18th International dCache Workshop Summary

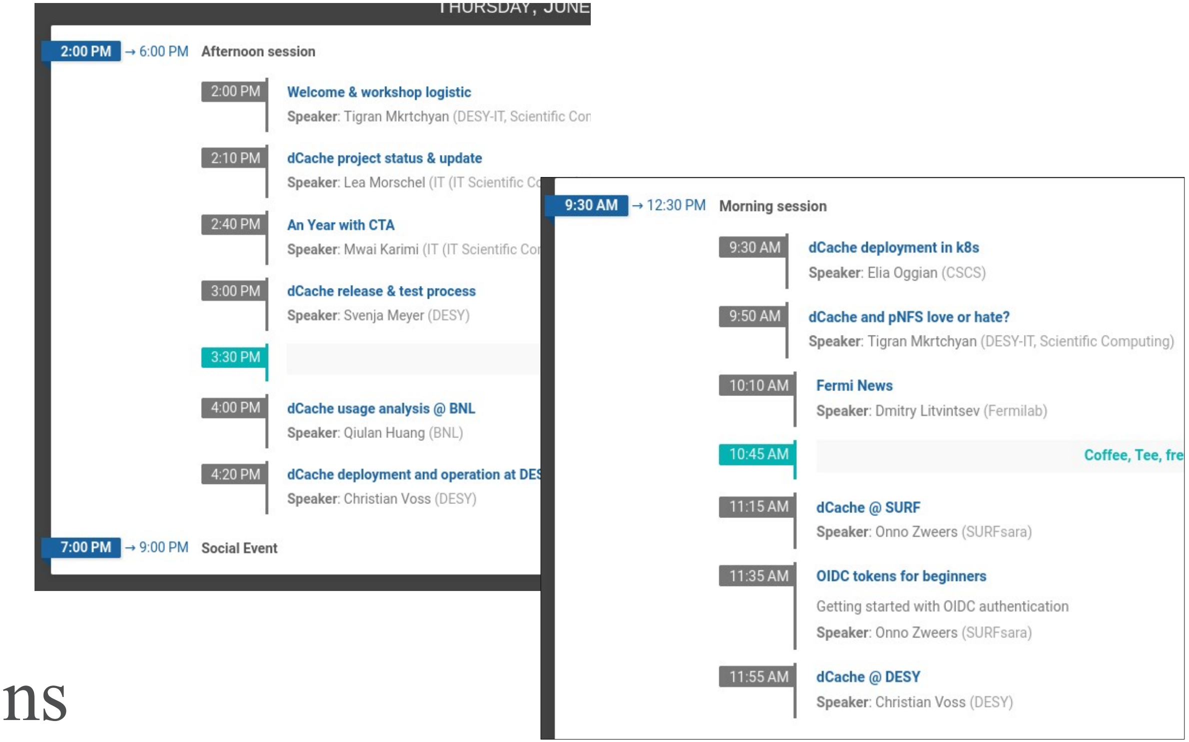
Tigran Mkrtchyan for the dCache collaboration

DESY. Fermilab neic NORDIC E-INFRASTRUCTURE COLLABORATION

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Sessions

- Participants
  - 23 in person
  - 23 online
- Agenda
  - two half-days
  - lot of open discussions
    - both sessions took 2 hours longer than planned



THURSDAY, JUNE

2:00 PM → 6:00 PM Afternoon session

2:00 PM Welcome & workshop logistic
Speaker: Tigran Mkrtchyan (DESY-IT, Scientific Cor

2:10 PM dCache project status & update
Speaker: Lea Morschel (IT (IT Scientific Co

2:40 PM An Year with CTA
Speaker: Mwai Karimi (IT (IT Scientific Cor

3:00 PM dCache release & test process
Speaker: Svenja Meyer (DESY)

3:30 PM

4:00 PM dCache usage analysis @ BNL
Speaker: Qiulan Huang (BNL)

4:20 PM dCache deployment and operation at DES
Speaker: Christian Voss (DESY)

7:00 PM → 9:00 PM Social Event

9:30 AM → 12:30 PM Morning session

9:30 AM dCache deployment in k8s
Speaker: Elia Oggian (CSCS)

9:50 AM dCache and pNFS love or hate?
Speaker: Tigran Mkrtchyan (DESY-IT, Scientific Computing)

10:10 AM Fermi News
Speaker: Dmitry Litvintsev (Fermilab)

10:45 AM Coffee, Tee, fre

11:15 AM dCache @ SURF
Speaker: Onno Zweers (SURFsara)

11:35 AM OIDC tokens for beginners
Getting started with OIDC authentication
Speaker: Onno Zweers (SURFsara)

11:55 AM dCache @ DESY
Speaker: Christian Voss (DESY)

Main Topics

- Project status
- dCache-CTA integration
- Large deployment trouble shooting
- Monitoring
- Tokens

Project Status

People

DESY.

- Karen Hoyos
- Svenja Meyer
- Tigran Mkrtchyan
- Lea Morschel
- Marina Sahakyan

Fermilab

- Chris Green
- Dmitry Litvintsev

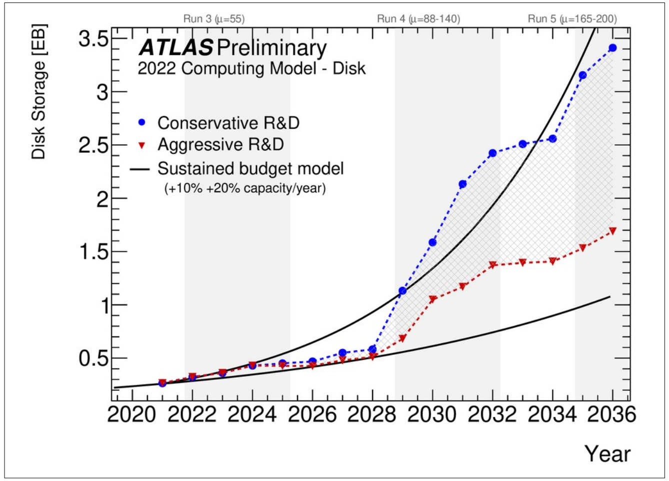
neic NORDIC E-INFRASTRUCTURE COLLABORATION

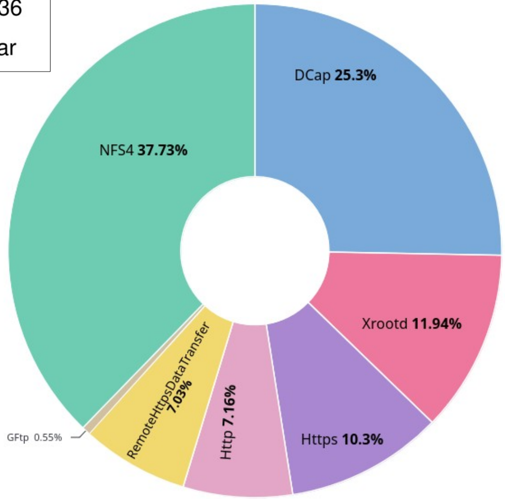
- Krishnaveni Chitrapu
- Darren Starr

Σ people != Σ FTE

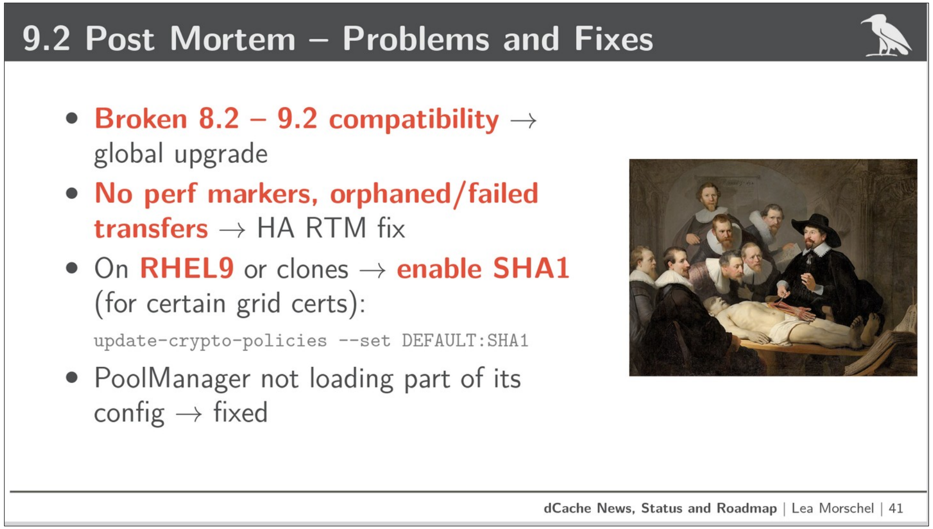
The Challenges

- 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





9.2 Post Mortem



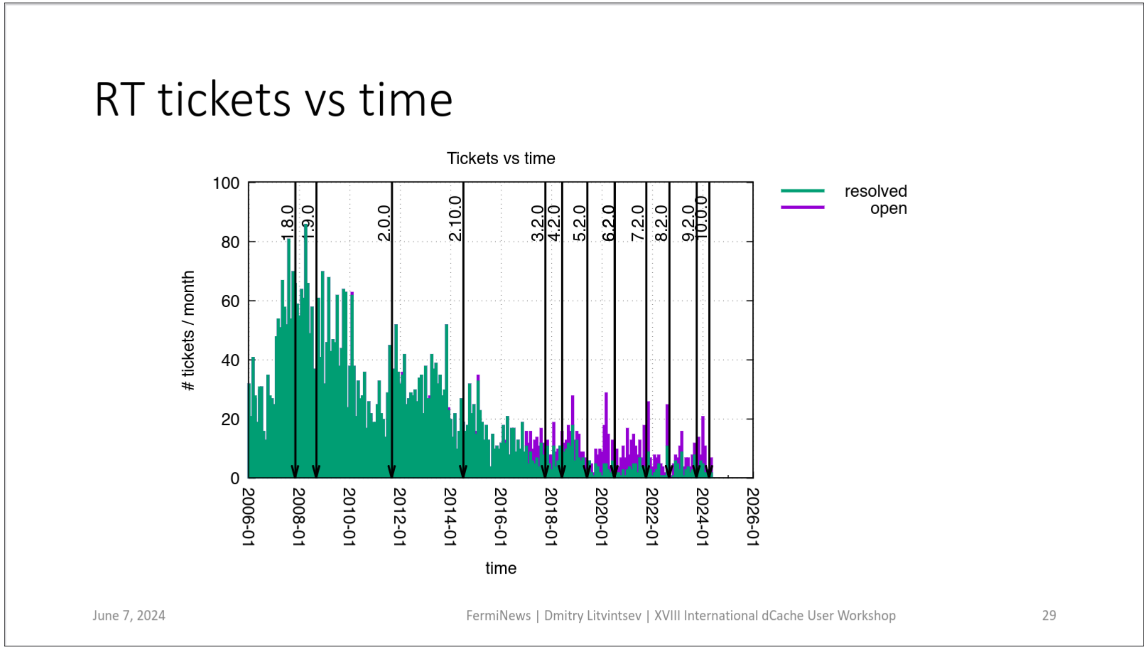
9.2 Post Mortem – Problems and Fixes

- **Broken 8.2 – 9.2 compatibility** → global upgrade
- **No perf markers, orphaned/failed transfers** → HA RTM fix
- On **RHEL9** or clones → **enable SHA1** (for certain grid certs):

```
update-crypto-policies --set DEFAULT:SHA1
```

