

dCache News & Roadmap

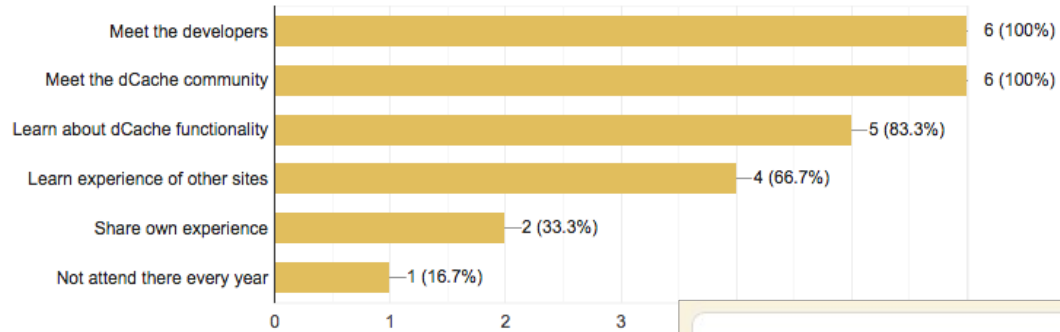
Tigran Mkrtchyan for



Nordic e-Infrastructure
Collaboration

Agenda

The reasons of workshop participation



If there was only one presentation, which title should it have?

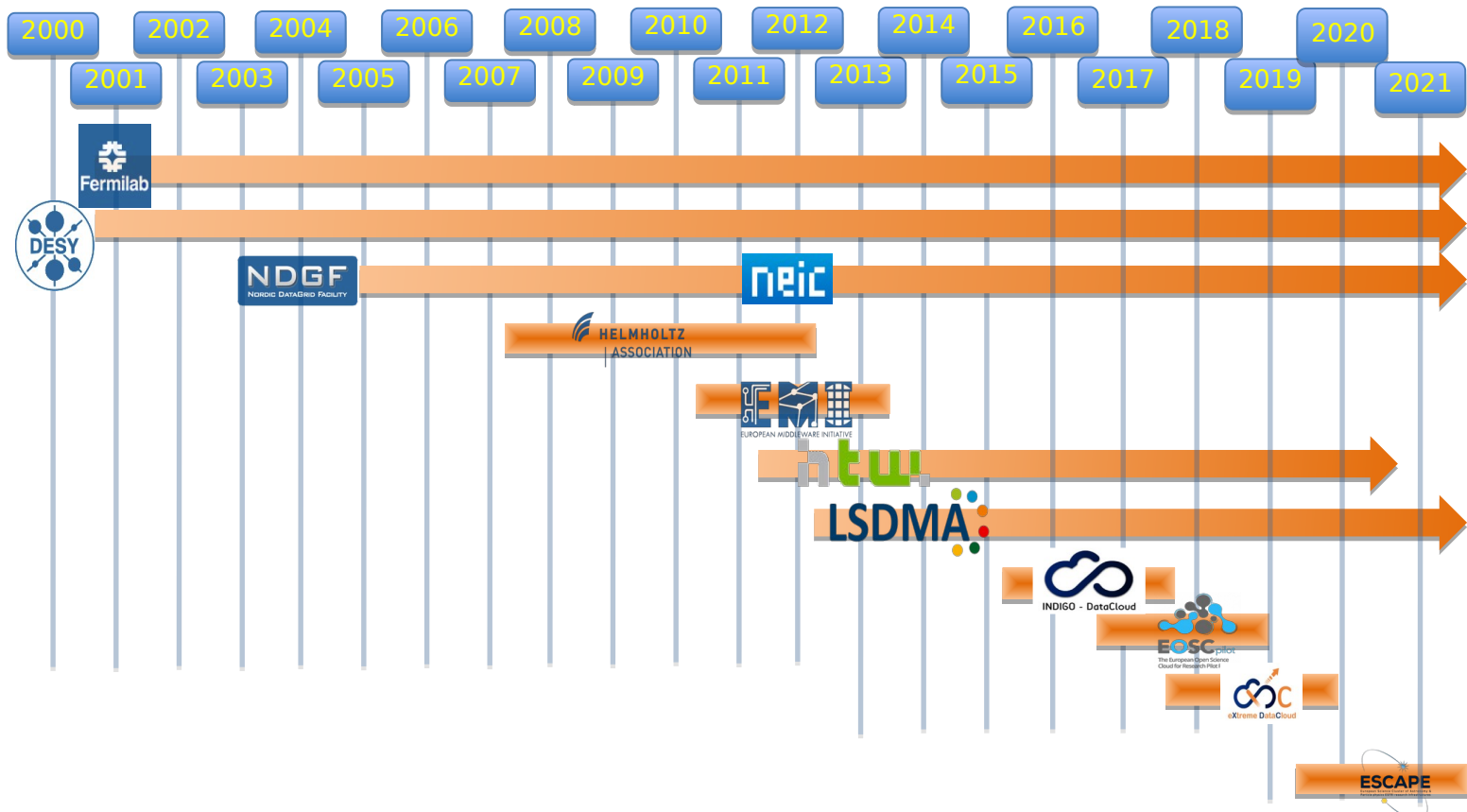
dCache News

Upcoming amazeballs features

Roadmap

Project Fundings & Team

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



- WLCG / HEP
 - Almost all sites run dCache for WLCG
- DUNE / FS / XFEL
 - Big local customer and stakeholder
- Astro- Bio- Life-science
 - Almost all sites have on-site communities



High Speed
Data Ingest



Data management
& workflow control

dCache.org

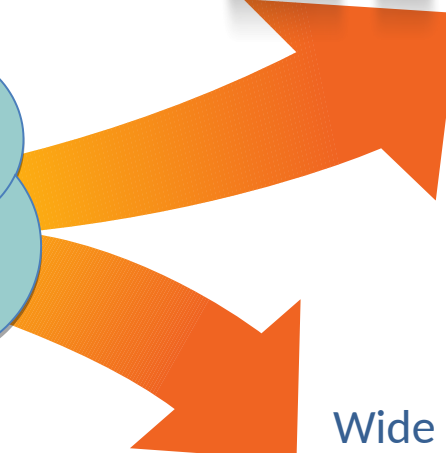
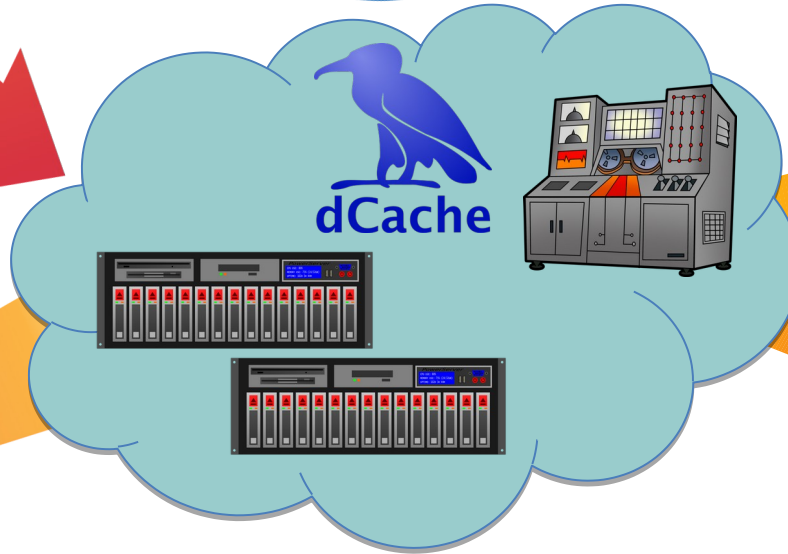
Batch processing



