205404 free, 7084532 used, 25258960 buff/cache
KiB Swap: 8191996 total, 8188148 free, 3848 used, 18183296 avail Mem
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
30933	root	20	0	16.4g	63084	22580	S	246.5	2.0	3:54.37	/usr/lib/jvm/java-11-openjdk-11.0.15.0.9-2.el7_9.x86_64/bin/java -XX:Co
7011	postgres	20	0	262056	2692	1188	S	45.5	0.0	31:26.71	postgres: logger
31092	postgres	20	0	7792496	352348	349380	R	19.6	1.1	0:15.57	postgres: postgres chimera-test 127.0.0.1(40994) SELECT
31045	postgres	20	0	7793104	356000	352860	S	19.3	1.1	0:15.71	postgres: postgres chimera-test 127.0.0.1(40992) SELECT
31048	postgres	20	0	7793104	351708	348552	S	19.3	1.1	0:15.57	postgres: postgres chimera-test 127.0.0.1(40993) SELECT
31052	postgres	20	0	7793104	350712	347568	S	19.3	1.1	0:15.71	postgres: postgres chimera-test 127.0.0.1(40995) SELECT
31059	postgres	20	0	7792548	349620	346624	S	19.3	1.1	0:15.64	postgres: postgres chimera-test 127.0.0.1(40928) SELECT
31062	postgres	20	0	7792548	352516	349528	R	19.3	1.1	0:15.65	postgres: postgres chimera-test 127.0.0.1(40934) SELECT
31064	postgres	20	0	7792496	352556	349588	S	19.3	1.1	0:15.58	postgres: postgres chimera-test 127.0.0.1(40938) SELECT
31066	postgres	20	0	7792496	351912	348932	S	19.3	1.1	0:15.62	postgres: postgres chimera-test 127.0.0.1(40942) SELECT
31068	postgres	20	0	7792548	351736	348768	S	19.3	1.1	0:15.63	postgres: postgres chimera-test 127.0.0.1(40946) SELECT
31076	postgres	20	0	7792548	354100	351120	S	19.3	1.1	0:15.65	postgres: postgres chimera-test 127.0.0.1(40962) SELECT
31082	postgres	20	0	7792548	358060	355076	S	19.3	1.1	0:15.67	postgres: postgres chimera-test 127.0.0.1(40974) SELECT
31085	postgres	20	0	7792548	354636	351660	S	19.3	1.1	0:15.65	postgres: postgres chimera-test 127.0.0.1(40980) idle in transaction
31086	postgres	20	0	7792548	356300	353320	S	19.3	1.1	0:15.50	postgres: postgres chimera-test 127.0.0.1(40982) SELECT
31089	postgres	20	0	7792548	351996	349020	R	19.3	1.1	0:15.59	postgres: postgres chimera-test 127.0.0.1(40988) SELECT
31100	postgres	20	0	7792556	355064	352084	S	19.3	1.1	0:15.63	postgres: postgres chimera-test 127.0.0.1(41010) SELECT
31039	postgres	20	0	7793104	354112	350964	R	18.9	1.1	0:15.51	postgres: postgres chimera-test 127.0.0.1(40990) BIND
31041	postgres	20	0	7793164	348012	344864	R	18.9	1.1	0:15.65	postgres: postgres chimera-test 127.0.0.1(40991) SELECT
31043	postgres	20	0	7793104	350088	346924	S	18.9	1.1	0:15.61	postgres: postgres chimera-test 127.0.0.1(40992) SELECT
31044	postgres	20	0	7793104	353500	350348	S	18.9	1.1	0:15.63	postgres: postgres chimera-test 127.0.0.1(40990) SELECT
31046	postgres	20	0	7793104	350364	347212	R	18.9	1.1	0:15.56	postgres: postgres chimera-test 127.0.0.1(40904) idle in transaction
31047	postgres	20	0	7793104	356932	353788	R	18.9	1.1	0:15.60	postgres: postgres chimera-test 127.0.0.1(40906) SELECT
31050	postgres	20	0	7793104	362412	359252	S	18.9	1.1	0:15.59	postgres: postgres chimera-test 127.0.0.1(40912) SELECT



- Two main gaps to fill
  - Space allocation
  - Tape operation
- Two alternatives to replace
  - User and Group based Quota system
  - WLCG tape recall API



- **Quota  $\neq$  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?