- PoolManager not loading part of its config → fixed

dCache News, Status and Roadmap | Lea Morschel | 41

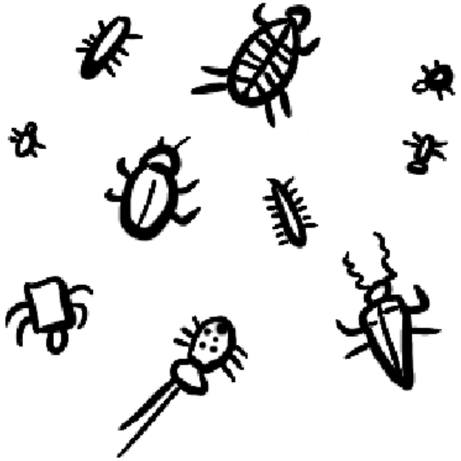


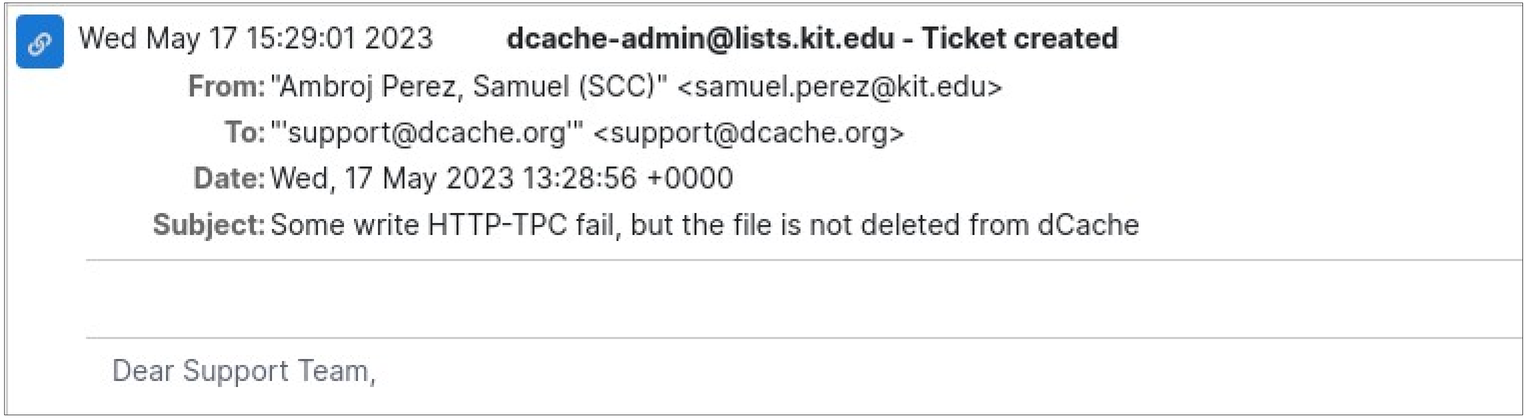
RT tickets vs time

June 7, 2024 FermiNews | Dmitry Litvintsev | XVIII International dCache User Workshop 29

Types of Bugs

- Low hanging fruits
  - introduced by new developments
  - often under time pressure by experiments (mostly related to tokens)
- Zero-day issues
  - scaling problems, race conditions
  - hard to reproduce

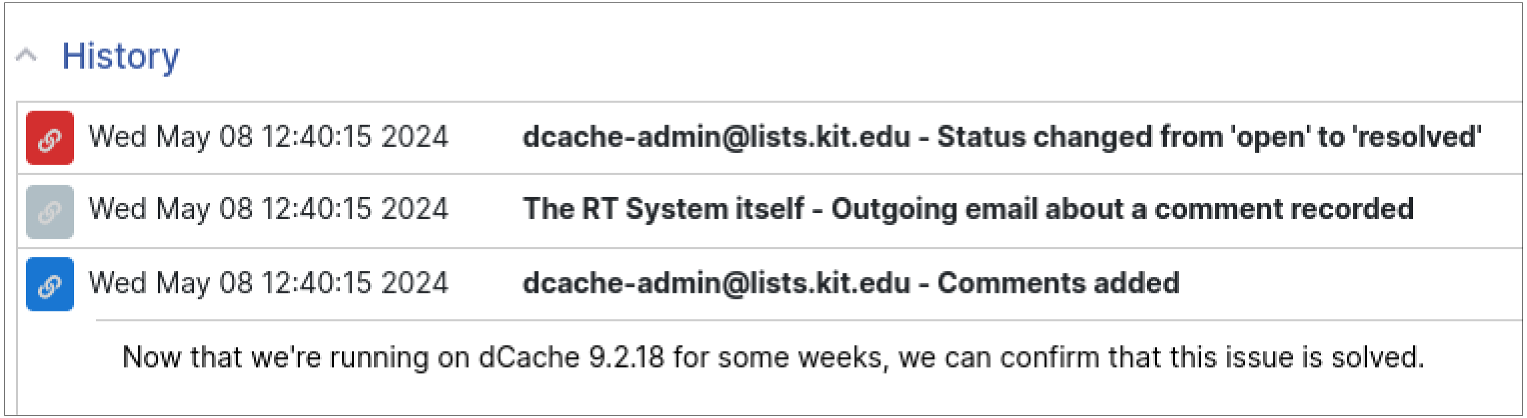




Wed May 17 15:29:01 2023 **dcache-admin@lists.kit.edu - Ticket created**

From: "Ambroj Perez, Samuel (SCC)" <samuel.perez@kit.edu>
To: "'support@dcache.org'" <support@dcache.org>
Date: Wed, 17 May 2023 13:28:56 +0000
Subject: Some write HTTP-TPC fail, but the file is not deleted from dCache

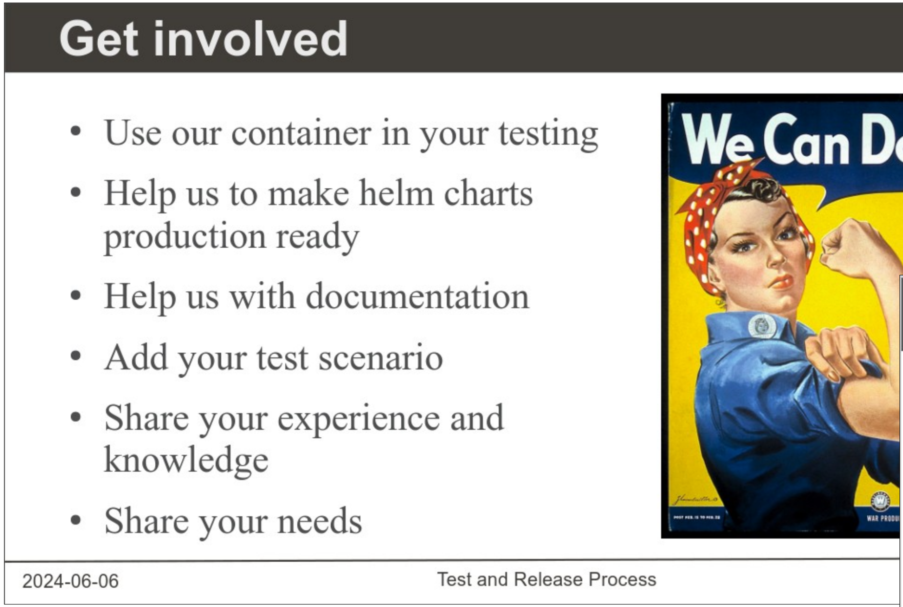
Dear Support Team,



History

Wed May 08 12:40:15 2024	**dcache-admin@lists.kit.edu - Status changed from 'open' to 'resolved'**
Wed May 08 12:40:15 2024	**The RT System itself - Outgoing email about a comment recorded**
Wed May 08 12:40:15 2024	**dcache-admin@lists.kit.edu - Comments added**

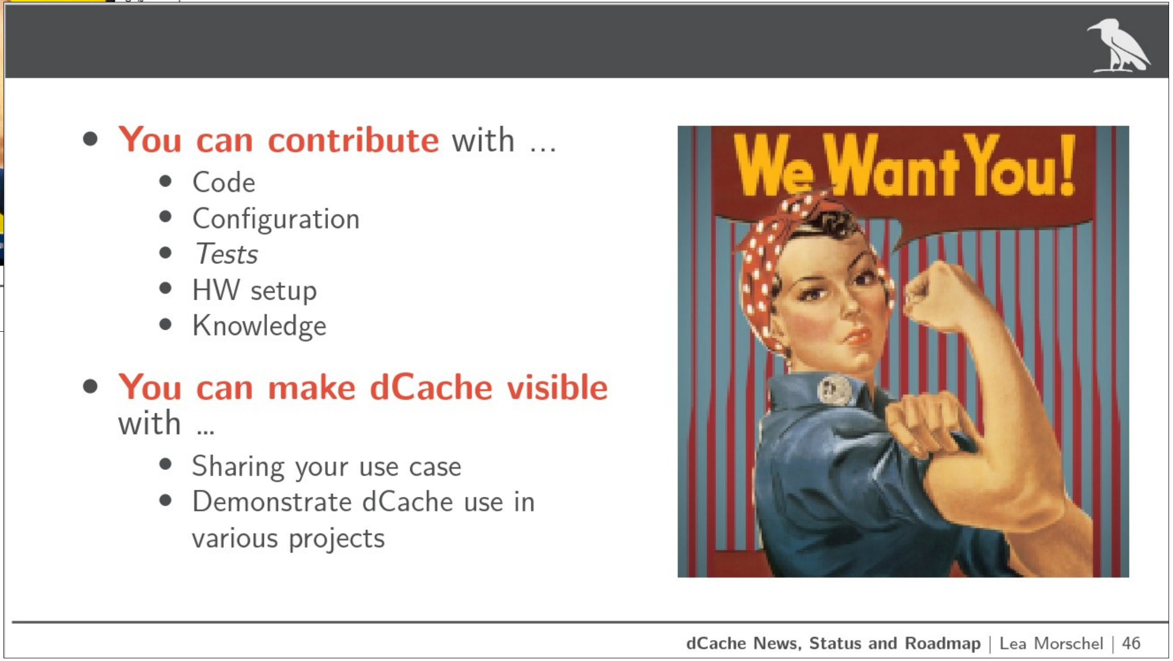
Now that we're running on dCache 9.2.18 for some weeks, we can confirm that this issue is solved.