Interactive analysis

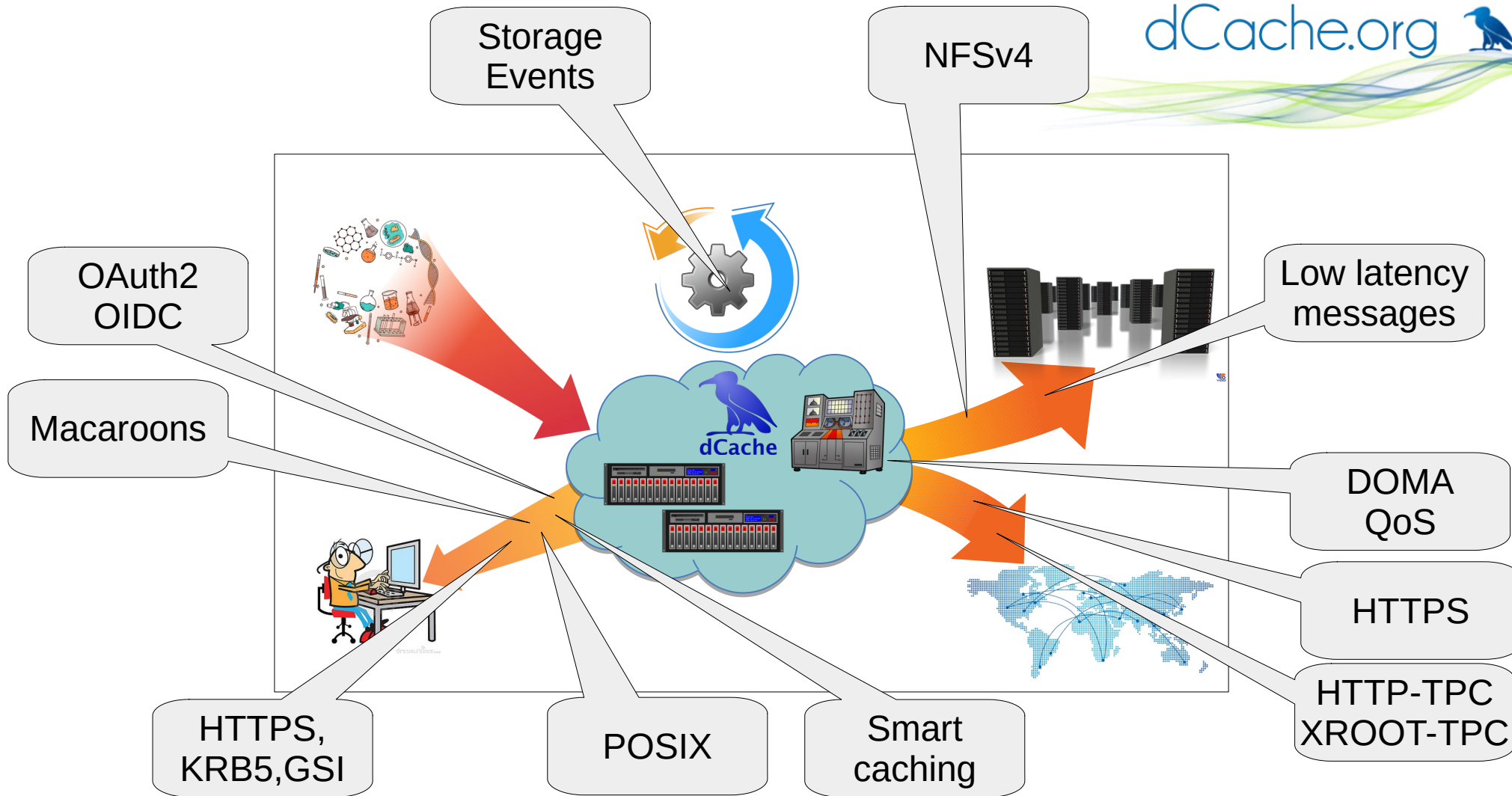


dreamstime.com



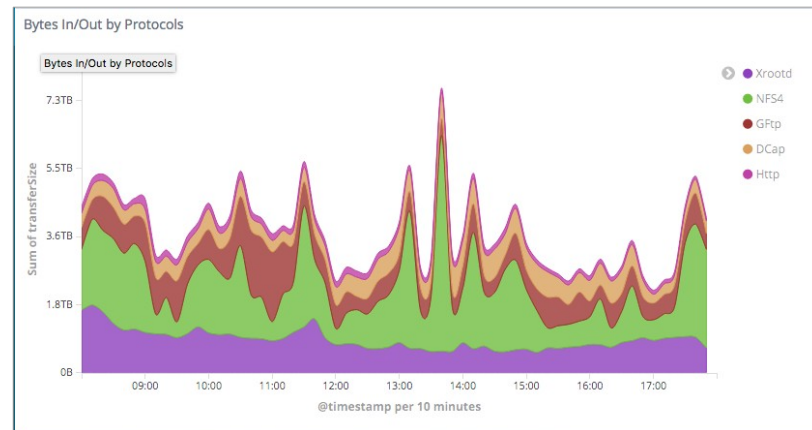
Wide Area Transfers





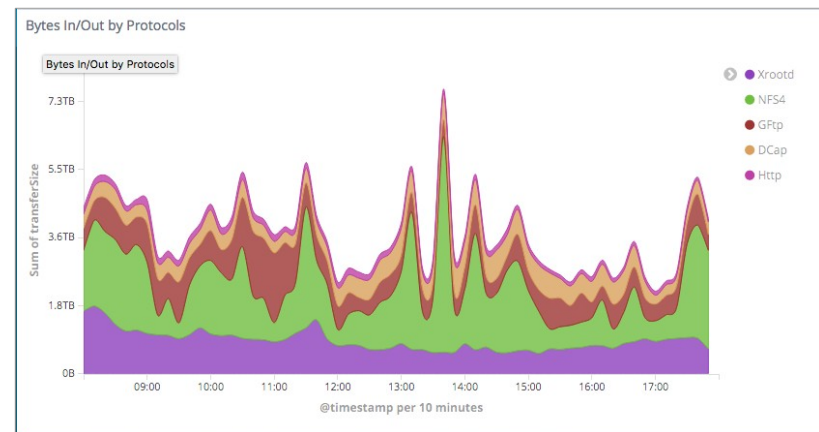
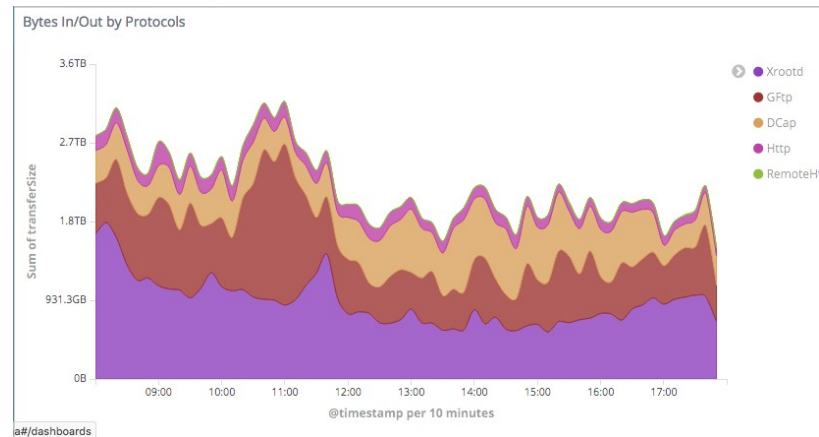
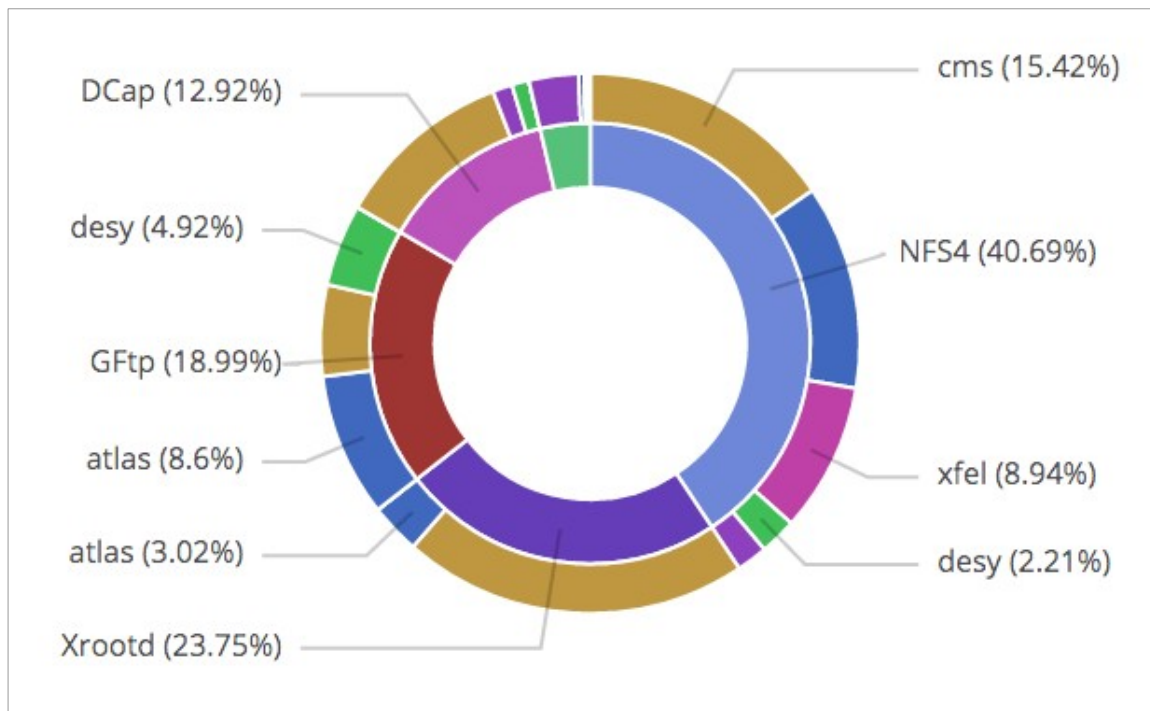
User Workflow Shift

- More non HEP tools and POSIX access
 - ROOT => Jupyter Notebook
 - Apache Spark