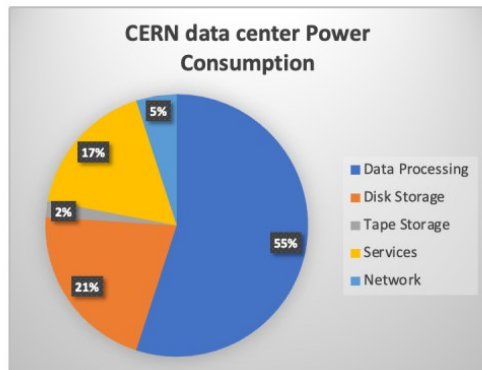


## WLCG data centers power consumption

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



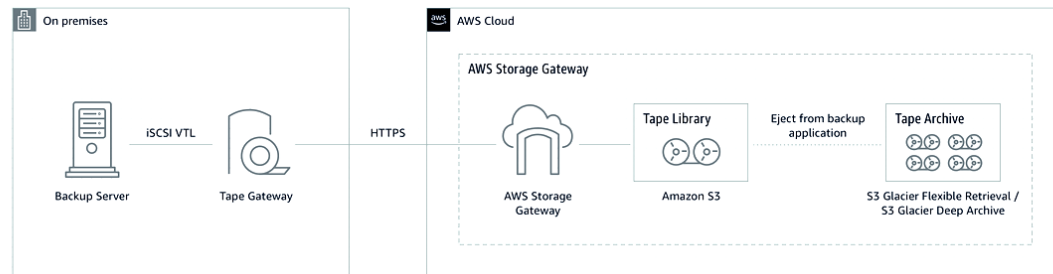
Shameless stolen from Simone Campana

9/11/2022

### Tape Gateway

### AWS Snowball with Tape Gateway

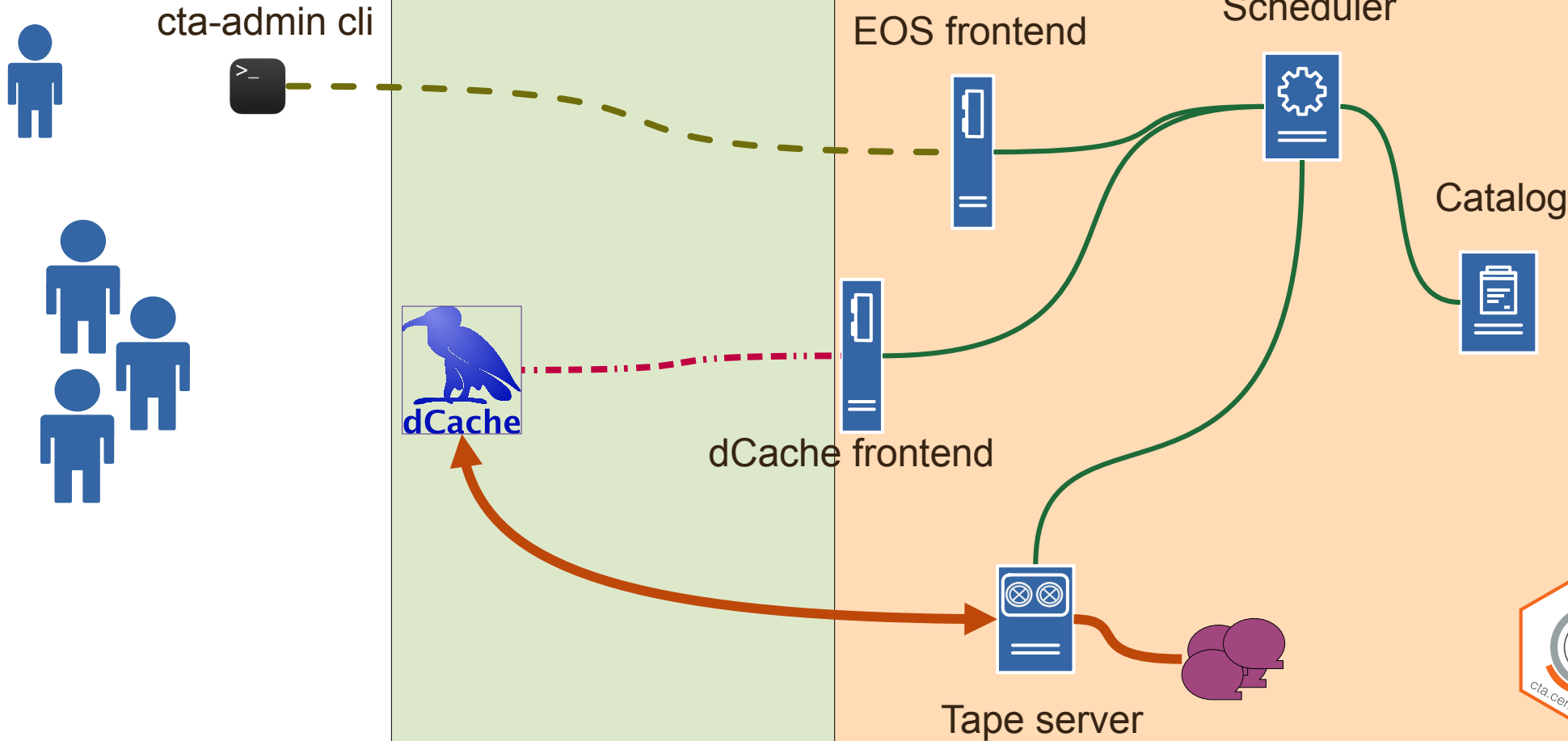
Back up and archive on-premises data to virtual tapes on AWS using your network.



