



dCache Project Updates

dCache user workshop, CC-IN2P3, Lyon

Tigran Mkrtchyan for the dCache collaboration



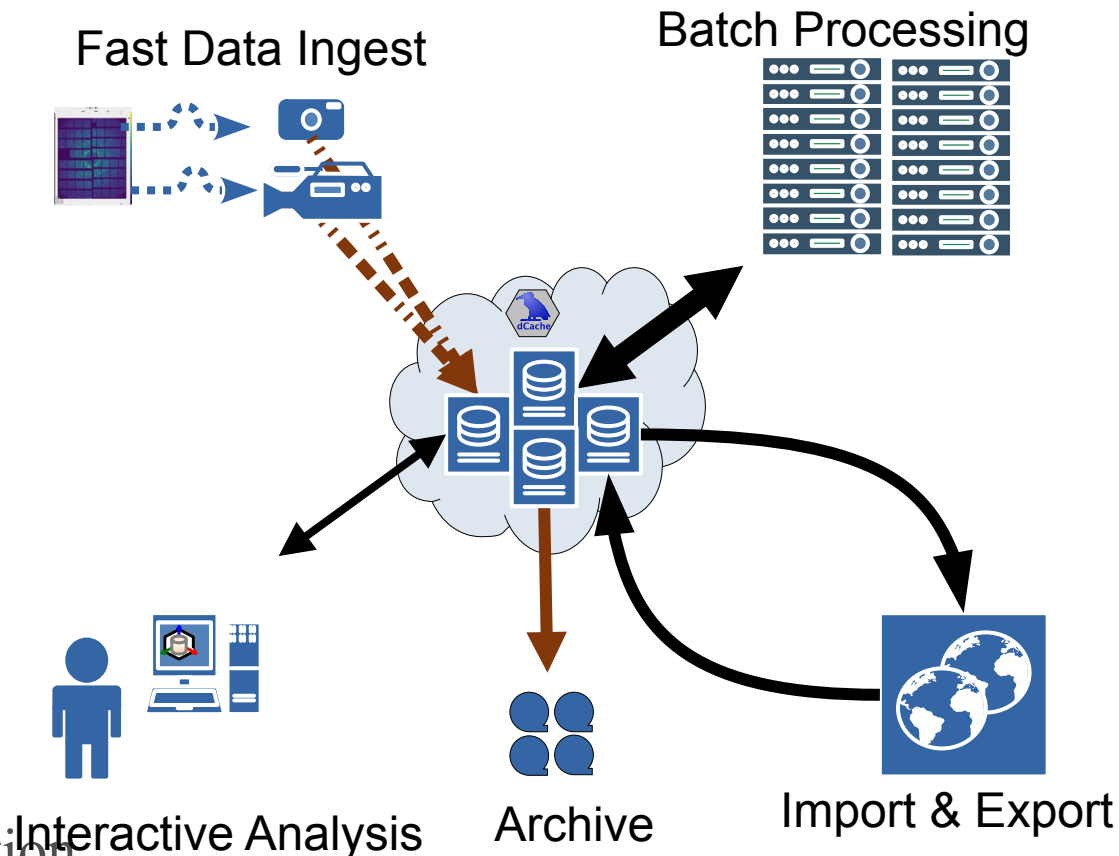
HELMHOLTZ

RESEARCH FOR
GRAND CHALLENGES

Mission Statement



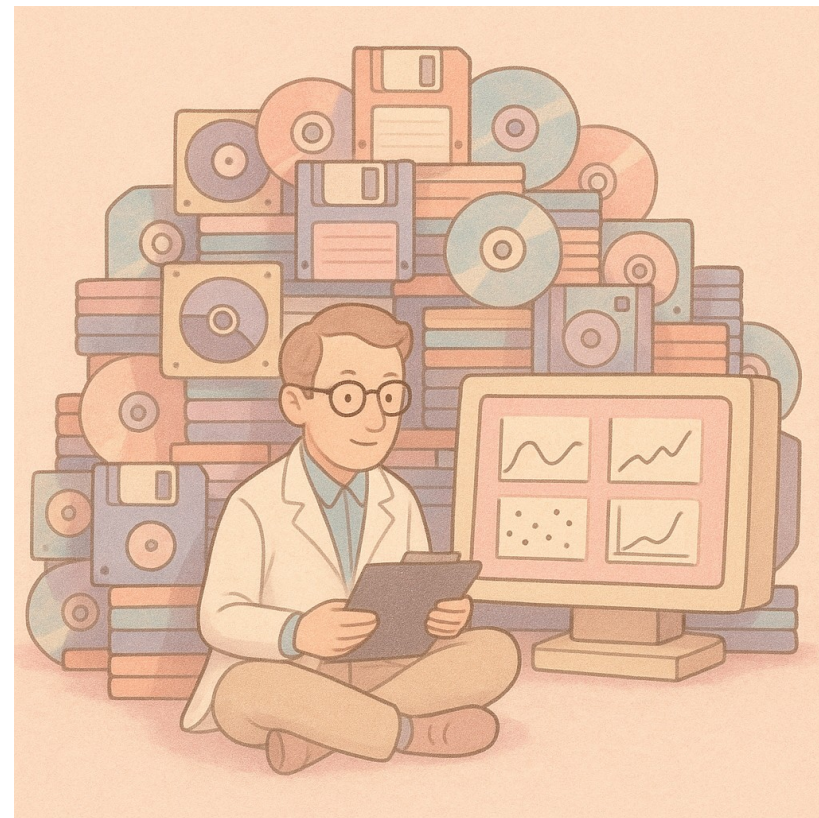
- Ingest
 - Multiple parallel streams.
 - High data rate.
 - Large number of files.
- Analysis
 - High CPU efficiency.
 - Chaotic user access.
 - Standard access protocols.
- Sharing & Exchange
 - Effective WAN access.
 - In-flight data protection.
 - Federated Identity handling.
- Long-Term Archival
 - High reliability.
 - Automatic technology migration.