Get involved

- Use our container in your testing
- Help us to make helm charts production ready
- Help us with documentation
- Add your test scenario
- Share your experience and knowledge
- Share your needs

2024-06-06 Test and Release Process



- **You can contribute** with ...
  - Code
  - Configuration
  - *Tests*
  - HW setup
  - Knowledge
- **You can make dCache visible** with ...
  - Sharing your use case
  - Demonstrate dCache use in various projects

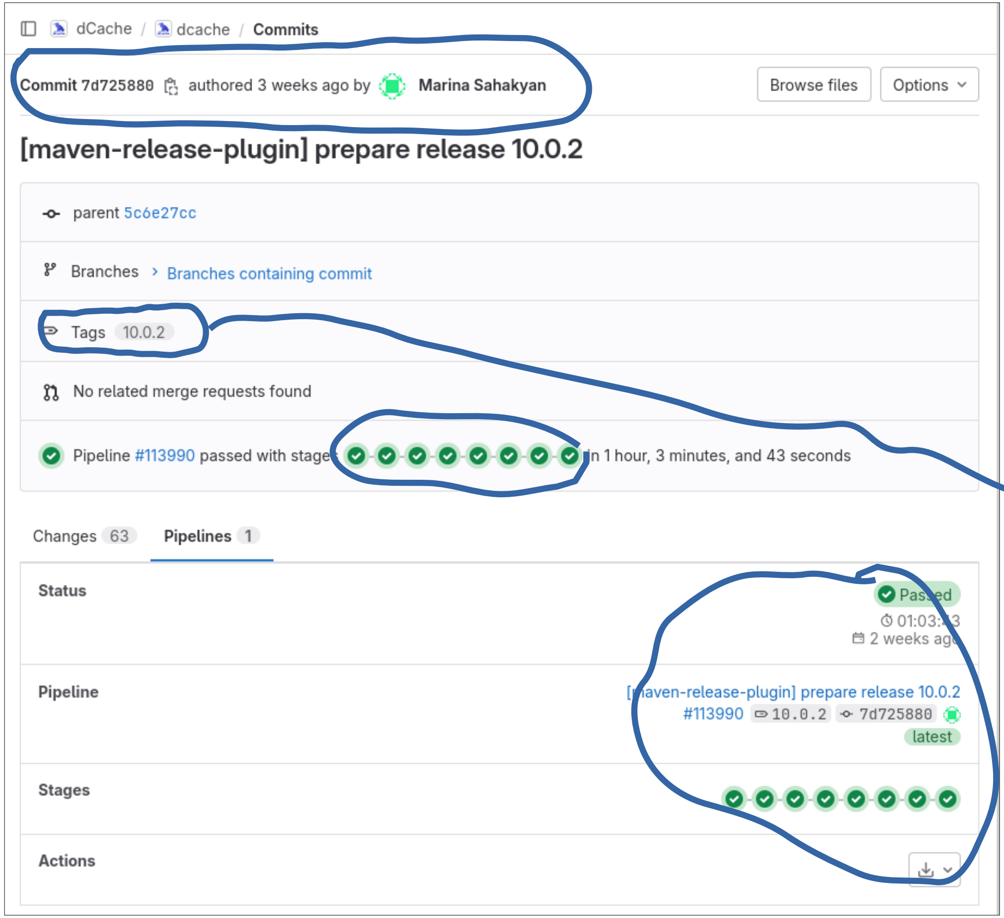
dCache News, Status and Roadmap | Lea Morschel | 46

Release & Test

Why to Know

- Get to know what we test and what we don’t
- Re-use our setup on your testbed
    - Get to know new functionality
- Re-run our test for your custom builds
- Extend our tests with your test case
    - Add your site setup

Outline



dCache / dcache / **Commits**

Commit 7d725880 authored 3 weeks ago by **Marina Sahakyan**

[maven-release-plugin] prepare release 10.0.2

parent 5c6e27cc

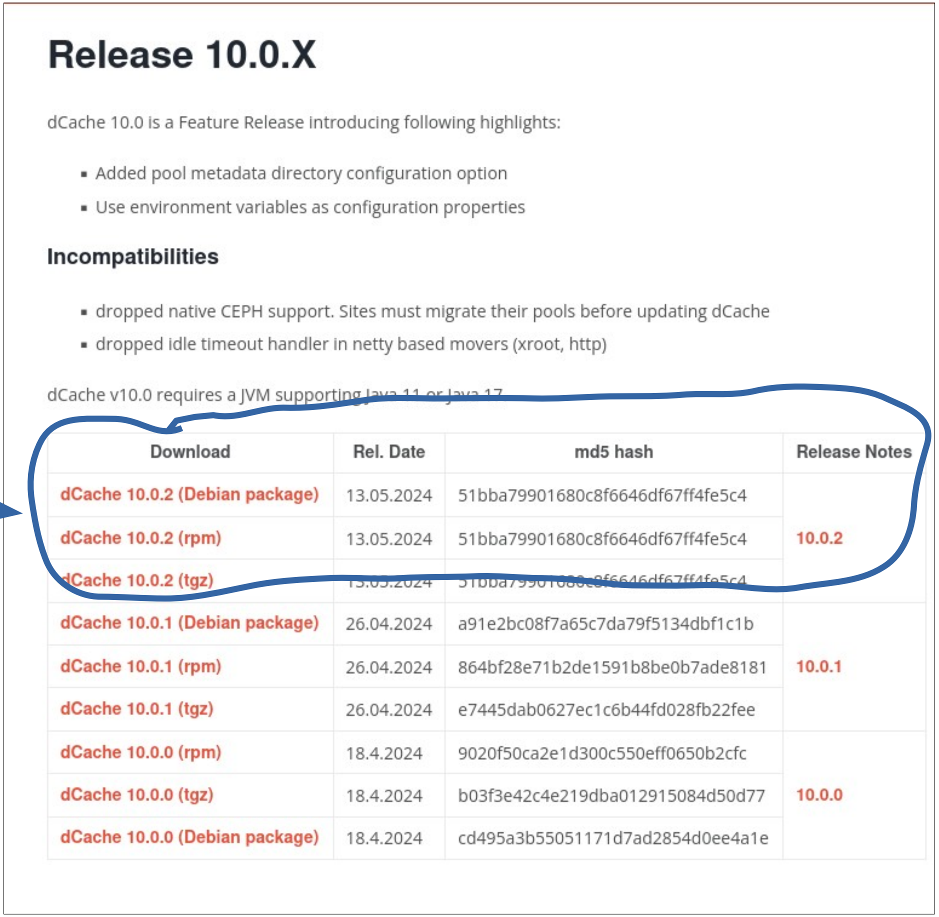
Branches › Branches containing commit

Tags 10.0.2

No related merge requests found

Pipeline #113990 passed with stages ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ in 1 hour, 3 minutes, and 43 seconds

Status	✓ Passed 01:03:43 2 weeks ago
Pipeline	[maven-release-plugin] prepare release 10.0.2 #113990 10.0.2 7d725880 latest
Stages	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓
Actions	



Release 10.0.X

dCache 10.0 is a Feature Release introducing following highlights:

- Added pool metadata directory configuration option
- Use environment variables as configuration properties

Incompatibilities

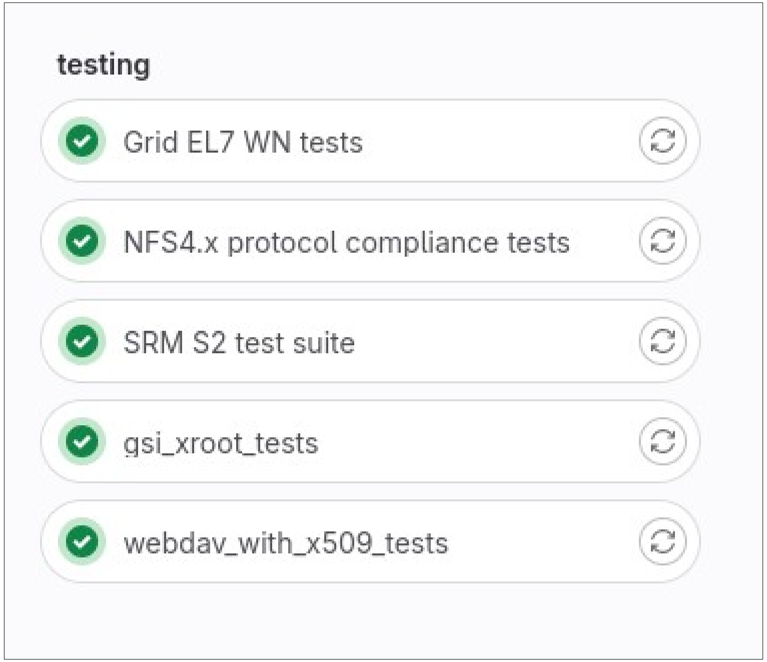
- dropped native CEPH support. Sites must migrate their pools before updating dCache
- dropped idle timeout handler in netty based movers (xroot, http)

dCache v10.0 requires a JVM supporting Java 11 or Java 17

Download	Rel. Date	md5 hash	Release Notes
dCache 10.0.2 (Debian package)	13.05.2024	51bba79901680c8f6646df67ff4fe5c4	10.0.2
dCache 10.0.2 (rpm)	13.05.2024	51bba79901680c8f6646df67ff4fe5c4	
dCache 10.0.2 (tgz)	13.05.2024	51bba79901680c8f6646df67ff4fe5c4	
dCache 10.0.1 (Debian package)	26.04.2024	a91e2bc08f7a65c7da79f5134dbf1c1b	10.0.1
dCache 10.0.1 (rpm)	26.04.2024	864bf28e71b2de1591b8be0b7ade8181	
dCache 10.0.1 (tgz)	26.04.2024	e7445dab0627ec1c6b44fd028fb22fee	
dCache 10.0.0 (rpm)	18.4.2024	9020f50ca2e1d300c550eff0650b2cfc	10.0.0
dCache 10.0.0 (tgz)	18.4.2024	b03f3e42c4e219dba012915084d50d77	
dCache 10.0.0 (Debian package)	18.4.2024	cd495a3b55051171d7ad2854d0ee4a1e	

Tests

- Grid toolkit with EL7
  - dccp, gridftp, srm, gfal-xxx, 3rd-party copy
- SRM spec compatibility tests
  - test suite since srm-2.0 deplyment
- xroot-gsi test
- Simple WebDAV with x509
- NFS protocol compatibility
  - No kernel client tests!