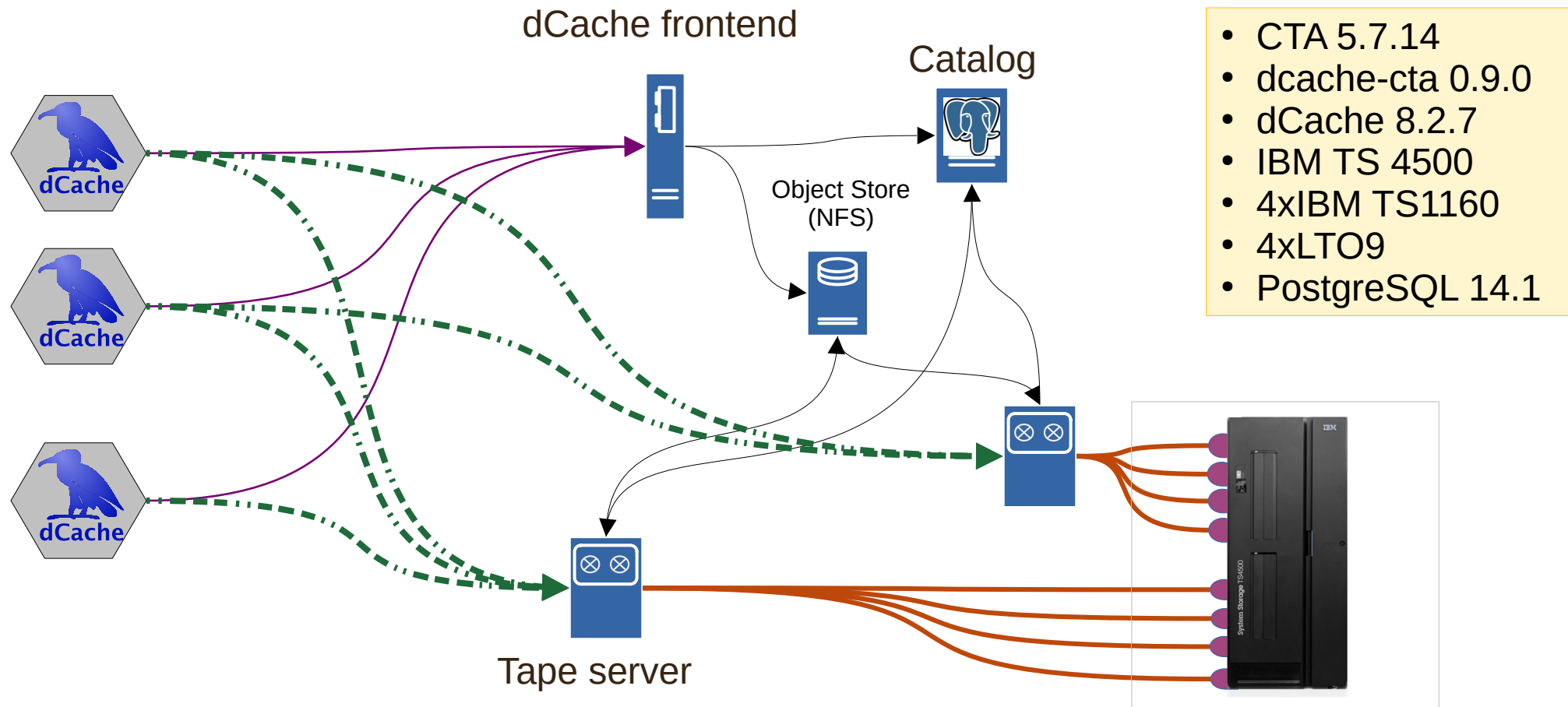
Use Tape Gateway to replace physical tapes on premises with virtual tapes on AWS—reducing your data storage costs without changing your tape-based backup workflows. Tape Gateway supports all leading backup applications and caches virtual tapes on premises for low-latency data access. It compresses your tape data, encrypts it, and stores it in a virtual tape library in Amazon Simple Storage Service (Amazon S3). From there, you can transfer it to either Amazon S3 Glacier Flexible Retrieval or Amazon S3 Glacier Deep Archive to help minimize your long-term storage costs.



# Integration with CTA



# Production Deployment at DESY





- 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)



Thursday, June 1, 2023

Bulk v2 & WLCG Tape API Redux

**HELMHOLTZ**

RESEARCH FOR  
GRAND CHALLENGES1/



# dCache and WLCG

Albert

## 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.


### The Bulk service and WLCG TAPE API support in dCache

Authors: ALBERT ROSSI, Dmitry Livshinsev (FNAL); Svenja Meyer, Paul Millar, Tigran Mkrtchyan, Lea Morschel, Marina Sahakyan (DESY); Krishnaveni Chitrappu (NSC)

**Staging via REST API and Bulk service**

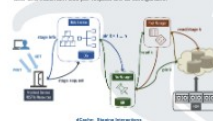
As part of the WLCG collaboration data staging, dCache will receive requests for the TAPE REST API allowing for discovery, read of data from tape. dCache implements the standard responses to the Frontend service, relaying the requests to a general-purpose Bulk service for handling. Currently, the result of a request is also returned as a REST that queries to the Frontend which relays the information from the Bulk service. Available attributes: filename, number of requests per user and maximum files per request are all configurable.