Challenges



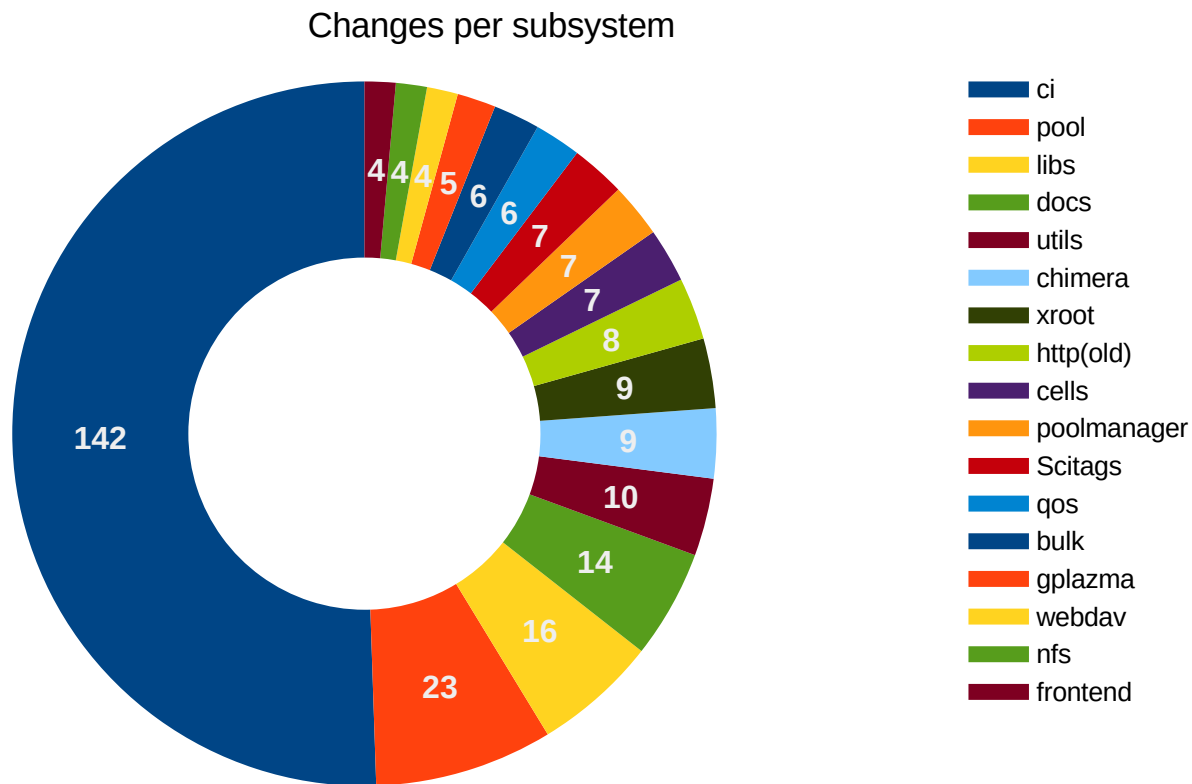
- Data is going to grow... A lot...
 - High ingest data rates
 - More movements between sites
- Shared Computing Resources
 - Analysis Facilities
 - Grid Farms
 - HPC
 - Cloud resources (CPU&Storage)
- Standard analysis tools
 - ROOT
 - Jupyter Notebooks
 - Non-ROOT tools
- Competing Tape Operations