 - HDF5
- Growth of interactive analysis
 - Analysis Facilities
- Industry standard AuthN
 - OpenID Connect
 - OAuth2
 - Federated IDP
- Hybrid Clouds
- New 3rd-party transfers protocols
- Integration with HPC clusters

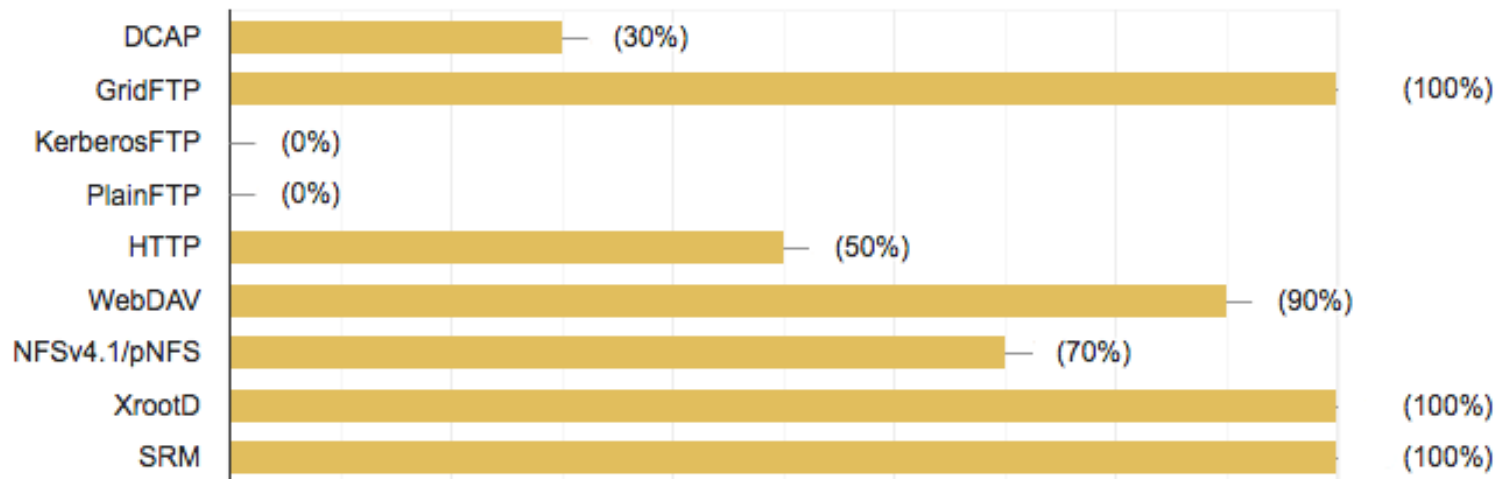


Protocol Usage (DESY-HH)

NFS access dominated by Atlas & CMS



Access protocol used



Breaking changes!