**REST API: SWAGGER**  
(https://example.org:8880/swagger/)



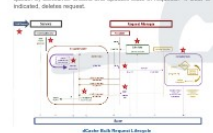
The dCache Frontend provides a SWAGGER page which describes the REST API and allows the user to try out the various resources.

**WLCG TAPE: A special case of the generic Bulk request**



**Lifecycle of a Bulk request**

- 1) Service receives request message.
- 2) Service issues unauthenticated request, updates message.
- 3) Consumer receives and requests from scheduler.
- 4) Request submitted to handler, which creates and starts consumer job. Job is submitted to manager, which handles using the activity's thread executor.
- 5) Job processes request targets by requesting locations of individual processing resources (e.g. individual), performing activity, processing completion on callback, saving and/or updating.
- 6) Consumer receives completed job from queue.
- 7) Consumer checks and updates state of requests. If clear is indicated, updates request.



**SWAGGER**

```

GET /bulk/requests
POST /bulk/requests
GET /bulk/requests/{id}
DELETE /bulk/requests/{id}
  
```

**WLCG TAPE**

```

GET /tape/requests
POST /tape/requests
GET /tape/requests/{id}
DELETE /tape/requests/{id}
  
```

© 2018 Fermilab. Bulk (v2) request (2018) on WLCG TAPE stage request (draft)

FERMILAB PROTON DATA CENTER

This work is produced by Fermi Research Alliance, LLC under Contract No. DE-AC02-09OR21400 with the U.S. Department of Energy. Publicly accessible under the DOE Public Access Plan. DOE Public Access Plan

Fermilab National Accelerator Laboratory Fermilab ENERGY

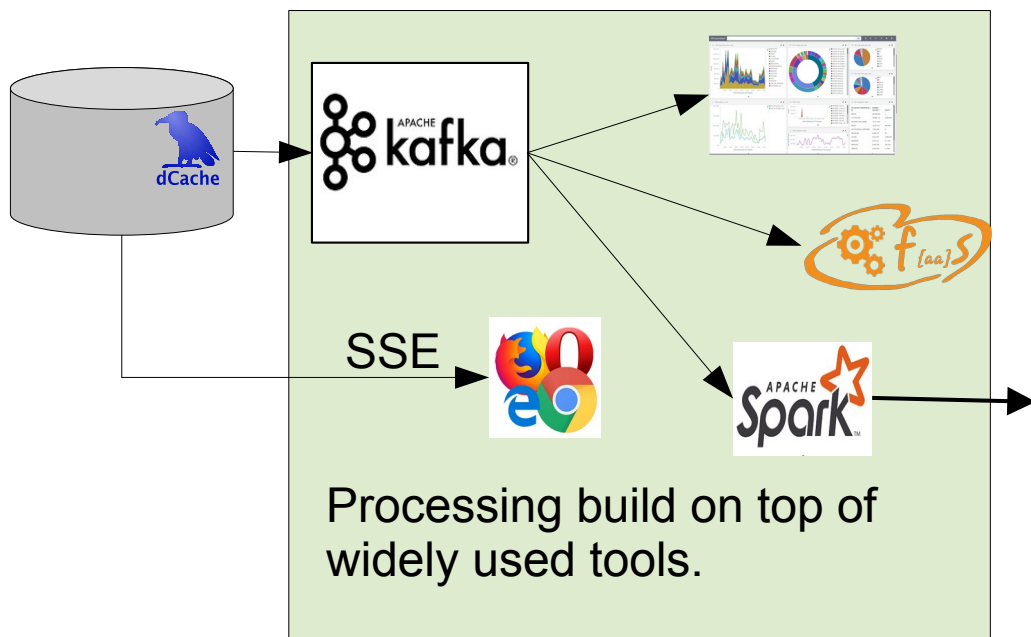
Thursday, June 1, 2023

Bulk v2

Thursday, June 1, 2023

Bulk v2 & WLCG Tape API Redux

# 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



## 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



*By: Mattias Wadenstein*



## 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







## 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 agency 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.org>

## Storage operations

- Local site admins maintain hardware, filesystem, operating system, networking, kernel tuning
  - Provides one unprivileged 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



## Services

- 32 PB disk installed
  - Tier-1 pledged disk +
    - (Including some purchases 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>

## 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



## Operations

re, filesystem,  
nel tuning  
of storage to the

Is

kes cooperation  
error (investigation by  
ps fix)