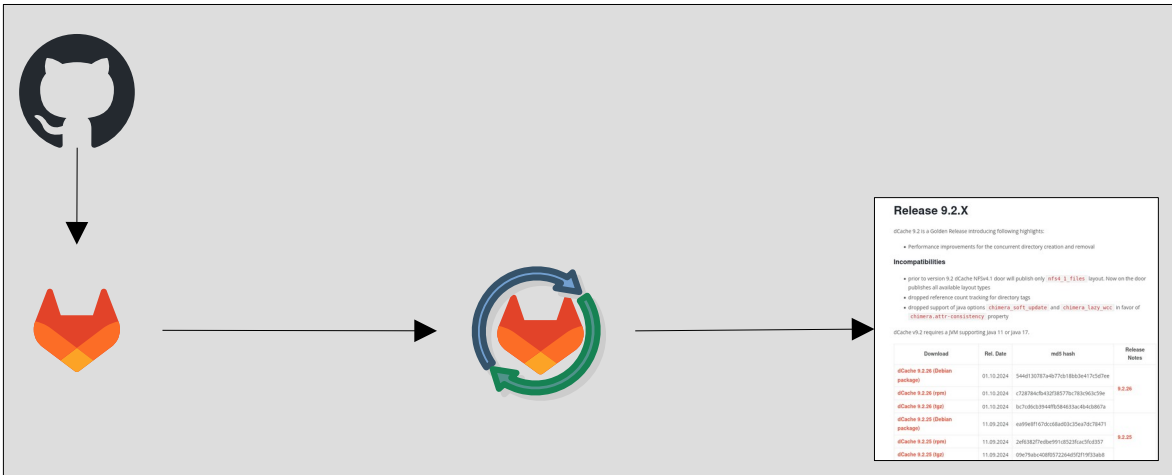


- Left 🙄
 - Svenja – other project by DESY
 - Lea – Looking for an attractive Ph.D. project
 - Krishnaveni (NDGF) → Industry
- Join 🎉
 - Darren Starr (NDGF)

Where the Effort Goes?



-
- The diagram illustrates the release process of iCache 9.2.X. It begins with the GitHub logo, followed by a cartoon character in a tuxedo holding a document, and then a box titled "Release 9.2.X" containing release highlights, incompatibilities, and a table of release notes.
- Release 9.2.X**
- iCache 9.2.X is Golden Release introducing following highlights:
- Performance improvements for the concurrent directory creation and removal
- Incompatibilities**
- pin to version 9.2 of iCache Refact 1 door will publish only `ifex_3_7346` input. Note on the door publisher of available input types
 - dropped reference count tracking for directory logs
 - dropped support of pin system, `idbname`, `idfs`, `idtype` and `idbname`, `idfs`, `idtype` in favor of `idbname`, `idfs`, `idtype`, property
- iCache v9.2 requires a JVM supporting Java 11 or Java 17.
- | | Download | Rel. Date | hash | md5 hash | Release Notes |
|-------------------------------|------------|----------------------------------|------|----------|---------------|
| iCache 9.2.0 (Golden package) | 01.10.2024 | 5446f303833480750830461750576e | | | |
| iCache 9.2.0 (api) | 01.10.2024 | c12f56a9a3a37a85776a70b30a30a30a | | | 9.2.0 |
| iCache 9.2.0 (api) | 01.10.2024 | bc70d85399595086833a6a8a0807a | | | |
| iCache 9.2.0 (Golden package) | 11.02.2024 | 6a99f8f7376c68a083325a7c76767671 | | | |
| iCache 9.2.0 (api) | 11.02.2024 | 2a9f33827c0b9f91f0522c3a303037 | | | 9.2.0 |
| iCache 9.2.0 (api) | 11.02.2024 | 69676da9a088722a6a707f0f0f0f0f | | | |





ci: disable unsupported pynfs tests

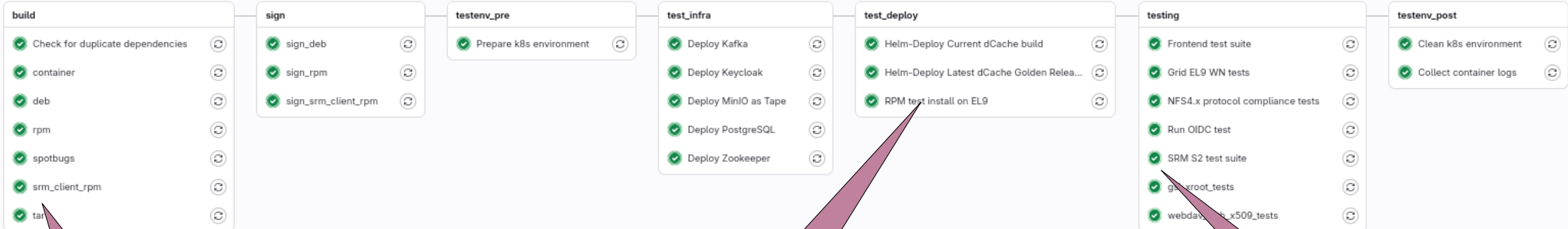
✓ Passed guthub-sync created pipeline for commit 6814424a 6 days ago, finished 6 days ago

For master

latest branch 28 jobs 34 minutes 49 seconds, queued for 3 seconds

Pipeline Jobs 28 Tests 5717

Group jobs by Stage Job dependencies



Did we brake
it for devs?