- Recommended version to use
 - We test with OpenJDK, others should work as well
- Required for 6.2
- Some (new) error messages to be ignored

WARNING: An illegal reflective access operation ...

WARNING: Illegal reflective access by ...

WARNING: Please consider reporting this ...

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

Zookeeper

- zookeeper > 3.5.7 for dcache 6.2
 - Better stability
 - Dynamic cluster re-configuration
 - TLS out-of-box
- Embedded zookeeper not affected
 - don't use embedded zk!
- dcache.org rpms

```
[dcache-org-zk]
```

```
name=dCache.ORG zookeeper packages
```

```
baseurl=https://download.dcache.org/nexus/repository/zookeeper-rpms/el7/  
noarch
```

```
gpgcheck=0
```

```
enabled=1
```

PostgreSQL \geq 9.5

- Older versions not supported by PostgreSQL
- Rule of Thumb: use latest version
- Better integration with dCache

~~ERROR: duplicate key value violates unique constraint "t_dirs_pkey"~~

- BerkeleyDBMetaDataRepository is default
- Defacto default configuration at sites
 - Better performance
 - FileSystem friendly
- Adjust config if needed

`pool.plugins.meta=`

`org.dcache.pool.repository.meta.file.FileMetaDataRepository`

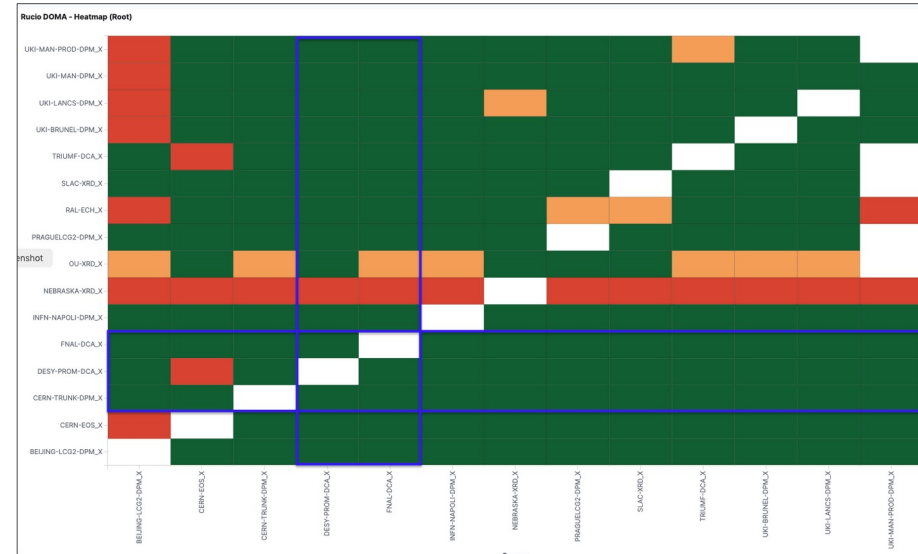
Pool Tags

- Can't use `zone` as pool tag
`pool.tags=zone=my-zone`
- More info later....

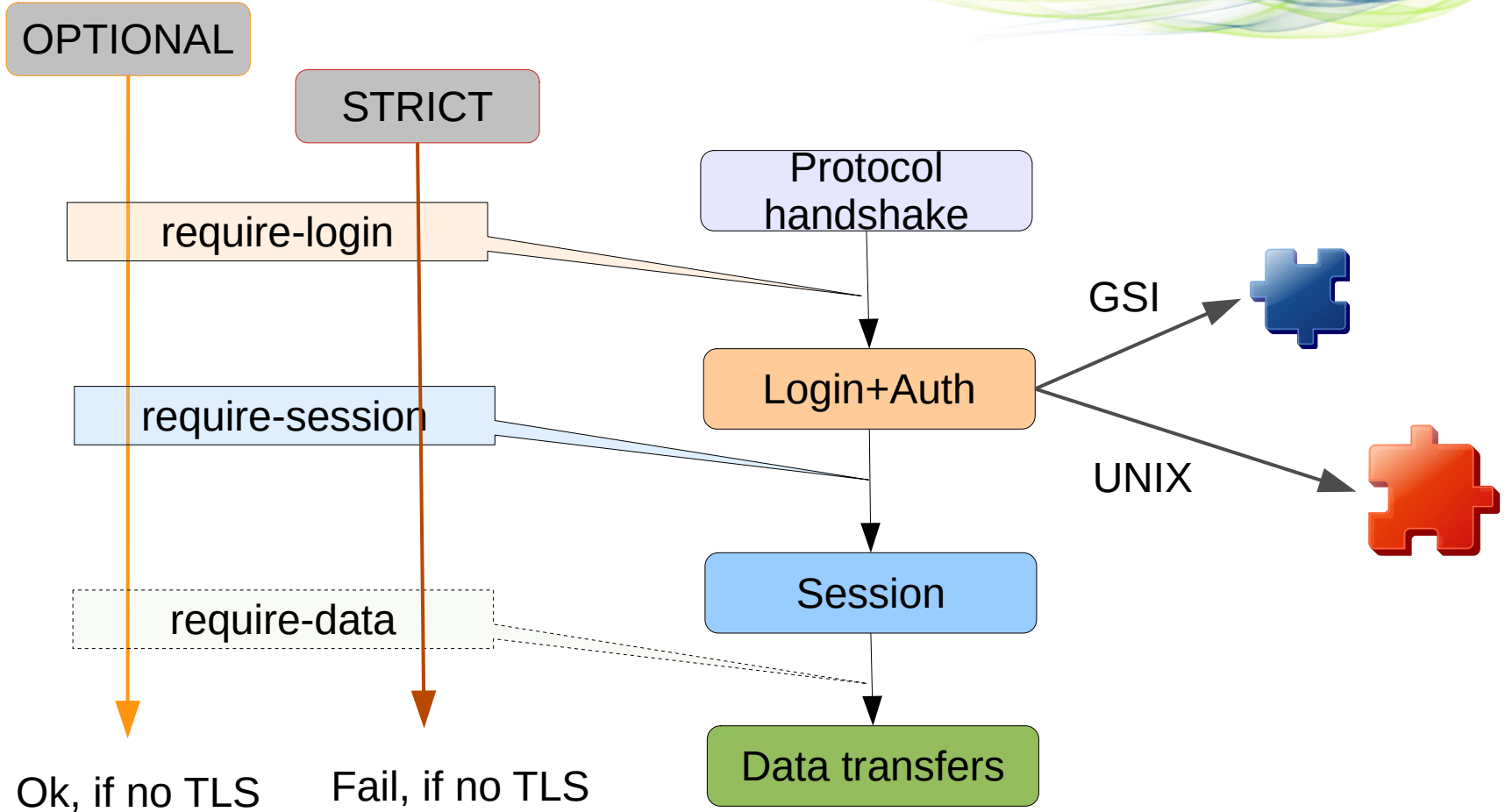
Xrootd & Co.