SPEAKER | Mattias Wadenstein <maswan@ndgf.org>

12





## 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*

*By: Fabio Hernandez*

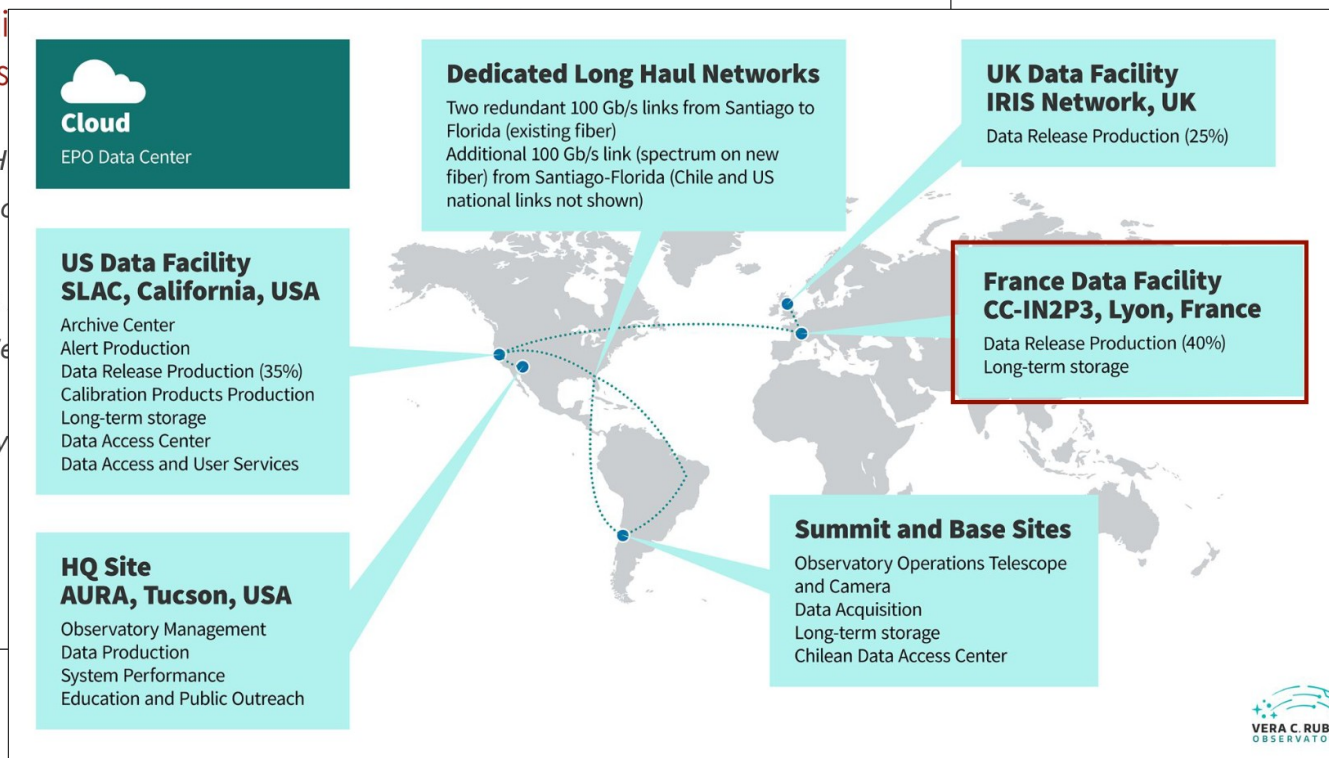
# dCache for Rubin LSST



## HISTORY OF dCache AT CC-IN2P3

- Started getting familiar with dCache
- Currently operating services
  - **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  
1.4 PB, 7 servers, 50 M objects
  - **NESSIE** [v8.2.16]  
for R&D purposes, currently in production  
3 servers, 300 TB
  - **Rubin LSST**  
the subject of this talk

georget | hernandez | le boulic'h



# dCache for Rubin LSST



## HISTORY OF dCache AT CC IN2P3

- Started getting files from the cloud
- Currently operating for the Rubin LSST
- **LHC** [v8.2.16]  
shared by ATLAS, CMS  
38 PB, 157 servers, 15000 files
- **EGEE** [v7.2.27]  
shared by CTA, Juno, ...  
1.4 PB, 7 servers, 50000 files
- **NESSIE** [v8.2.16]  
for R&D purposes, currently not used  
3 servers, 300 TB
- **Rubin LSST**  
the subject of this talk