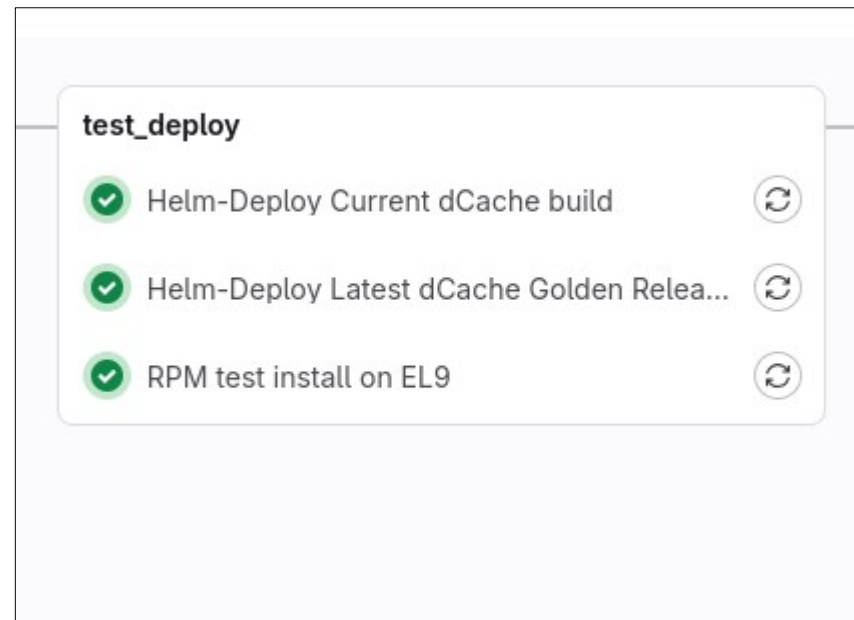
Did we brake
it for admins?

Did we brake
it for users?

Test Deployment



- Full dCache deployment for a typical site
- Version-mix for backward compatibility test



Still Tested Manually



- Kernel NFS I/O
 - fio, mdtest, xfs-tests
- DB schema migration
- HA, Fail-over
- ... (your specific use case)
 - Share your *curl command*





- Updated possibility for custom checks
 - `pool.check-health-command=/my/custom/check`
 - `0` → OK, `1` → Read-Only, else → Disable
- Direct I/O for scrub jobs
- max space as as percent of filesystem
 - `set max diskspace 94%`
- ``st/rh ls <pnfsid>``
- ``st/rh kill`` can be issued on ‘canceled’ requests



- Yet another mechanism to monitor data movement between sites
- Propagate client-supplied SciTag to fireflies
- dCache sends (UDP) fireflies on transfer begin/end
- The experiment ID can be enforced by config



```
pool.enable.firefly=false
```

```
pool.firefly.destination=<transfer's endpoint>
```

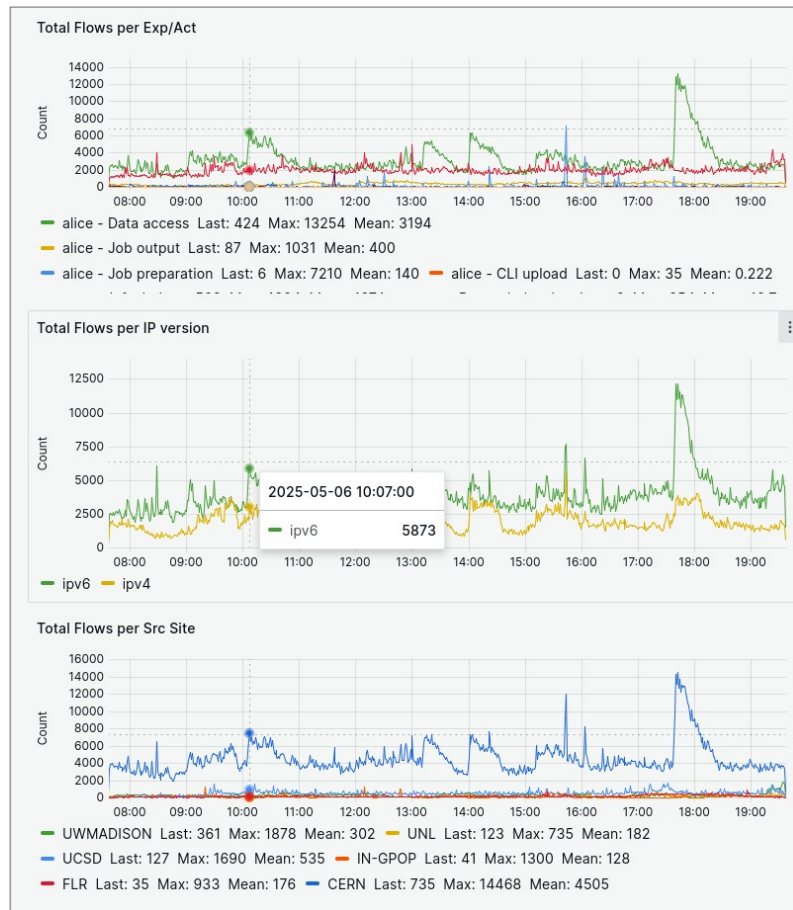
```
pool.firefly.excludes=a.b.c.d/16
```

```
pool.firefly.vo-mapping=atlas:2, cms:3, lsst:9
```

```
pool.firefly.storage-statistics=false
```