3rd Party Copy

- XROOTD
 - Source/destination support
 - GSI + delegation, SciToken
 - Inter-op with SLAC xrootd client & server (DPM, EOS)
- HTTP
 - Source/destination support
 - 3rd vendor HTTP server as destination
 - X509, Macaroon and SciToken support
- dCache 5.2.x is the LTS version with all required changes
 - recommended version by DOMA-TPC WG



xrootd.security.tls



- Door

```
xrootd.security.tls.mode=OPTIONAL
```

```
xrootd.security.tls.require-session=true
```

```
xrootd.plugins=gplazma:gsi,authz:none
```

- Pool

```
pool.mover.xrootd.security.tls.mode=OPTIONAL
```

```
pool.mover.xrootd.security.tls.require-tpc=true
```

```
pool.mover.xrootd.tpc-authn-plugins=gsi,unix
```

Rule of Thumb

- GSI
 - mode=**OPTIONAL**
 - required-**session**=true
- SciTokens
 - mode=**STRICT**
 - required-**login**=true

DataLakes, fjords & deserts

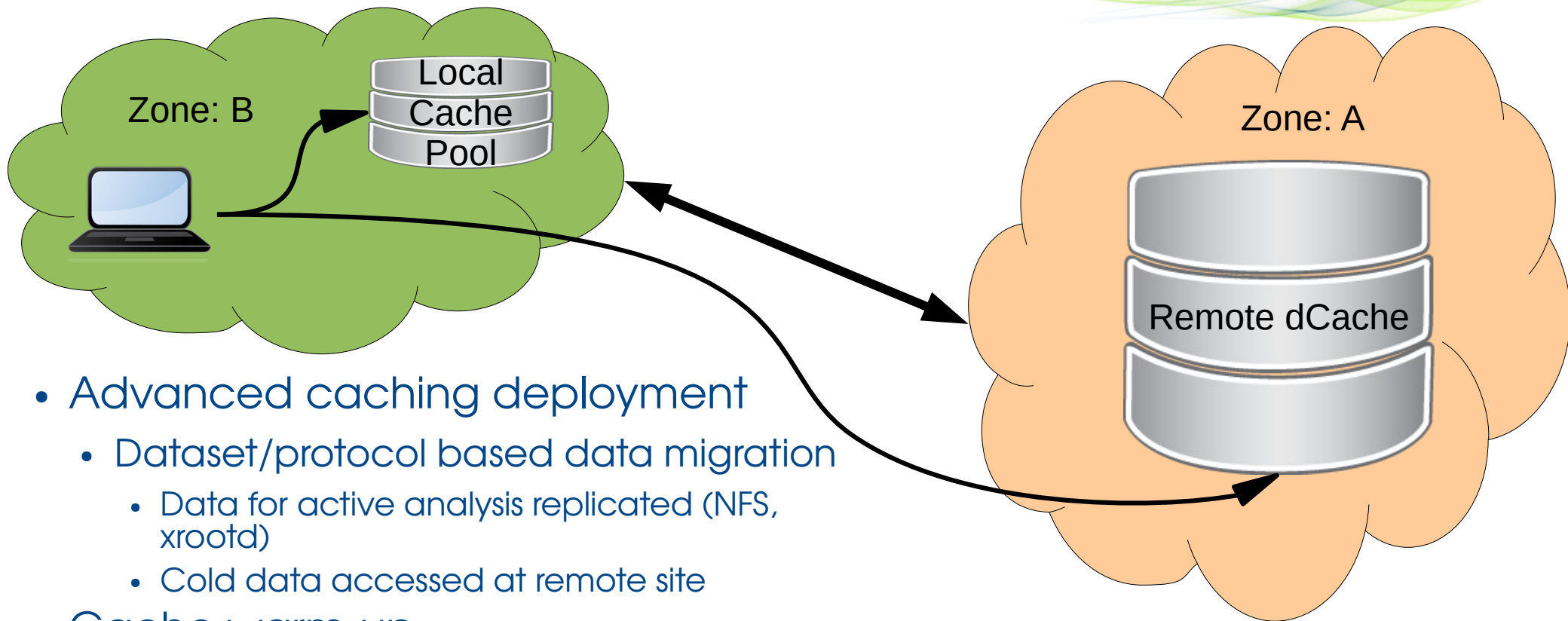
Zones: Geo-Location

- Geo-location aware unit
- Dynamically groups services together
- Available in replication rules
- Network topology aware internal communication
 - *Always prefer local resources*
 - *Disconnected operation*

```
set storage unit data:resilient@osm -required=2 -onlyOneCopyPer=zone
```

```
create pgroup caching-pools -dynamic -tags=zone=A
```

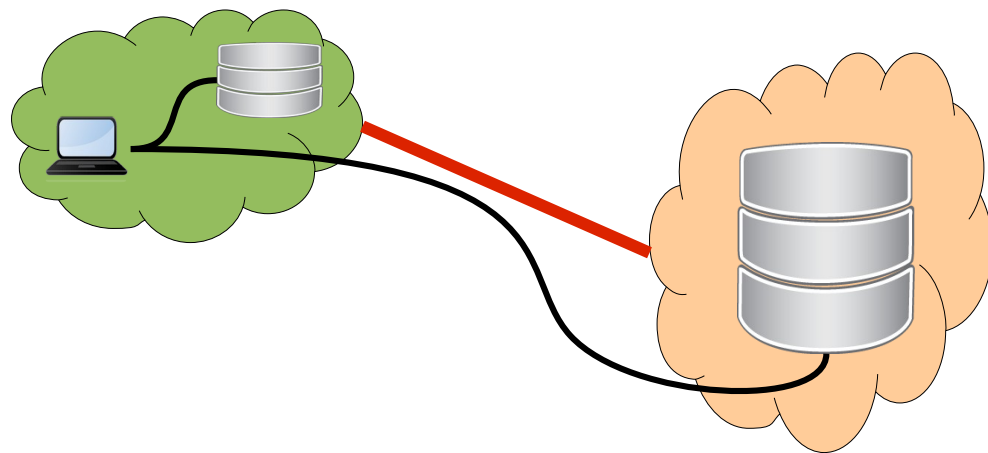
Caching/Cloud Bursting



- Advanced caching deployment
 - Dataset/protocol based data migration
 - Data for active analysis replicated (NFS, xrootd)
 - Cold data accessed at remote site
- Cache warm-up

In-transit Encryption

- HTTPS on redirect (upload/download)
 - Like NFS with krb5i and krb5p
- HTTPS on internal copy
 - Pool-to-pool over WAN
 - *Zone awareness*



Config

```
dcache.zone=extern
```

```
[${host.name}]
```

```
[${host.name}/pool]
```

```
pool.name=pool-${host.name}
```

```
pool.path=/pool
```

Read-only Doors

- Many workloads like pipelines
 - producer != consumer
- Many sites have hot stand-by chimera DB replica
 - **if not – you should!**
- Move read-only load away from master DB