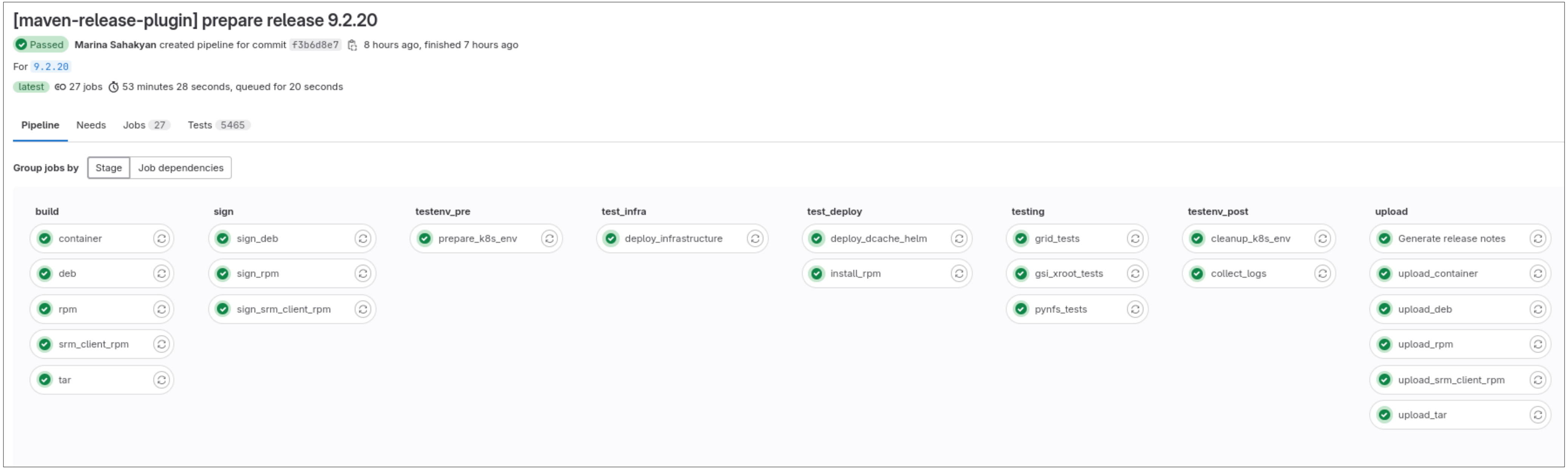
testing

- ✓ Grid EL7 WN tests
- ✓ NFS4.x protocol compliance tests
- ✓ SRM S2 test suite
- ✓ gsi_xroot_tests
- ✓ webdav_with_x509_tests

Tested Manually

- Kernel NFS I/O
  - fio, mdtest, xfs-tests
- HSM interface
  - script, CTA
- DB schema migration
- REST API & frontend
- Migration module
- HA, Fail-over
- Backward compatibility
- …

The Full ‘Thing’



[maven-release-plugin] prepare release 9.2.20

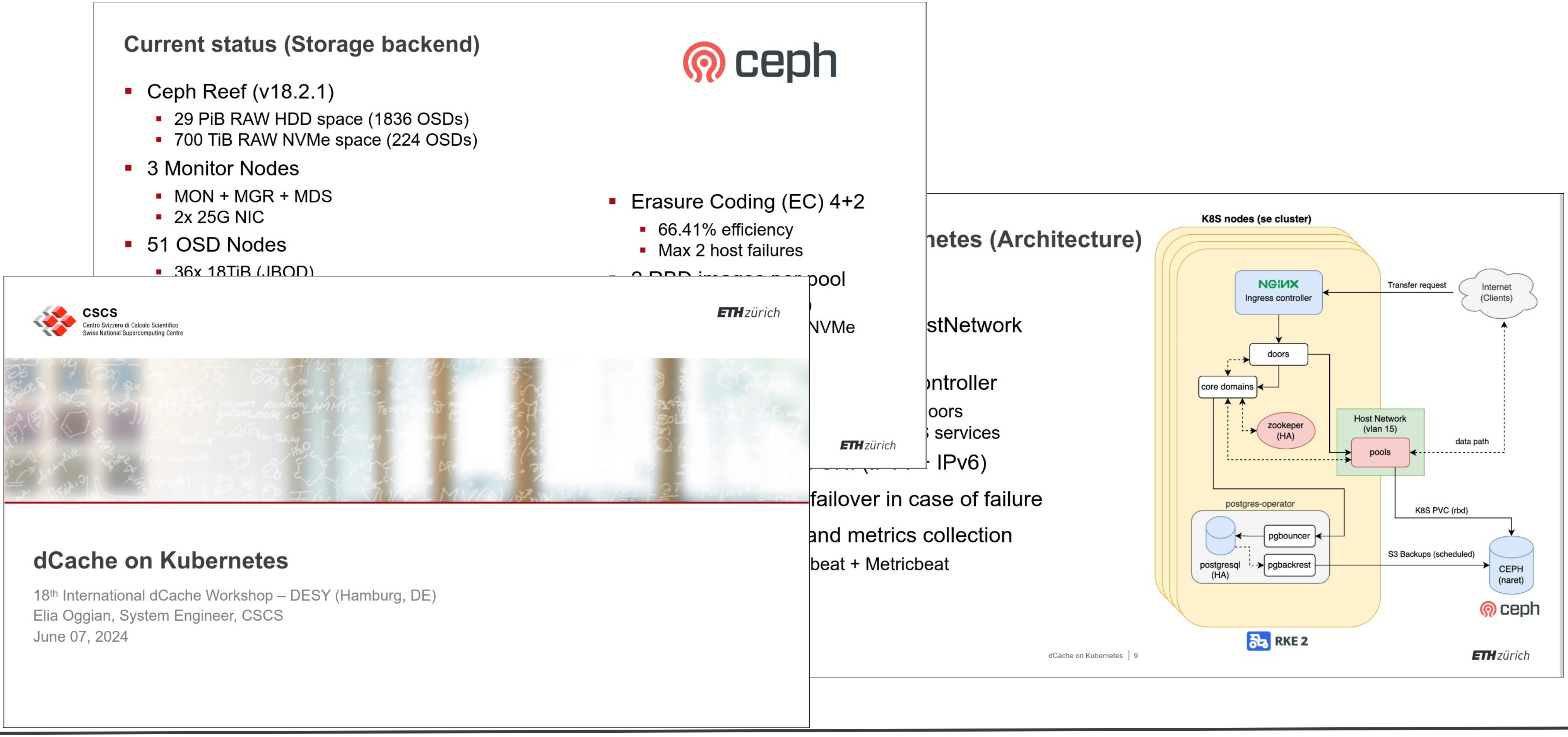
Passed **Marina Sahakyan** created pipeline for commit `f3b6d8e7` 8 hours ago, finished 7 hours ago

For `9.2.20`

latest 27 jobs 53 minutes 28 seconds, queued for 20 seconds

build	sign	testenv_pre	test_infra	test_deploy	testing	testenv_post	upload
container	sign_deb	prepare_k8s_env	deploy_infrastructure	deploy_dcache_helm	grid_tests	cleanup_k8s_env	Generate release notes
deb	sign_rpm			install_rpm	gsi_xroot_tests	collect_logs	upload_container
rpm	sign_srm_client_rpm				pynfs_tests		upload_deb
srm_client_rpm							upload_rpm
tar							upload_srm_client_rpm
							upload_tar

dCache on Kubernetes



Current status (Storage backend)

- Ceph Reef (v18.2.1)
  - 29 PiB RAW HDD space (1836 OSDs)
  - 700 TiB RAW NVMe space (224 OSDs)
- 3 Monitor Nodes
  - MON + MGR + MDS
  - 2x 25G NIC
- 51 OSD Nodes
  - 36x 18TiB (JBOD)

- Erasure Coding (EC) 4+2
  - 66.41% efficiency
  - Max 2 host failures
- 2 RBD images per pool

NVMe

ETH zürich

netes (Architecture)

stNetwork

ontroller

oors

services

IPv6)

failover in case of failure

and metrics collection

beat + Metricbeat

dCache on Kubernetes | 9

ETH zürich

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

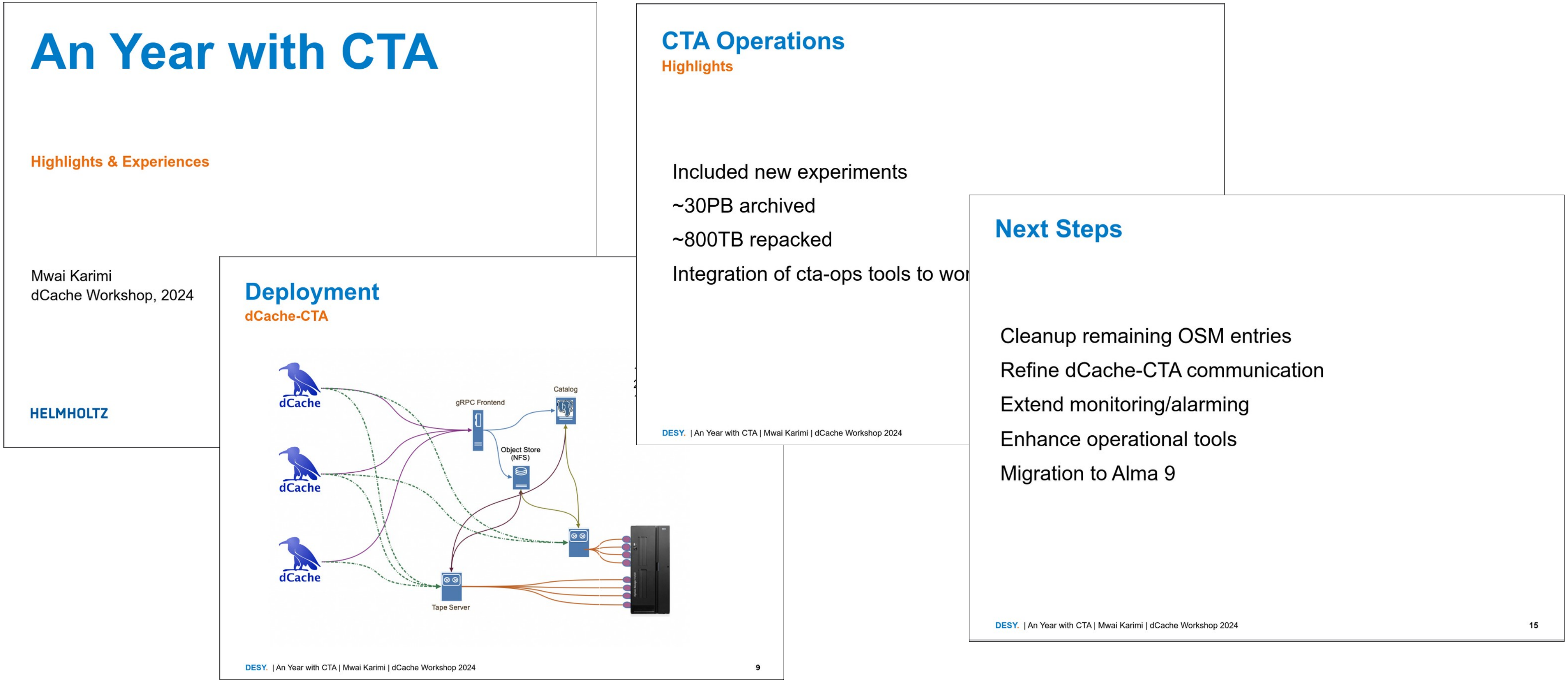
ETH zürich

dCache on Kubernetes

18th International dCache Workshop – DESY (Hamburg, DE)
Elia Oggian, System Engineer, CSCS
June 07, 2024