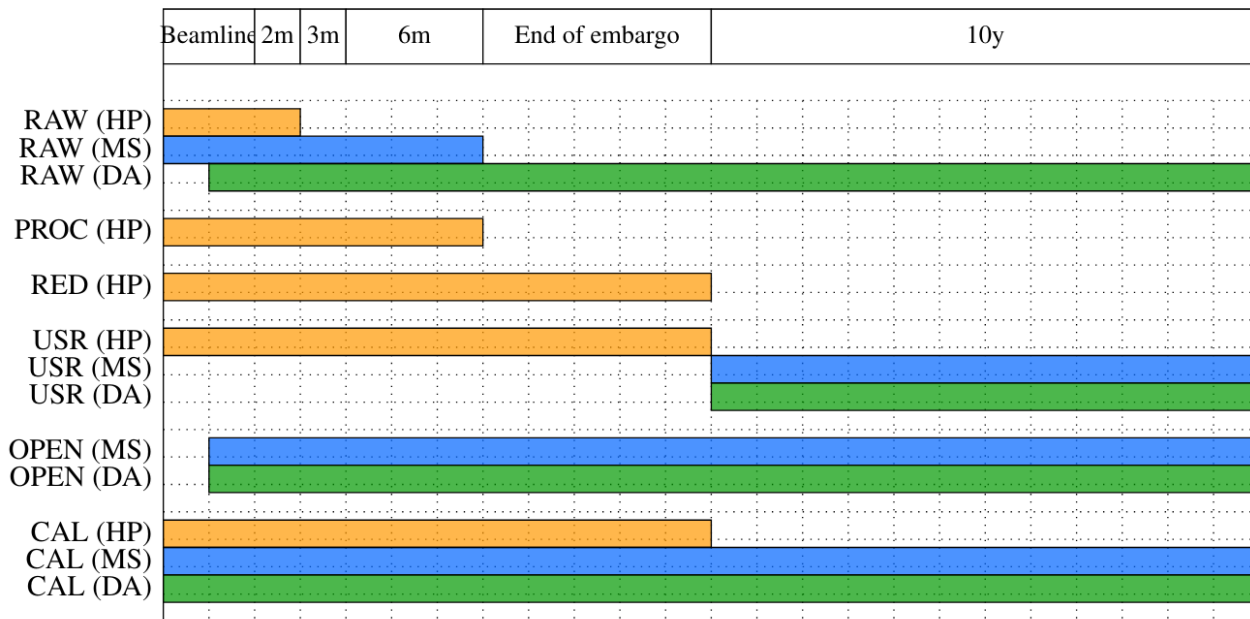
```
"version": 1,
"flow-lifecycle": {
  "state": "start", ...
},
"flow-id": {
  "afi": "ipv6", "protocol": "udp",
  "src-ip": "fe80::1", "src-port": 1234,
  "dst-ip": "fe80::2", "dst-port": 56789
},
"context": {
  "experiment-id": 3333,
  "activity-id": 12,
  "application": "grid-ftp v99.99"
}
```



Policy driven QoS system



```
"name": "xfel-raw-policy",  
"states": [  
  {  
    "duration": "P6M",  
    "media": 1x DISK, 1x  
  },  
  {  
    "media": 2x HSM  
  }  
]
```



Srouce: https://www.xfel.eu/users/policies/index_eng.html

"HSM-only" on DISK



- What if has QoS policy is "HSM-only" and the file is still on DISK?
 - QoS policy will be reflected in the namespace
 - dCache will reject READ access
 - Extra QoS transition by admin required to make it accessible



- Admin friendly.
- Quick way to prototype site-specific mapping/authentication.
- Dynamically loaded.

```
def py_map(principals):  
    if "user:tigran" in principals:  
        principals.add("uid:1234")  
        return True  
    else:  
        return False
```

JFR: Door Message Timing



```
org.dcache.door.Messages {  
    startTime = 09:45:35.672  
    duration = 37.8 ms  
    message = "PnfsGetFileAttributes"  
    Source = "NFS-nairi@dCacheDomain"  
    Destination = "PnfsManager"  
    eventThread = "proc-9" (javaThreadId = 73)  
}
```

-

perf top



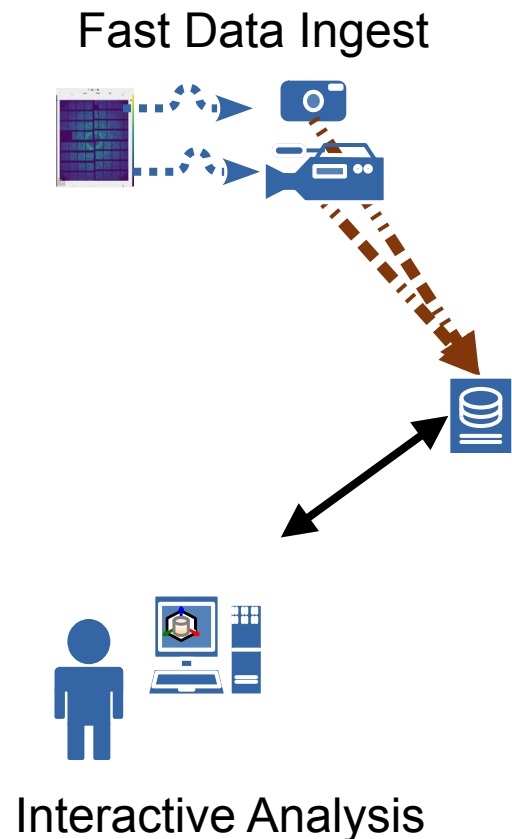
Samples: 68K of event 'cycles:P', 4000 Hz, Event count (approx.): 23084479093
lost: 0/0 drop: 0/0