PnfsManager

```
[ro-core- $\{host.name\}$ ]
```

```
dcache.queue.pnfsmanager=PnfsManager-RO
```

```
dcache.service.pnfsmanager=PnfsManager-RO
```

```
[ro-core- $\{host.name\}$ /pnfsmanager]
```

```
pnfsmanager.db.schema.auto=false
```

```
chimera.db.url=jdbc:postgresql:// $\{chimera.db.host\}$ /  
 $\{chimera.db.name\}$ ?
```

```
targetServerType=preferSlave&ApplicationName= $\{$   
 $\{pnfsmanager.cell.name\}$ 
```

Door

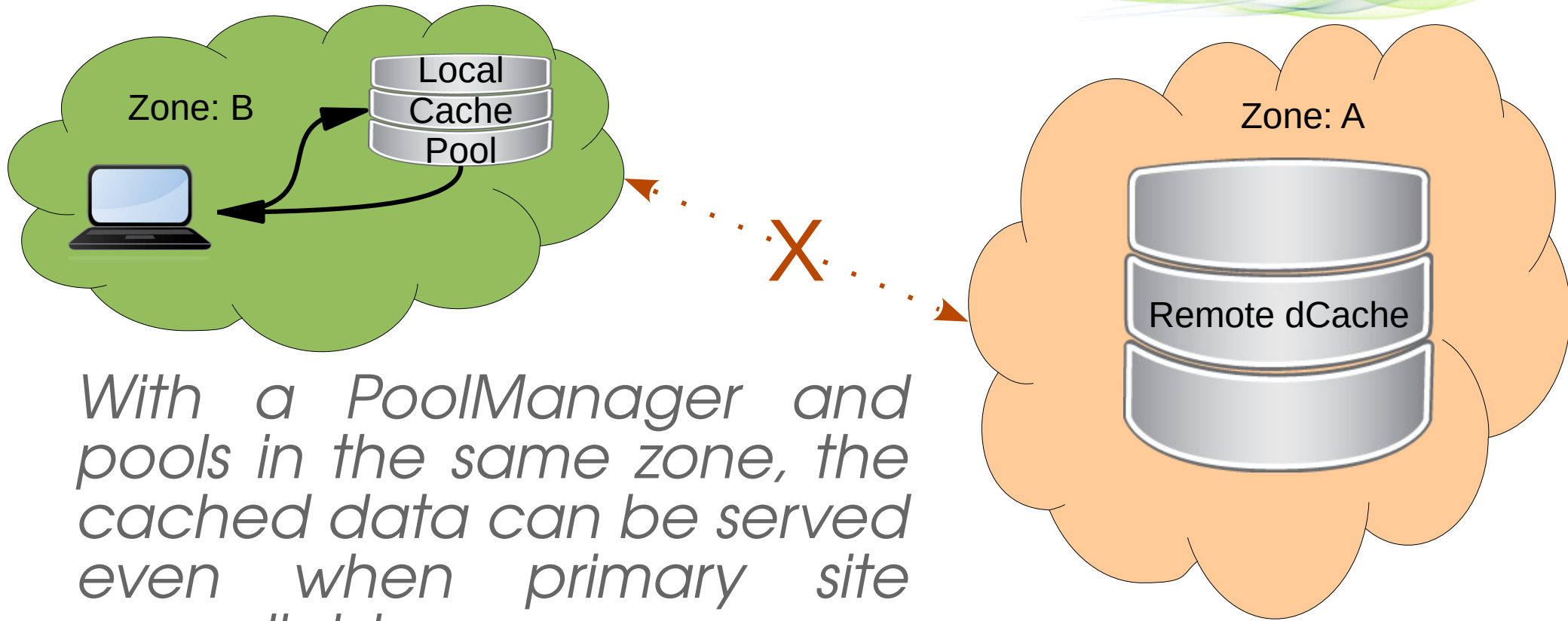
```
[ro-door- $\{host.name\}$ ]
```

```
dcache.queue.pnfsmanager=PnfsManager-RO
```

```
dcache.service.pnfsmanager=PnfsManager-RO
```

```
[ro-door- $\{host.name\}$ /nfs]
```

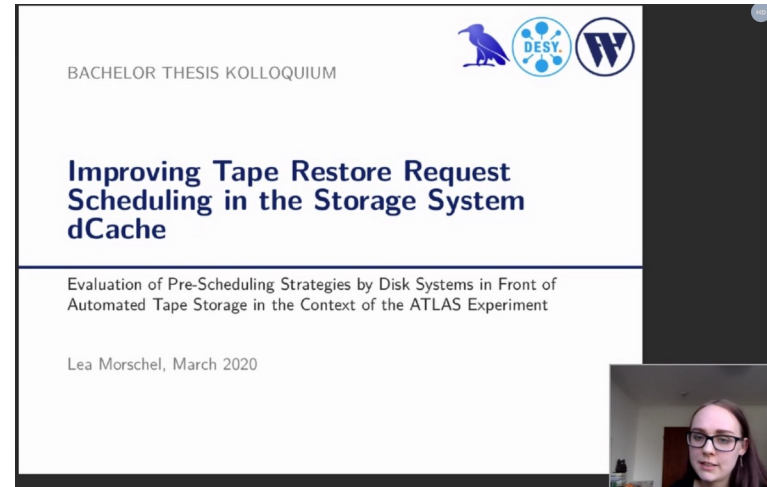
```
[ro-door- $\{host.name\}$ /xrootd]
```






With a PoolManager and pools in the same zone, the cached data can be served even when primary site unavailable.

HSM & QoS

- Atlas tape carousel => WLCG Tape carousel
 - Bad habits die hard
- Tape driver interface evolution
 - Docu update
 - New drivers
 - Functionality enhancements

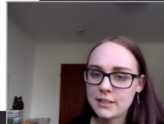


BACHELOR THESIS KOLLOQUIUM   

Improving Tape Restore Request Scheduling in the Storage System dCache

Evaluation of Pre-Scheduling Strategies by Disk Systems in Front of Automated Tape Storage in the Context of the ATLAS Experiment

Lea Morschel, March 2020



- Group requests by tapes (on pools)
 - Requests to single tape submitted together
- Group tapes by pools (on PinManager)
 - Access to single tape sent to a single pool
 - Requests from multiple pools to single tape coordinated
 - Some sites `fixed` by using single stage pool