CERN Tape Archive

CTA at DESY



An Year with CTA

Highlights & Experiences

Mwai Karimi
dCache Workshop, 2024

HELMHOLTZ

Deployment

dCache-CTA

DESY. | An Year with CTA | Mwai Karimi | dCache Workshop 2024 9

CTA Operations

Highlights

Included new experiments

~30PB archived

~800TB repacked

Integration of cta-ops tools to wor

DESY. | An Year with CTA | Mwai Karimi | dCache Workshop 2024

Next Steps

Cleanup remaining OSM entries

Refine dCache-CTA communication

Extend monitoring/alarming

Enhance operational tools

Migration to Alma 9

DESY. | An Year with CTA | Mwai Karimi | dCache Workshop 2024 15

CTA at Fermilab

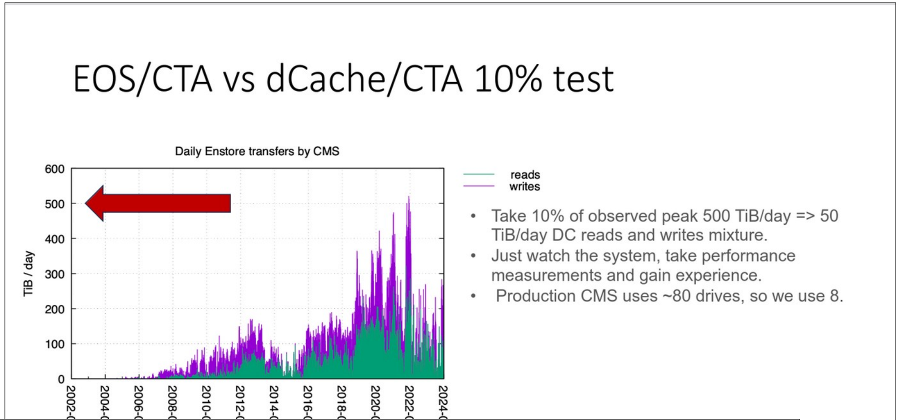
FERMI NEWS

FERMILAB A U.S. DEPARTMENT OF ENERGY LABORATORY

Dmitry Litvintsev

18th International dCache user workshop

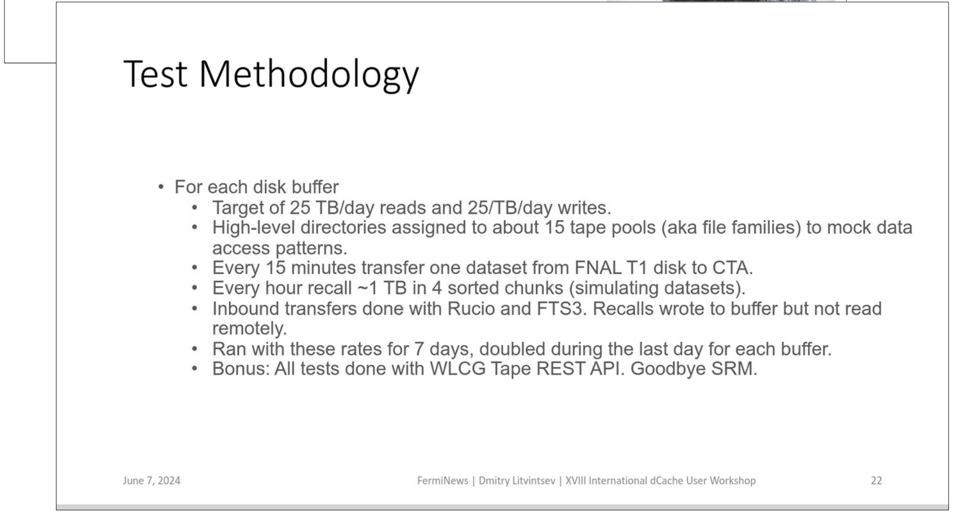
DESY, Hamburg, June 7, 2024



EOS/CTA vs dCache/CTA 10% test

- Take 10% of observed peak 500 TiB/day => 50 TiB/day DC reads and writes mixture.
- Just watch the system, take performance measurements and gain experience.
- Production CMS uses ~80 drives, so we use 8.

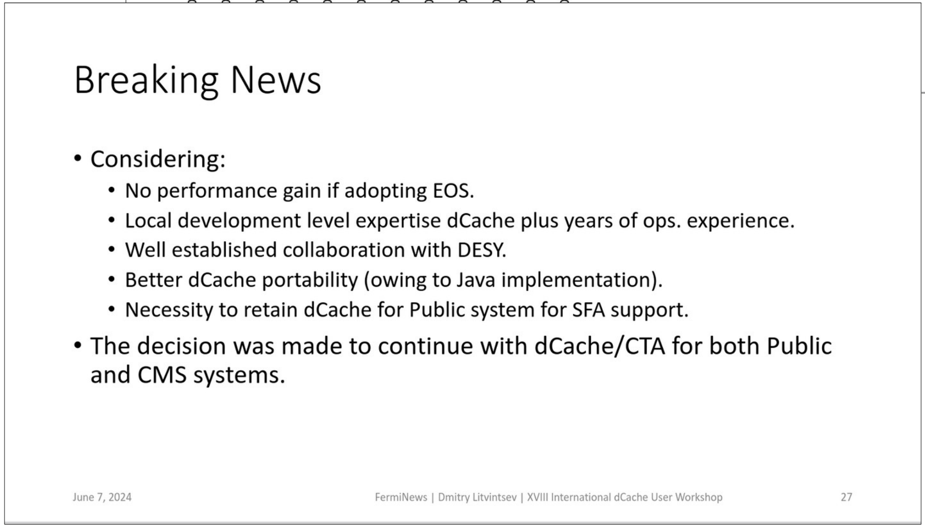
20



Test Methodology

- For each disk buffer
  - Target of 25 TB/day reads and 25/TB/day writes.
  - High-level directories assigned to about 15 tape pools (aka file families) to mock data access patterns.
  - Every 15 minutes transfer one dataset from FNAL T1 disk to CTA.
  - Every hour recall ~1 TB in 4 sorted chunks (simulating datasets).
  - Inbound transfers done with Rucio and FTS3. Recalls wrote to buffer but not read remotely.
  - Ran with these rates for 7 days, doubled during the last day for each buffer.
  - Bonus: All tests done with WLCG Tape REST API. Goodbye SRM.

June 7, 2024 FermiNews | Dmitry Litvintsev | XVIII International dCache User Workshop 22



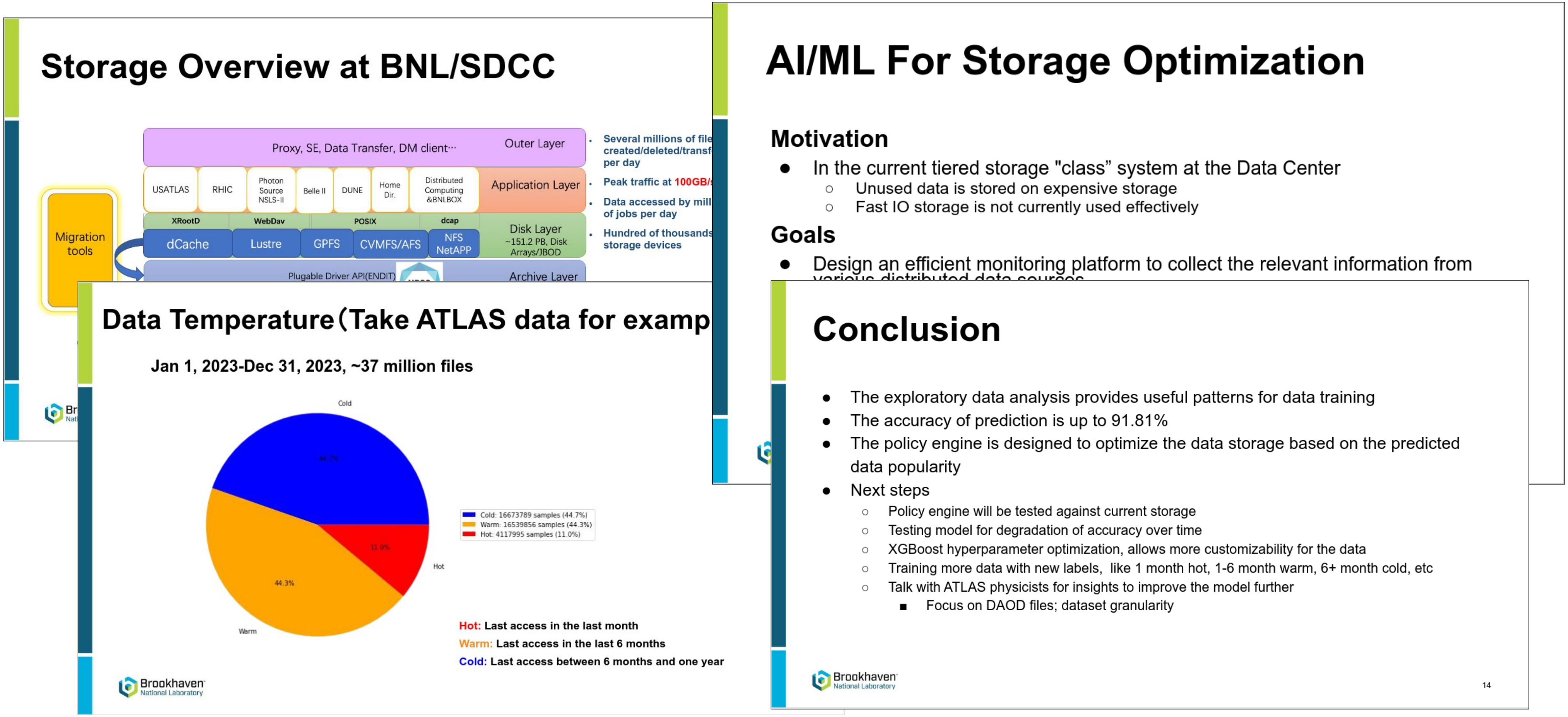
Breaking News

- Considering:
  - No performance gain if adopting EOS.
  - Local development level expertise dCache plus years of ops. experience.
  - Well established collaboration with DESY.
  - Better dCache portability (owing to Java implementation).
  - Necessity to retain dCache for Public system for SFA support.
- The decision was made to continue with dCache/CTA for both Public and CMS systems.