**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)

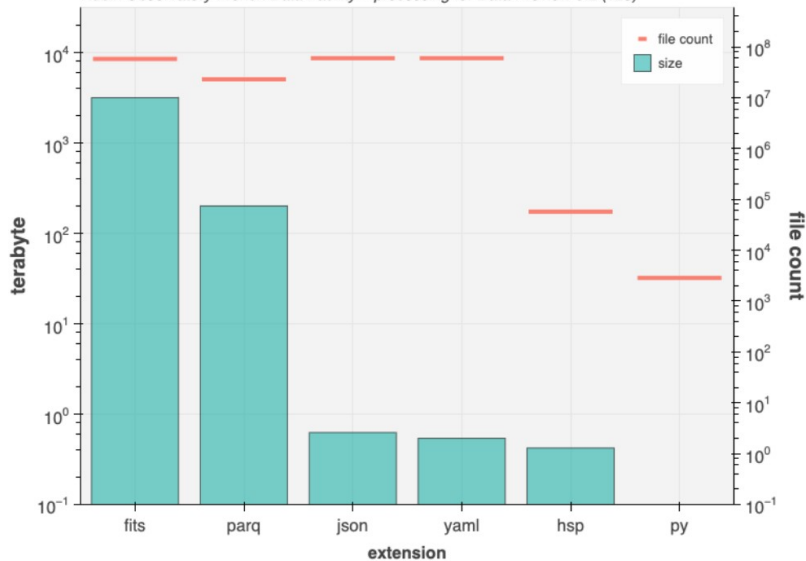
### UK Data Facility IRIS Network, UK

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 boulic'h

georget | hernandez | le boulic'h

CCIN2P3 16

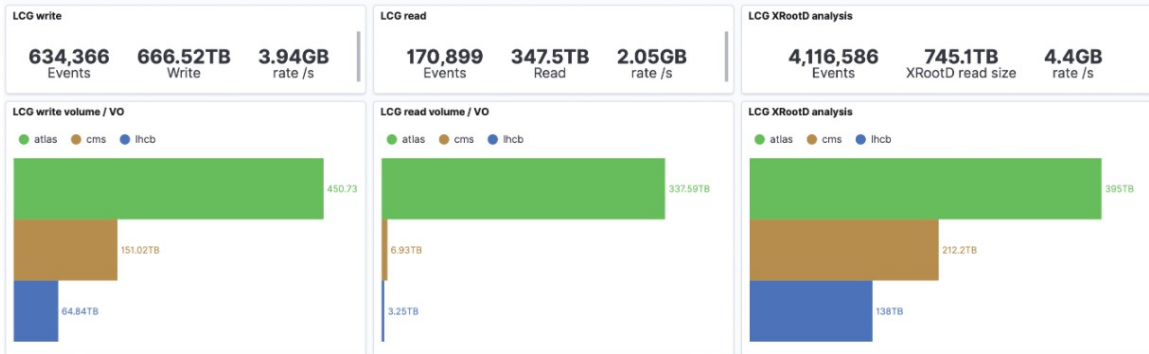
# dCache for Rubin LSST



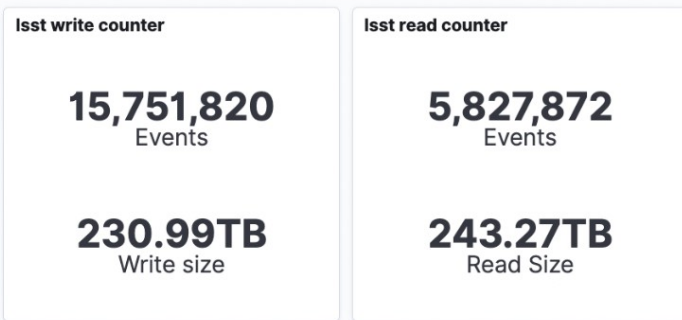
## HISTORY OF dCache AT CC IN2P3

- Started getting...
- Currently op...
- LHC [v8.2.16] shared by ATLAS 38 PB, 157 servers
- EGEE [v7.2.27] shared by CTA 1.4 PB, 7 servers
- NESSIE [v8.2.16] for R&D purposes 3 servers, 300...
- Rubin LSST the subject of...

## 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 boulic'h

georget | hernandez | le boulic'h

georget | hernandez | le boulic'h

CCIN2P3 22

CCIN2P3 16

# dCache at BNL

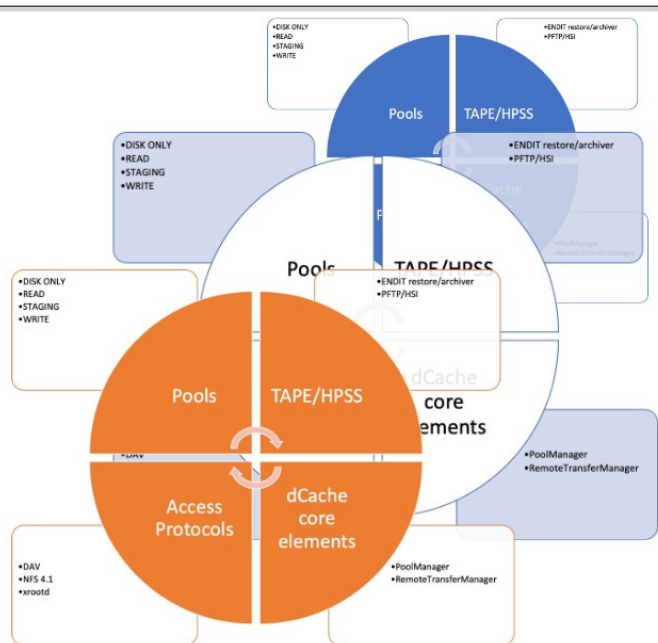


dCache.org

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



Storage technology flexible to support SC individual requirements

8



## Towards an Improved dCache Operation

### Areas of work:

- Enhancing software for interaction among dCache and TAPE HPSS systems
  - [ENDIT](#) archiver/retriever
- Improving dCache data access workflows for client access
  - Non firewalled Xrootd client access for write/read
  - DUAL IPv4/IPv6 dCache application stack configuration
- Extending monitoring for dCache operations
- Evolving dCache along with infrastructure