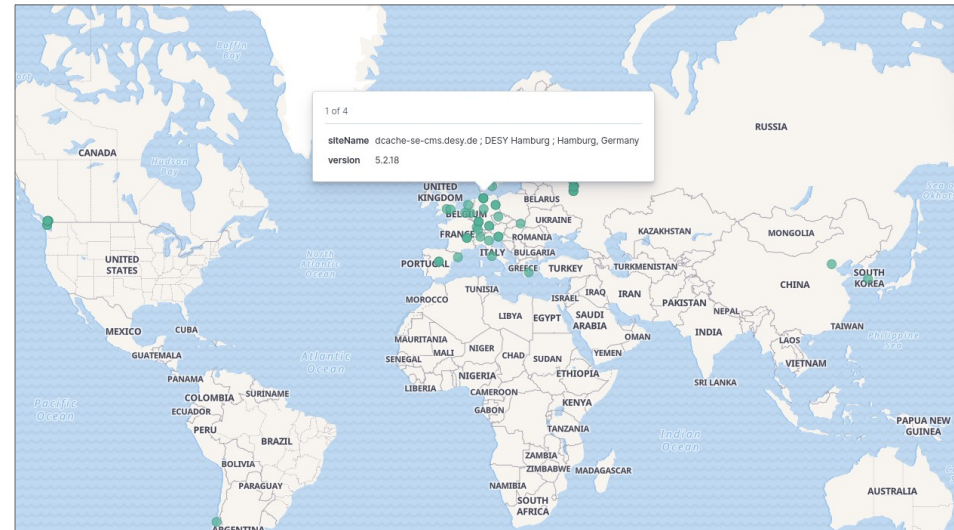
HSM, Tape, QoS

- New role for resilience
 - QoS capabilities
- Bulk operations
 - Like SRM, but different
 - QoS transitions directory/storage group based

dCache.org

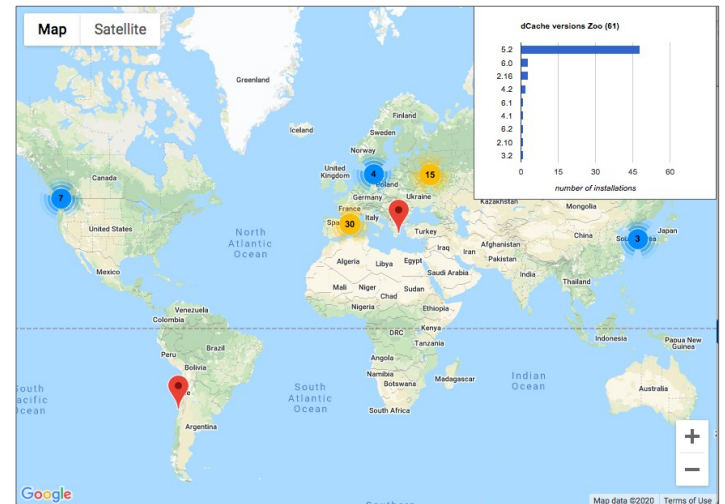
Watching you!

- See “the community”
 - Show where dCache is used
 - Good PR for the project
- See versions around
 - Version popularity
 - Planing back-ports



Why Not dCache-map?

- Information comes from BDII
 - Only grid sites
 - Information limited by GLUE schema
 - In phaseout state
 - No US sites
- Uses google maps
 - GDPR concerns



What We Collect, Where

- Site Name
 - telemetry.instance.site-name=Friends Of dCache
- Geo-location
 - telemetry.instance.location.latitude=
 - telemetry.instance.location.longitude=
- dCache version
 - auto detected
- Online capacity
 - calculated
- Stored in DB at DESY
 - telemetry.destination.url=https://stats.dcache.org/collector

What We Collect

```
{  
  "siteName": "dcache-se-desy.desy.de ; DESY Hamburg",  
  "version" : "5.2.20",  
  "storage" : 3551478000,  
  "location": {  
    "lon" : 9.8772,  
    "lat" : 53.5772  
  }  
}
```

How to Enable

[telemetryDomain]

[telemetryDomain/telemetry]

```
telemetry.cell.enable=true
```


- Sites have to enable service explicitly
- No Personal data collected
- Data used by dCache developers only
 - we might use it in our presentations!
- Site information can be removed on demand
- **Consult your SITE SECURITY OFFICER before enabling!**

45 m \Rightarrow 7 m

Pool Start-up

- Pool sizes increase
- File sizes decrease
- Typical pool has >1M files
- dcache-lab – 4M files, (cold) startup time 45min.

Pool Start-up

- File System Information
 - list of entries
 - file size
 - modification time (atime)
- Metadata
 - State (precious, from client, ...)
 - Sticky flags
 - Storage group information



Pool Start-up

- File System Information
 - list of entries
- Metadata (BerkeleyDB)
 - file size
 - access time
 - State (precious, from client, ...)
 - Sticky flags
 - Storage group information



Pool Start-up

- File System Information
 - list of entries
- Metadata (BerkeleyDB)
 - file size
 - access time
 - State (precious, from client, ...)
 - Sticky flags
 - Storage group information



Pool Start-up Time Optimization

- Less operations on FS
- Parallel scan

```
pool.limits.scan-threads=8
```

- BerkeleyDB optimization options

```
pool.plugins.meta.db!je.checkpointer.wakeupInterval =  
60 s
```

```
pool.plugins.meta.db!je.checkpointer.bytesInterval = 0
```

```
pool.plugins.meta.db!je.cleaner.wakeupInterval= 0 s
```

```
pool.plugins.meta.db!je.log.fileCacheSize=1024
```

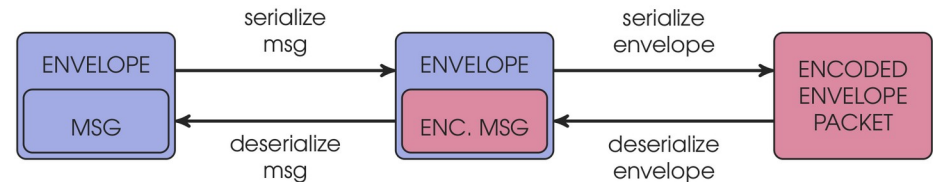
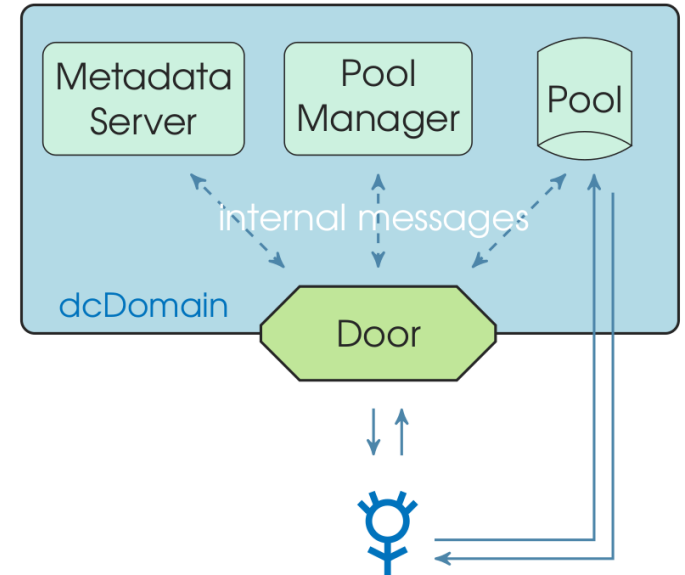
Last, but
not least!