June 7, 2024 FermiNews | Dmitry Litvintsev | XVIII International dCache User Workshop 27

Monitoring & Deployment

Storage Access Optimization @ BNL



Storage Overview at BNL/SDCC

Proxy, SE, Data Transfer, DM client··· | Outer Layer

USATLAS | RHIC | Photon Source NSLS-II | Belle II | DUNE | Home Dir. | Distributed Computing &BNLBOX | Application Layer

XRootD | WebDav | POSIX | dcap

dCache | Lustre | GPFS | CVMFS/AFS | NFS NetAPP | Disk Layer ~151.2 PB, Disk Arrays/JBOD

Plugable Driver API(ENDIT) | Archive Layer

Migration tools

- Several millions of file created/deleted/transf per day
- Peak traffic at 100GB/s
- Data accessed by mill of jobs per day
- Hundred of thousands storage devices

AI/ML For Storage Optimization

Motivation

- In the current tiered storage "class" system at the Data Center
  - Unused data is stored on expensive storage
  - Fast IO storage is not currently used effectively

Goals

- Design an efficient monitoring platform to collect the relevant information from various distributed data sources

Data Temperature（Take ATLAS data for examp

Jan 1, 2023-Dec 31, 2023, ~37 million files

Cold
44.7%
11.0%
Hot
44.3%
Warm
Cold: 16673789 samples (44.7%)
Warm: 16539856 samples (44.3%)
Hot: 4117995 samples (11.0%)

Hot: Last access in the last month

Warm: Last access in the last 6 months

Cold: Last access between 6 months and one year

Brookhaven National Laboratory

Conclusion

- The exploratory data analysis provides useful patterns for data training
- The accuracy of prediction is up to 91.81%
- The policy engine is designed to optimize the data storage based on the predicted data popularity
- Next steps
  - Policy engine will be tested against current storage
  - Testing model for degradation of accuracy over time
  - XGBoost hyperparameter optimization, allows more customizability for the data
  - Training more data with new labels, like 1 month hot, 1-6 month warm, 6+ month cold, etc
  - Talk with ATLAS physicists for insights to improve the model further
    - Focus on DAOD files; dataset granularity

Brookhaven National Laboratory 14

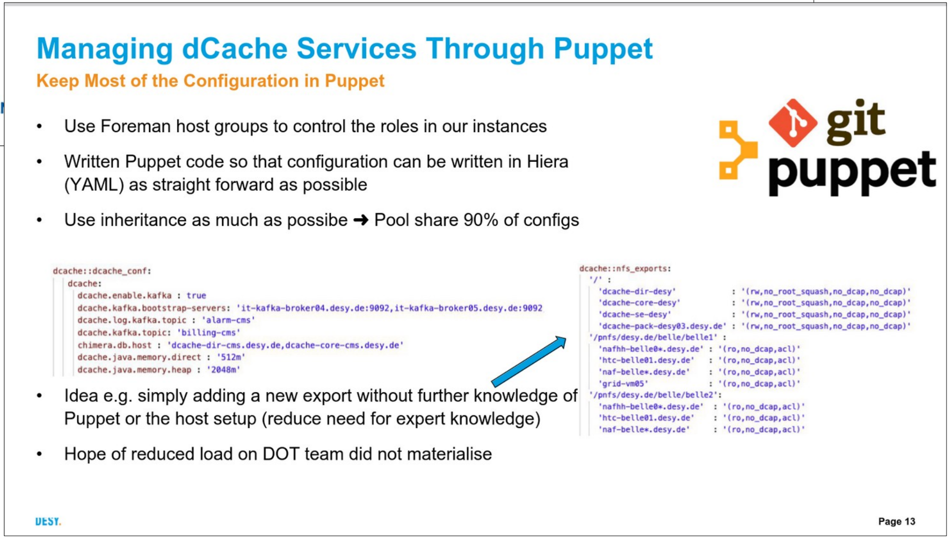
dCache at DESY

dCache-Operations at DESY

dCache Workshop Hamburg 06/07 June 2024

Christian Voß on behalf of the dCache Operations Team
DESY, 6th June 2024

HELI



Managing dCache Services Through Puppet

Keep Most of the Configuration in Puppet

- Use Foreman host groups to control the roles in our instances
- Written Puppet code so that configuration can be written in Hiera (YAML) as straight forward as possible
- Use inheritance as much as possibe ➜ Pool share 90% of configs

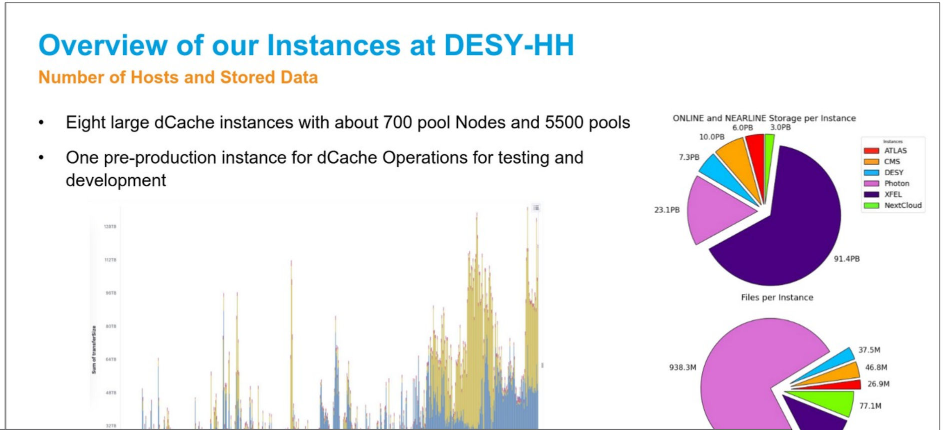
```
dcache::dcache_conf:
  dcache:
    dcache.enable.kafka : true
    dcache.kafka.bootstrap-servers: 'it-kafka-broker04.desy.de:9092,it-kafka-broker05.desy.de:9092
    dcache.log.kafka.topic : 'alarm-cms'
    dcache.kafka.topic: 'billing-cms'
    chimera.db.host : 'dcache-dir-cms.desy.de,dcache-core-cms.desy.de'
    dcache.java.memory.direct : '512m'
    dcache.java.memory.heap : '2048m'
```

```
dcache::nfs_exports:
  '/' :
    'dcache-dir-desy'                : '(rw,no_root_squash,no_dcap,no_dcap)'
    'dcache-core-desy'               : '(rw,no_root_squash,no_dcap,no_dcap)'
    'dcache-se-desy'                 : '(rw,no_root_squash,no_dcap,no_dcap)'
    'dcache-pack-desy03.desy.de'     : '(rw,no_root_squash,no_dcap,no_dcap)'
  '/pnfs/desy.de/belle/belle1' :
    'nafhh-belle0*.desy.de' : '(ro,no_dcap,acl)'
    'htc-belle01.desy.de'   : '(ro,no_dcap,acl)'
    'naf-belle*.desy.de'    : '(ro,no_dcap,acl)'
    'grid-vm05'             : '(ro,no_dcap,acl)'
  '/pnfs/desy.de/belle/belle2':
    'nafhh-belle0*.desy.de'  : '(ro,no_dcap,acl)'
    'htc-belle01.desy.de'    : '(ro,no_dcap,acl)'
    'naf-belle*.desy.de'     : '(ro,no_dcap,acl)'
```

- Idea e.g. simply adding a new export without further knowledge of Puppet or the host setup (reduce need for expert knowledge)
- Hope of reduced load on DOT team did not materialise

DESY.

Page 13

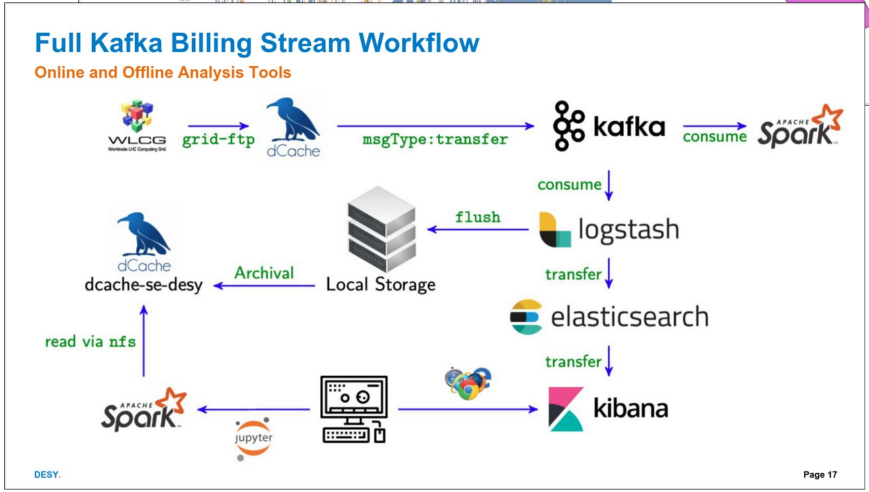


Overview of our Instances at DESY-HH

Number of Hosts and Stored Data

- Eight large dCache instances with about 700 pool Nodes and 5500 pools
- One pre-production instance for dCache Operations for testing and development

Page 8



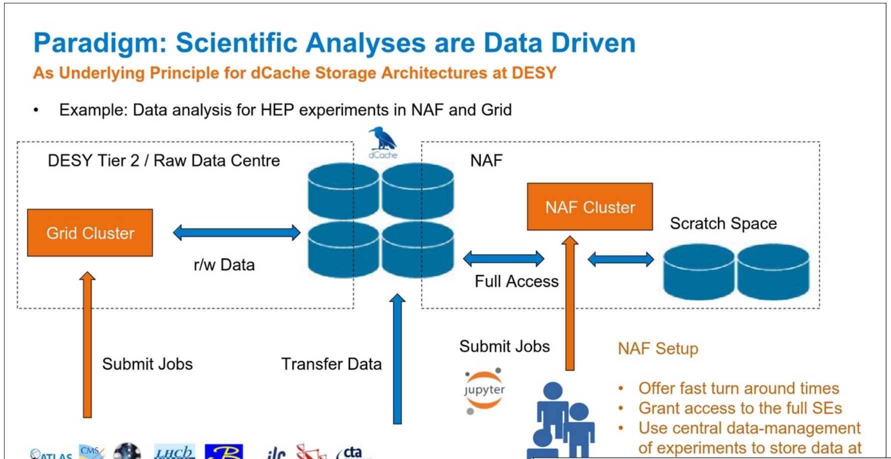
Full Kafka Billing Stream Workflow

Online and Offline Analysis Tools

DESY.

Page 17

dCache at DESY



Paradigm: Scientific Analyses are Data Driven

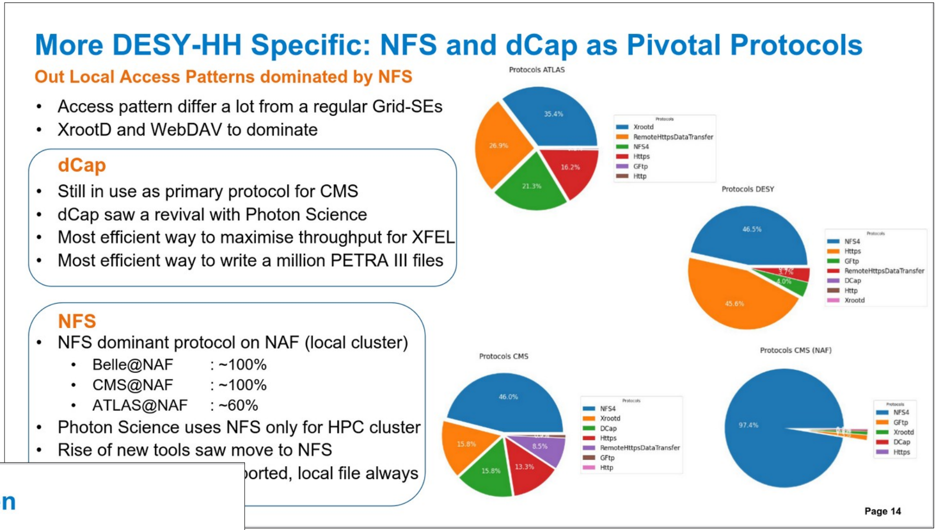
As Underlying Principle for dCache Storage Architectures at DESY

- Example: Data analysis for HEP experiments in NAF and Grid

NAF Setup

- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at

DESY.