Overhead	Shared Object	Symbol
3.83%	[JIT] tid 960615	[.] vtable chunks
3.21%	libjvm.so	[.] _ZNK17HeapRegionManager8reservedEv
2.25%	libjvm.so	[.] _ZN9MemRegionC2EPP12HeapWordImplS2_
1.79%	libjvm.so	[.] _ZN12SpaceMangler13mangle_...
1.16%	[JIT] tid 960615	[.] java.security.ProtectionDomain[]
1.13%	[JIT] tid 960615	[.] ch.qos.logback.core.spi.FilterReply
1.06%	[JIT] tid 960615	[.] StubRoutines (2)
0.97%	libjvm.so	[.] _ZNK12MemAllocator25allocate_...
0.59%	libjvm.so	[.] _ZNK23G1BiasedMapped...

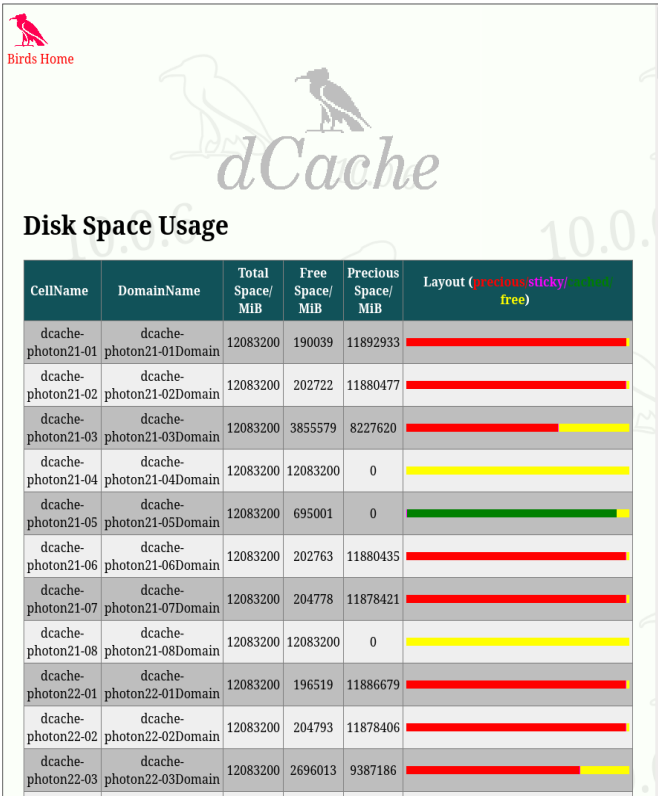
Tunable Namespace Consistency



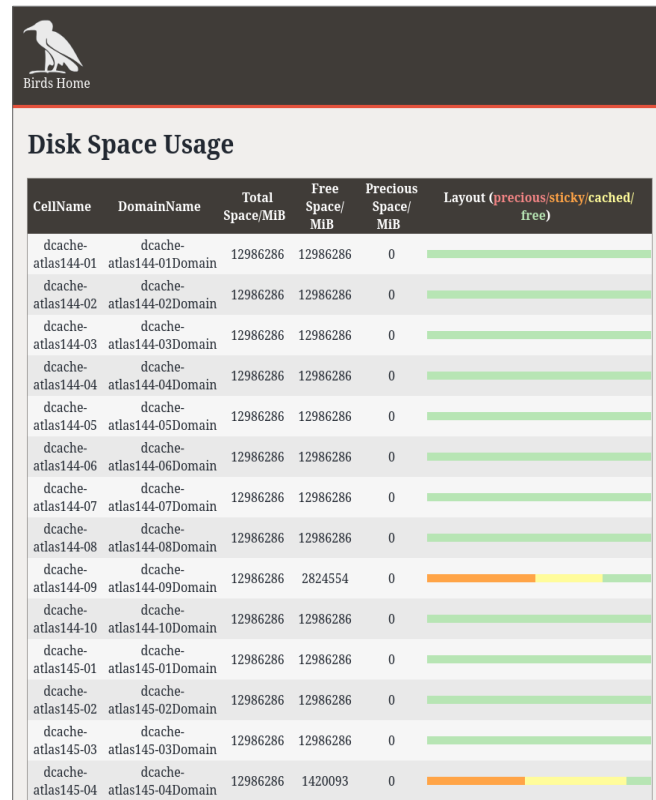
- **STRONG** (default in ≤ 11.0)
A creation of a filesystem object triggers simultaneous update of parent directory's mtime, ctime, nlink and generation attributes.
- **WEAK**
A creation of a filesystem object will eventually update (after 30 seconds) parent directory's mtime, ctime, nlink and generation attributes. Multiple concurrent modifications to a directory are aggregated into single attribute update.
- **SOFT** (default in $> 11.1?$)
Same as weak, but, reading of directory attributes takes into account pending attribute updates.



New Look and feel



VS.