- Some deployments use pools as buffers
- Sequential removal of cached files not effective
- Bulk cleanup for efficient data ingest
 - `pool.limits.sweeper-margin=0.0..1.0`
 - the value shows how much to clean related to pool size

New Message Serialization

- Interactive usage / HPC via NFS/WebDAV
 - Latency for the individual user becomes noticeable!
- Inter-cell communication comparable to IO times
- New experimental message serializer:
 - (De)serializing ~ 10% faster!
 - All components run the same minor version

dcache.broker.channel.msg-payload-serializer=**experimental**



systemd or sysvinit?

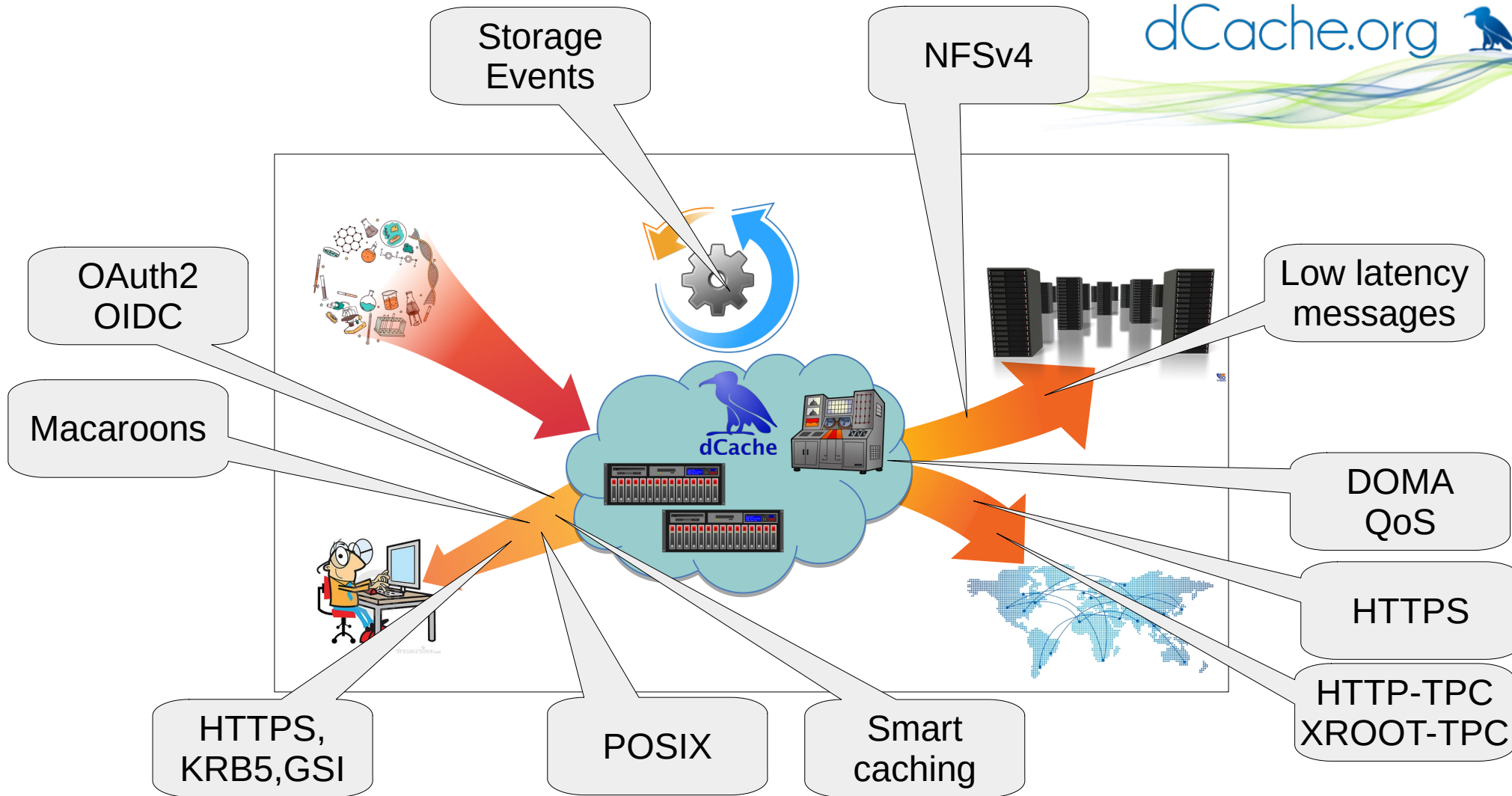
- Our Zoo
 - deb – systemd (only two sites?)
 - rpm - sysvinit
- Lot of custom functionality covered by systemd
 - permission drop
 - demonizing
 - auto-restart on crash
- We need your help to make it happen

Demand on Community effort

- You are the experts!
- How we can collect you knowledge?
 - HW selection
 - Best practice
 - Configuration management

Will be Removed Soon

- gPlazma
 - nis
 - nsswitch
- pool
 - file metadata
 - mongo metadata
- rrd plots



Support/Contact Channels

- **support [✉ dcache.org](mailto:support@dcache.org)**
 - User request tracking system
 - Place to ask for a help from developers
 - Accessible by all team member
- **security [✉ dcache.org](mailto:security@dcache.org)**
 - Request tracking system
 - To report security issues or incidents
 - Accessible by selected people
- **user-forum [✉ dcache.org](mailto:user-forum@dcache.org)**
 - Mailing list for sysadmins self-help group
 - To ask for an advice or share experience
 - Used by (almost) all sysadmins and developers
- **dev [✉ dcache.org](mailto:dev@dcache.org)**
 - Shared mailbox
 - An email to contact developers. Not for support
 - Developers can send e-mail from this address
- **srm-deployment [✉ dcache.org](mailto:srm-deployment@dcache.org)**
 - Tier – 1 coordination mailing list
- **meet.desy.de/xxx**
 - Weekly Tier 1 support video call
- **workshop [✉ dcache.org](mailto:workshop@dcache.org)**
 - Shared mailbox
 - An e-mail used to organize workshops
- **Github issues**
 - Request tracking system
 - To report software defects and feature requests
 - Public
- **Github pull-requests**
 - Request tracking system
 - To provide code changes
 - Public

Thank You!

Survey report

- Main usage – WLCG
 - but most of the sites have local communities
- The main value
 - flexibility, stability, documentation (Wow!)
- Disappointment
 - migration guide quality
 - Complexity
 - Closed development process
- All sites have other systems in parallel
 - Object Store
 - R/W POSIX-IO
 - simple usecase
- Happiness
 - 60% happy with the quality, 40% ok
 - 70% with stability, 30% ok
 - 50% ok with administration, 30% see need to improve, 10% think it's crap
 - 50% happy with performance, 30% ok
 - 70% see potential for resource utilization
- Support
 - In general Ok, but can be improved, especially response times.
- 90% seeking information in the book.
- 10% happy with the content of the book. 70% urge better documentation.
- FTP is going to die.
- Multi-protocol support is important.
 - NFS for local
 - WebDAV remote
 - xroot forced
- 70% are ok with installation procedure
 - easy start, hard to evolve
- 60% run (some) dcache components in VMs
- 50/50 release frequency
- 90% run golden/LTS release
- Most of the sites have more than one instance.