More DESY-HH Specific: NFS and dCap as Pivotal Protocols

Out Local Access Patterns dominated by NFS

- Access pattern differ a lot from a regular Grid-SEs
- XrootD and WebDAV to dominate

dCap

- Still in use as primary protocol for CMS
- dCap saw a revival with Photon Science
- Most efficient way to maximise throughput for XFEL
- Most efficient way to write a million PETRA III files

NFS

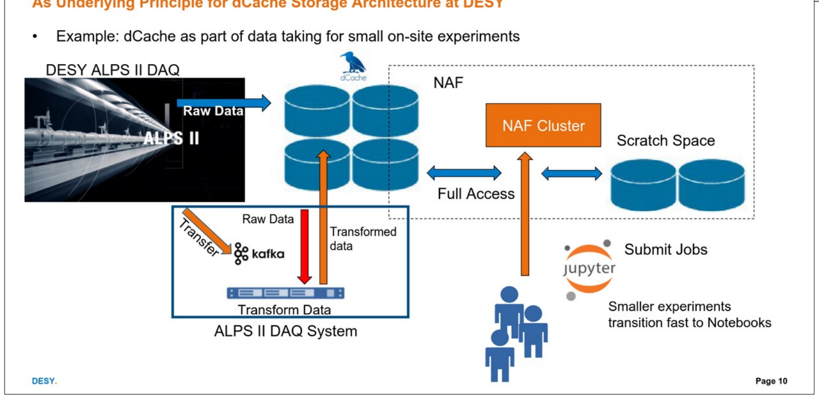
- NFS dominant protocol on NAF (local cluster)
  - Belle@NAF : ~100%
  - CMS@NAF : ~100%
  - ATLAS@NAF : ~60%
- Photon Science uses NFS only for HPC cluster
- Rise of new tools saw move to NFS
- ported, local file always

Page 14

Paradigm: Scientific Analyses are Data Driven

As Underlying Principle for dCache Storage Architecture at DESY

- Example: dCache as part of data taking for small on-site experiments

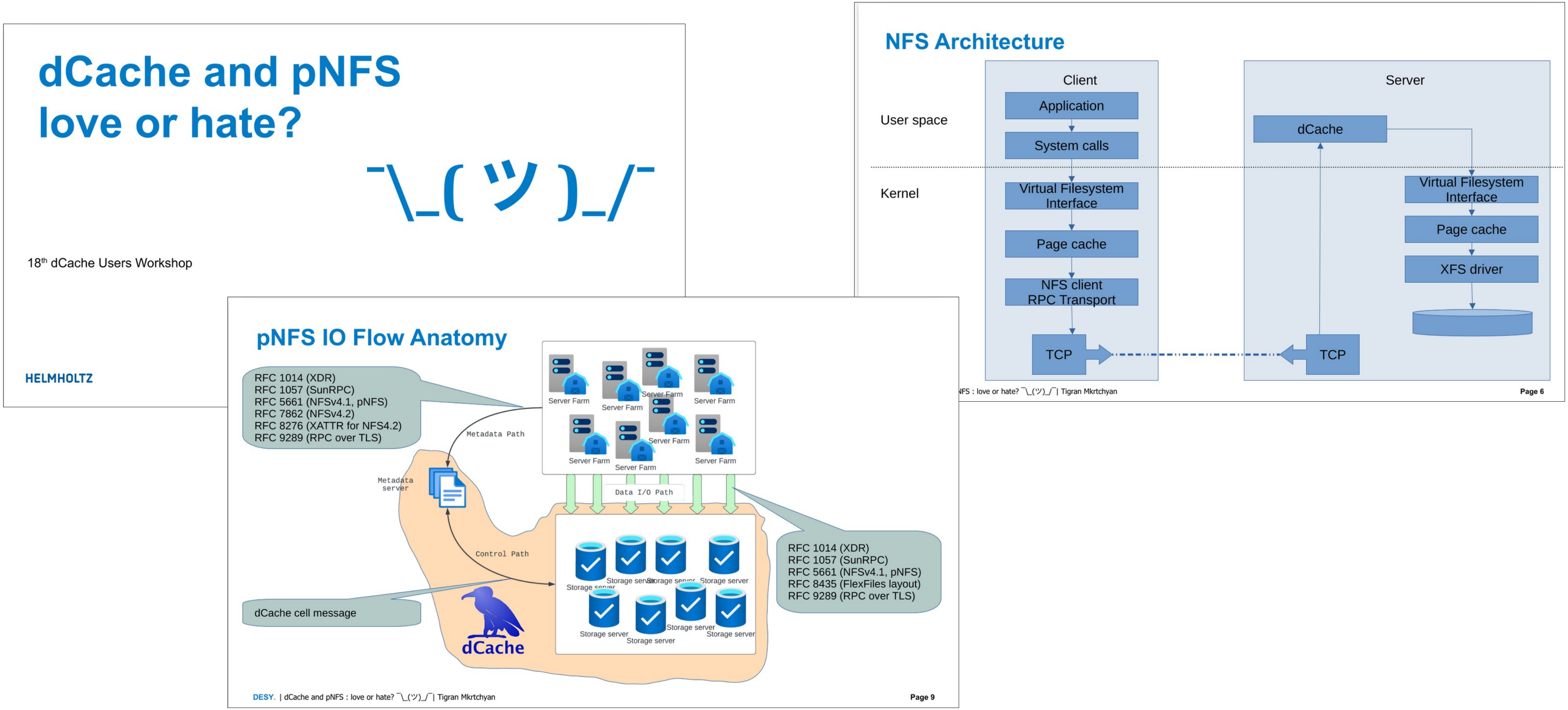


Smaller experiments transition fast to Notebooks

DESY.

Page 10

dCache with pNFS



dCache and pNFS love or hate?

¯_(ツ)_/¯

18th dCache Users Workshop

HELMHOLTZ

pNFS IO Flow Anatomy

RFC 1014 (XDR)
RFC 1057 (SunRPC)
RFC 5661 (NFSv4.1, pNFS)
RFC 7862 (NFSv4.2)
RFC 8276 (XATTR for NFS4.2)
RFC 9289 (RPC over TLS)

Server Farm
Metadata Path
Metadata server
Data I/O Path
Control Path
Storage server

dCache cell message

dCache

RFC 1014 (XDR)
RFC 1057 (SunRPC)
RFC 5661 (NFSv4.1, pNFS)
RFC 8435 (FlexFiles layout)
RFC 9289 (RPC over TLS)

DESY. | dCache and pNFS : love or hate? ¯_(ツ)_/¯ | Tigran Mkrtchyan

Page 9

NFS Architecture

User space
Kernel

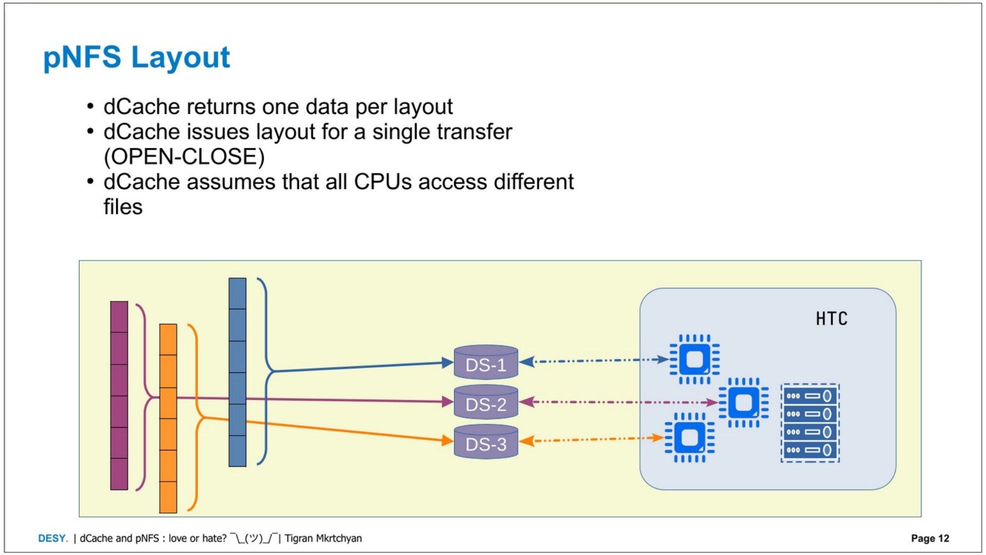
Client
Application
System calls
Virtual Filesystem Interface
Page cache
NFS client RPC Transport
TCP

Server
dCache
Virtual Filesystem Interface
Page cache
XFS driver
TCP

FS : love or hate? ¯_(ツ)_/¯ | Tigran Mkrtchyan

Page 6

dCache with pNFS

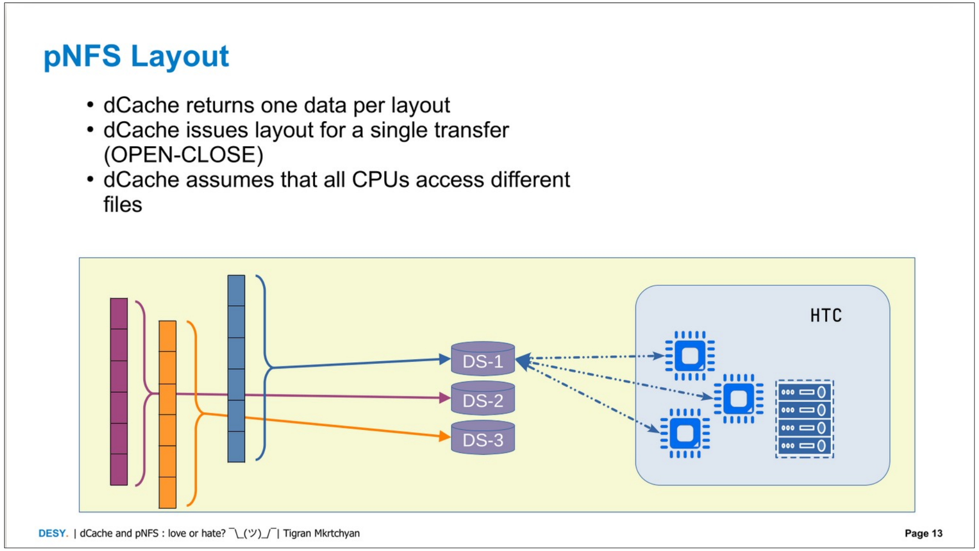


pNFS Layout

- dCache returns one data per layout
- dCache issues layout for a single transfer (OPEN-CLOSE)
- dCache assumes that all CPUs access different files

DESY. | dCache and pNFS : love or hate? ¯_(ツ)_/¯| Tigran Mkrtchyan

Page 12



pNFS Layout

- dCache returns one data per layout
- dCache issues layout for a single transfer (OPEN-CLOSE)
- dCache assumes that all CPUs access different files