Propose your style! Colors in: `/usr/share/dcache/httpd/static/styles/common.css`

Labeling



- Add custom labels to existing file
- Labels dynamically create file collections
- Collections exposed as virtual-directories
- Files can be accessed by regular path
- Files can be accessed by collections



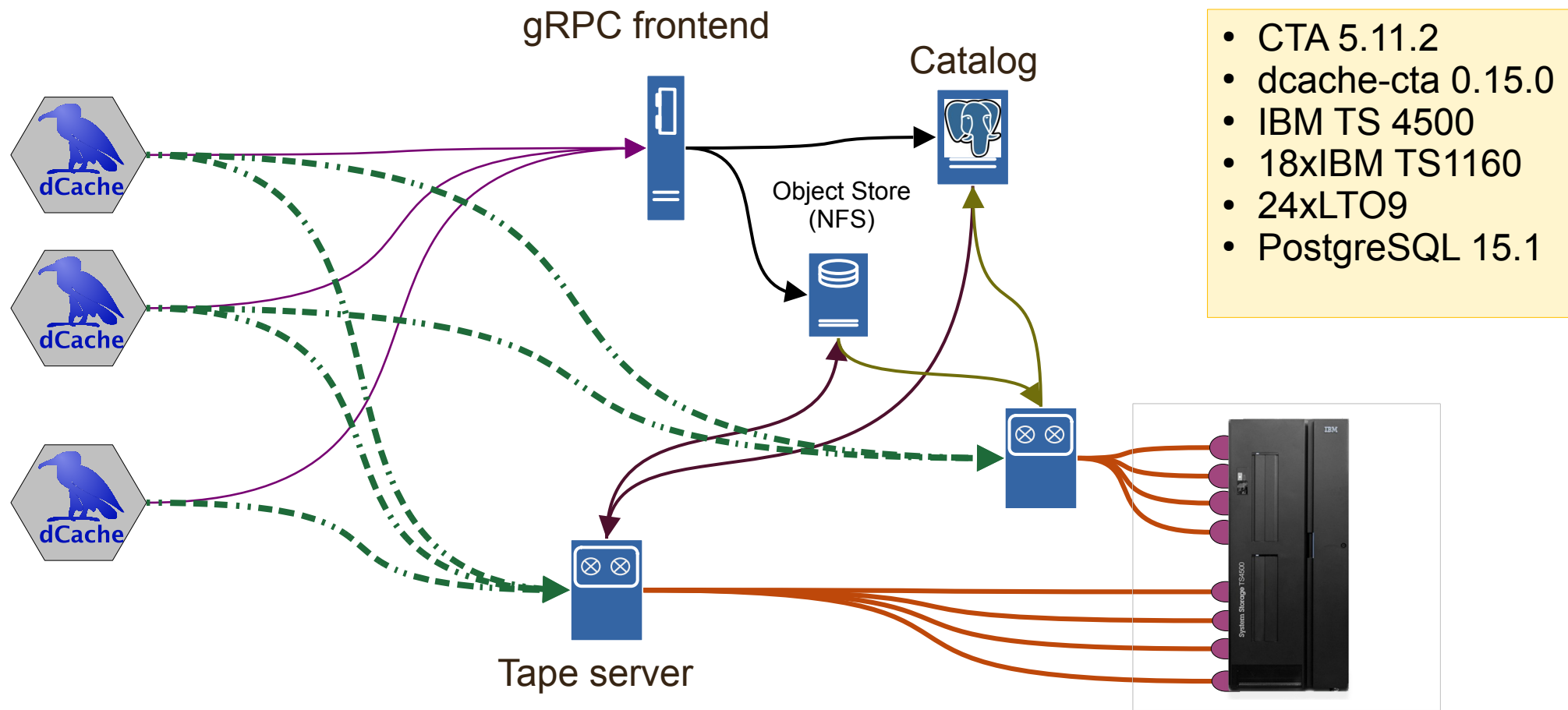
#Anna Bertha Ludwig

#hand

#1895

#x-ray

CTA Integration



Tokens carrying **storage.* OR compute.*** scopes are explicit AuthZ tokens and exempt from any further permission checks!

The Mapping and Access Handling



```
"wlcg.ver": "1.0",  
"sub": "2ea41db8-ecd7-47a1-8f86-7ac97ebcc466",  
"aud": "https://wlcg.cern.ch/jwt/v1/any",  
"nbf": 1662472763,  
"scope": "openid profile storage.read:/ storage.create:/",  
"iss": "https://cms-auth.web.cern.ch/",  
"exp": 1662476363,  
"iat": 1662472763,  
"jti": "88a1f764-8311-4210-a8cd-d5604b942d8c",  
"client_id": "106ceb94-e89c-421a-9c7d-1424e2fa4668",  
"wlcg.groups": [  
  "/cms",  
  "/cms/dcms",  
  "/cms/integration"  
]
```

Shameless stolen from Onno

<https://indico.desy.de/event/43473/contributions/167034/>

- Better HPC workload support
 - Read-delegations (details later today)
- NFS-over-TLS
 - `mount -o xprtsec=tls nfs-lab:/exports /mnt`



- Persistent port number per protocol on pools
- `storage.stage` scope support
- Environment variables passed as dCache properties.
- Json format for billing files (no need for kafka).
- Retry read request on 'pool up' for stages

Breaking Changes



- `nfs4_1_files` layout type dropped
 - (remove from export files `lt=nfsv4_1_files`)
- Removed transfermanager DB
 - Was never read :)
- Failing request in PoolManager will cancel stage request
- Files with QoS policy "HSM-only" will reject reads event if on disk



SRM... Are We Done?