# dCache at BNL



dCache.org

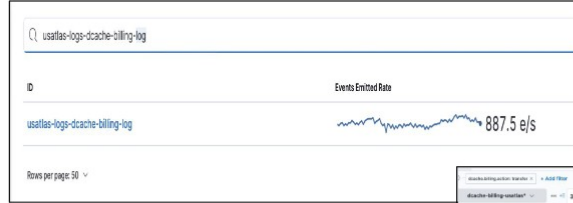
distributed s

## Towards an Improved

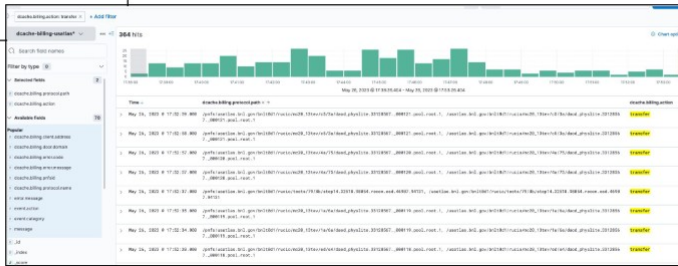
dCache insta-  
per SC

## 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



## 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)

File information							
PNFSID	type	inode	size	uid	gid	mtime	flag
000F405A8D545E4D7	DIR	408	4096	91162	3269109070		1

Locations				
Location	type	status	update	lastsync
dCache-9	DIR	ONLINE	2022-10-04 05:12:17	2022-10-04 05:12:17

Locations				
number	type	priority	update	lastsync
139423714	1	10	2022-10-04 05:12:17	2022-10-04 05:12:17

### Feature driven dashboards

Home	Home
Overview	Overview
File Transfer	File Transfer
Location Transfer	Location Transfer
Location Details	Location Details
Location	Location
Show Details	Show Details
Storage Info	Storage Info
Storage Info Details	Storage Info Details
Storage Info	Storage Info
Storage Info Details	Storage Info Details
Transfer Details	Transfer Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details
Transfer Info	Transfer Info
Transfer Info Details	Transfer Info Details

### Performance of dCache



## migration

9

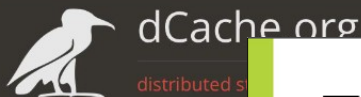
# dCache at BNL



## Monitoring Enhancement

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

Feature driven dashboards



### Tow

dCache installed per SC

### dCache and ELK Started to be Used

Filebeat / Logstash pipelines enabled for dom

usafes-logs-dcache-billing-log

ID Events Enabled

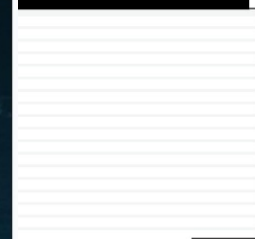
usafes-logs-dcache-billing-log

Rows per page: 50

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



1



Performance of dCache



Time	Event	Message	Severity
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO
May 24, 2023 @ 17:02:58.890	usafes-logs-dcache-billing-log	usafes-logs-dcache-billing-log: [INFO] ...	INFO