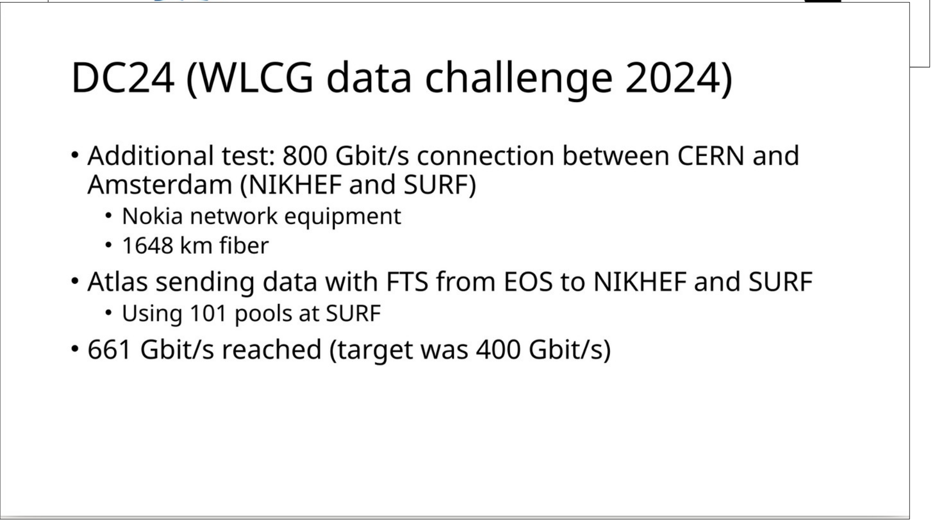
DESY. | dCache and pNFS : love or hate? ¯_(ツ)_/¯| Tigran Mkrtchyan

Page 13

dCache at SURF

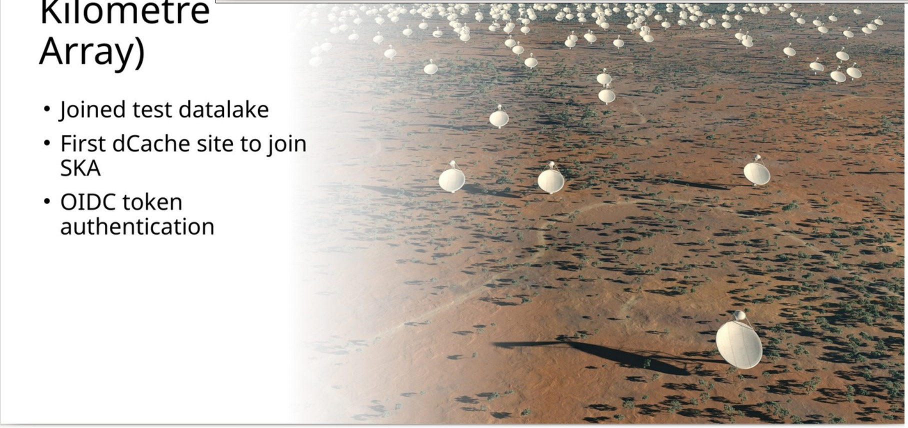
What went wrong and how we fixed some of it

Onno Zweers – dCache Workshop – 2024-05-07



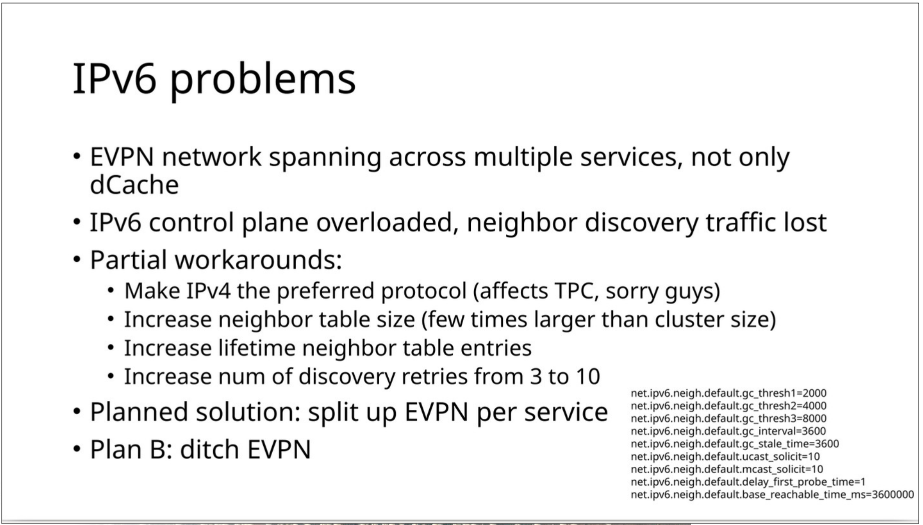
DC24 (WLCG data challenge 2024)

- Additional test: 800 Gbit/s connection between CERN and Amsterdam (NIKHEF and SURF)
  - Nokia network equipment
  - 1648 km fiber
- Atlas sending data with FTS from EOS to NIKHEF and SURF
  - Using 101 pools at SURF
- 661 Gbit/s reached (target was 400 Gbit/s)



SKA (Squa Kilometre Array)

- Joined test datalake
- First dCache site to join SKA
- OIDC token authentication

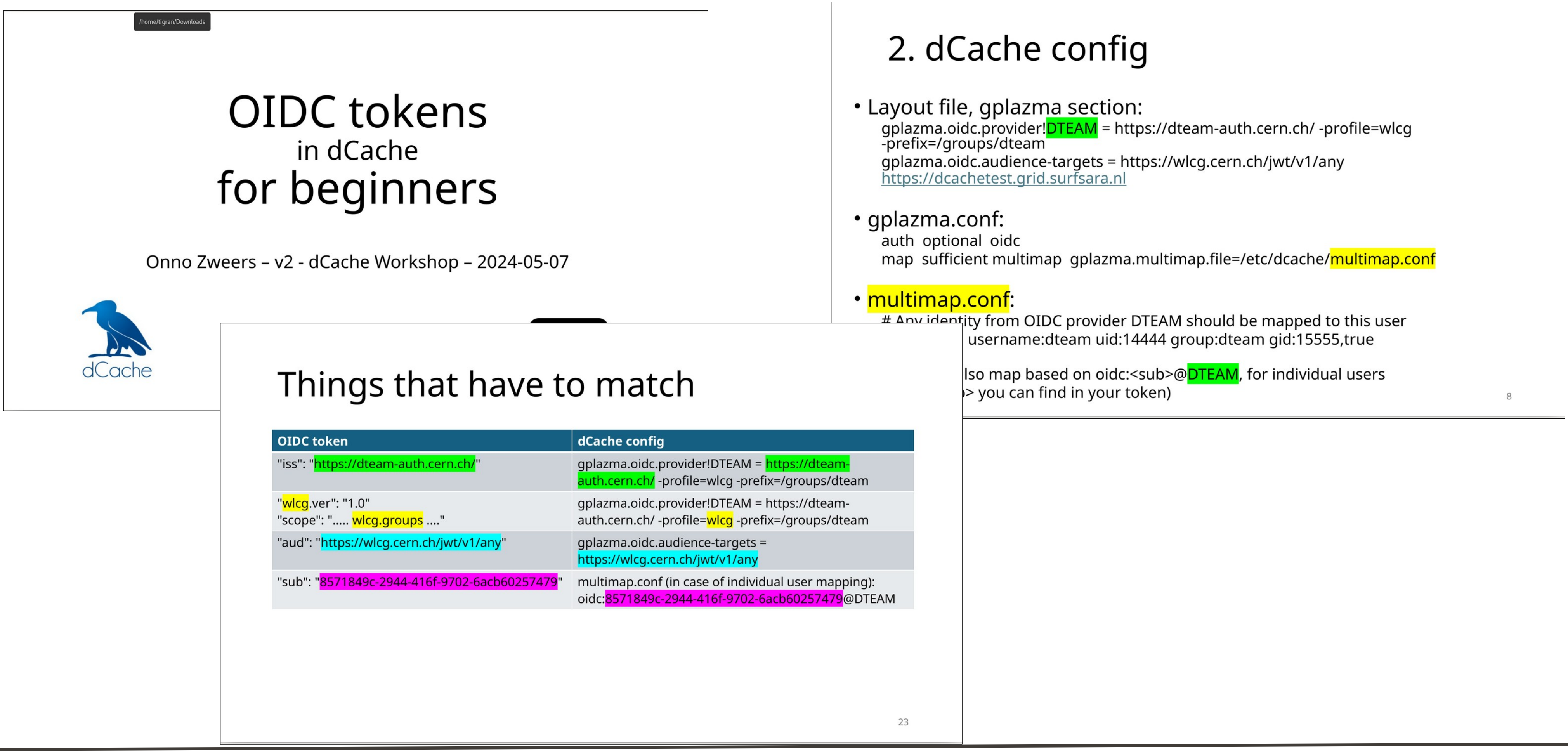


IPv6 problems

- EVPN network spanning across multiple services, not only dCache
- IPv6 control plane overloaded, neighbor discovery traffic lost
- Partial workarounds:
  - Make IPv4 the preferred protocol (affects TPC, sorry guys)
  - Increase neighbor table size (few times larger than cluster size)
  - Increase lifetime neighbor table entries
  - Increase num of discovery retries from 3 to 10
- Planned solution: split up EVPN per service
- Plan B: ditch EVPN

```
net.ipv6.neigh.default.gc_thresh1=2000
net.ipv6.neigh.default.gc_thresh2=4000
net.ipv6.neigh.default.gc_thresh3=8000
net.ipv6.neigh.default.gc_interval=3600
net.ipv6.neigh.default.gc_stale_time=3600
net.ipv6.neigh.default.ucast_solicit=10
net.ipv6.neigh.default.mcast_solicit=10
net.ipv6.neigh.default.delay_first_probe_time=1
net.ipv6.neigh.default.base_reachable_time_ms=3600000
```

Tokens…



OIDC tokens
in dCache
for beginners

Onno Zweers – v2 - dCache Workshop – 2024-05-07

2. dCache config

- Layout file, gplazma section:
  gplazma.oidc.provider!DTEAM = https://dteam-auth.cern.ch/ -profile=wlcg -prefix=/groups/dteam
  gplazma.oidc.audience-targets = https://wlcg.cern.ch/jwt/v1/any https://dcachetest.grid.surfsara.nl
- gplazma.conf:
  auth optional oidc
  map sufficient multimap gplazma.multimap.file=/etc/dcache/multimap.conf
- multimap.conf:
  # Any identity from OIDC provider DTEAM should be mapped to this user
  username:dteam uid:14444 group:dteam gid:15555,true
  lso map based on oidc:<sub>@DTEAM, for individual users
  > you can find in your token)

8

Things that have to match

OIDC token	dCache config
"iss": "https://dteam-auth.cern.ch/"	gplazma.oidc.provider!DTEAM = https://dteam-auth.cern.ch/ -profile=wlcg -prefix=/groups/dteam
"wlcg.ver": "1.0" "scope": "….. wlcg.groups …."	gplazma.oidc.provider!DTEAM = https://dteam-auth.cern.ch/ -profile=wlcg -prefix=/groups/dteam
"aud": "https://wlcg.cern.ch/jwt/v1/any"	gplazma.oidc.audience-targets = https://wlcg.cern.ch/jwt/v1/any
"sub": "8571849c-2944-416f-9702-6acb60257479"	multimap.conf (in case of individual user mapping): oidc:8571849c-2944-416f-9702-6acb60257479@DTEAM

23

Questions?