- Should not run by Tier-2s
 - No tape
- Should not run by Tier-1s
 - Tape REST-API

gPlazma modules to go...



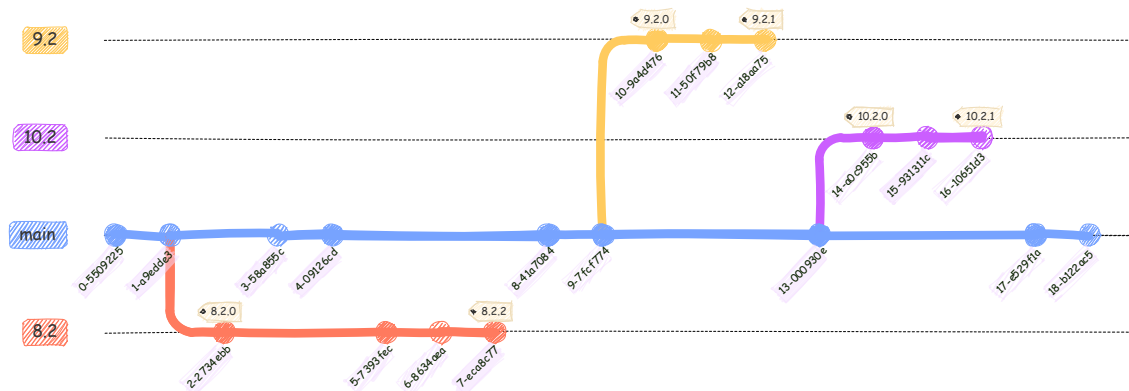
- ~~argus~~
 - nis
 - nsswitch
- } $\Rightarrow$ ldap
- gridmapfile
 - vorolemap
 - kpwd
- } $\Rightarrow$ multimap



dCache versions



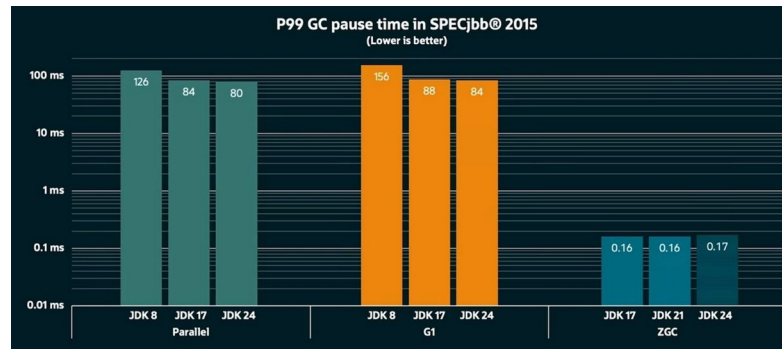
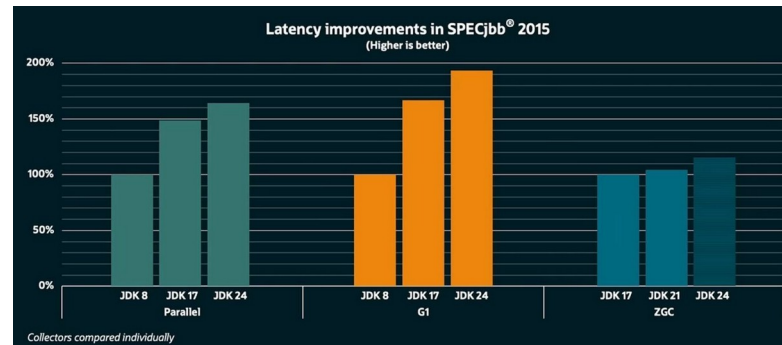
- 9.2 In Maintenance mode
 - Security fixes only
 - Support till fall 2025
- 10.2 Current **Golden Release** (LTS)
 - Since Oct. 2024
 - Support till fall 2026
- 11.x Series started



Java Dependencies



- 17 is required for $\geq$ dCache-10.1
- 21 not required
 - Used in our CI and containers
- 25 recommended/required ~11.2/12.0
 - Depends on our adoption speed and availability in distros



<https://inside.java/2025/05/14/javaone-garbage-collection/>

Technical Directions



- Scaleout
 - Namespace.
 - Number of pools (SW/HW).
- BULK operations
- QoS
- Token-based Authentication
- Better *Analysis Facility* support
 - POSIX access and compliance.
 - HPC workload support (DDoS protection).
- Tape integration





- You can contribute with ...
 - Code
 - Configuration
 - Testing
 - HW setup
 - Knowledge
 - Hosting Events
- You can make dCache visible with ...
 - Sharing your use case
 - Demonstrate dCache use in various events



Thank you!



Questions?



- Mimic a small site
 - Tape library
 - MinIO
 - IdP
 - Keycloak
 - X509-CA
 - PostgreSQL
 - ZooKeeper
 - Kafka
 - Worker nodes