9

# OIDC and Token-based Access



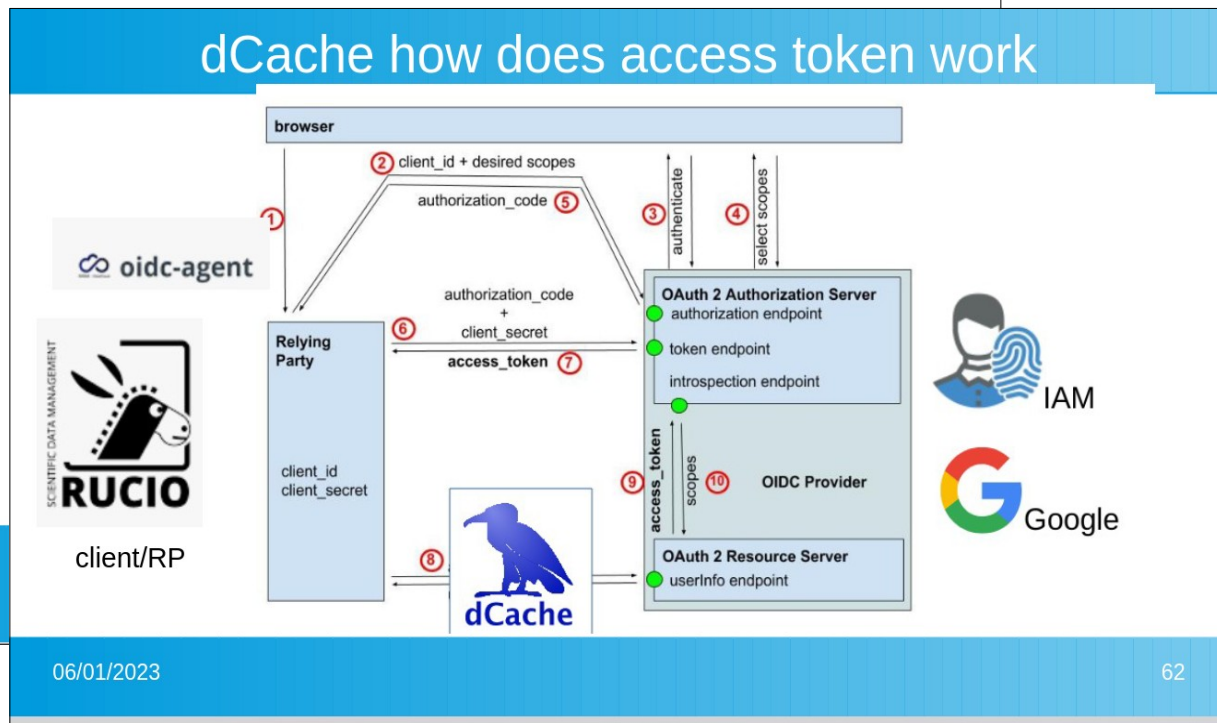
OIDC and all that Jazz

06/01/2023

1

*By: Marina Sahakyan*

# OIDC and Token-based Access



06/01/2023

06/01/2023

62



## dCache how does access token work



06/01/2023

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

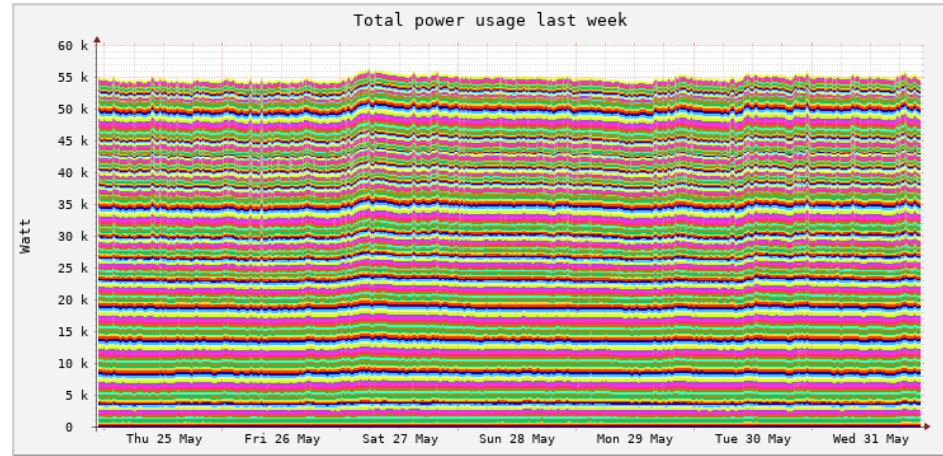
19

# Sustainability

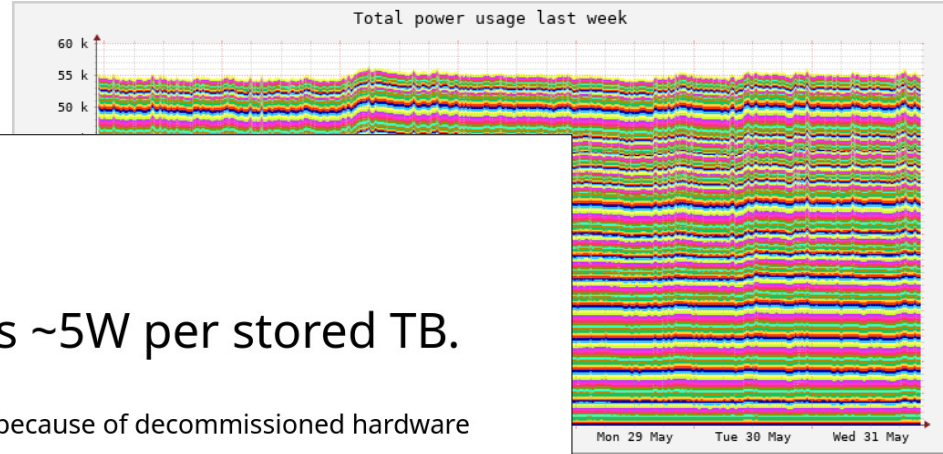


*By: Onno Zweers*

# Sustainability



By: Onno Zweers



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





O

At th

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](mailto:support@dCache.org), [user-forum@dCache.org](mailto:user-forum@dCache.org)

***Developers:***

[dev@dCache.org](mailto:dev@dCache.org)



- 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>