dCache.org

Active storage for science and cloud

Tigran Mkrtchyan for dCache People

CS3 2019, Roma

neic Nordic e-Infrastructure Collaboration

eXtreme DataCloud

DESY.

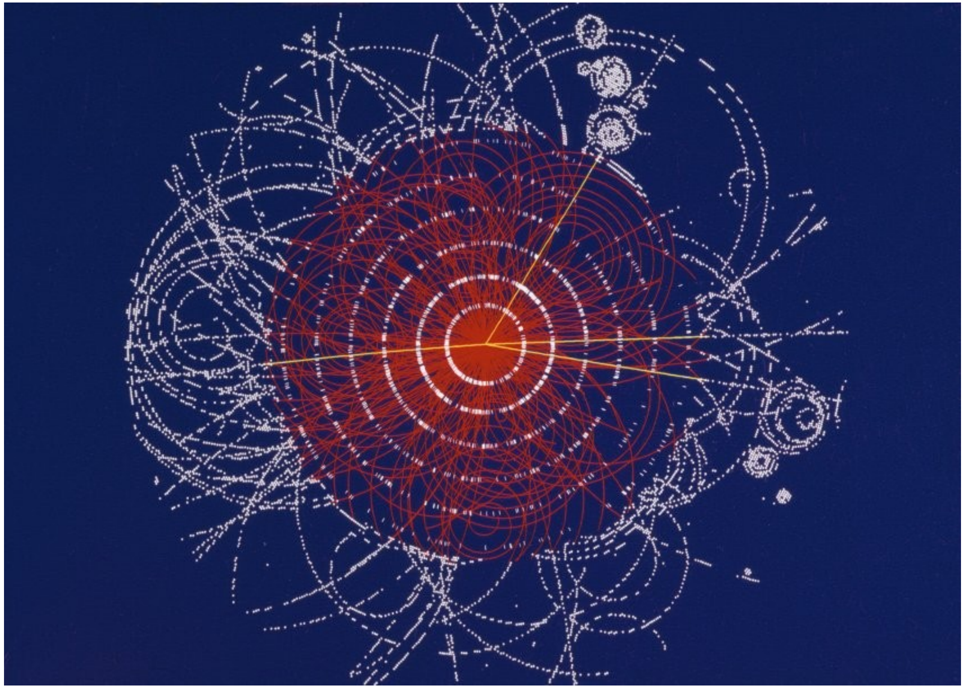
LSDMA

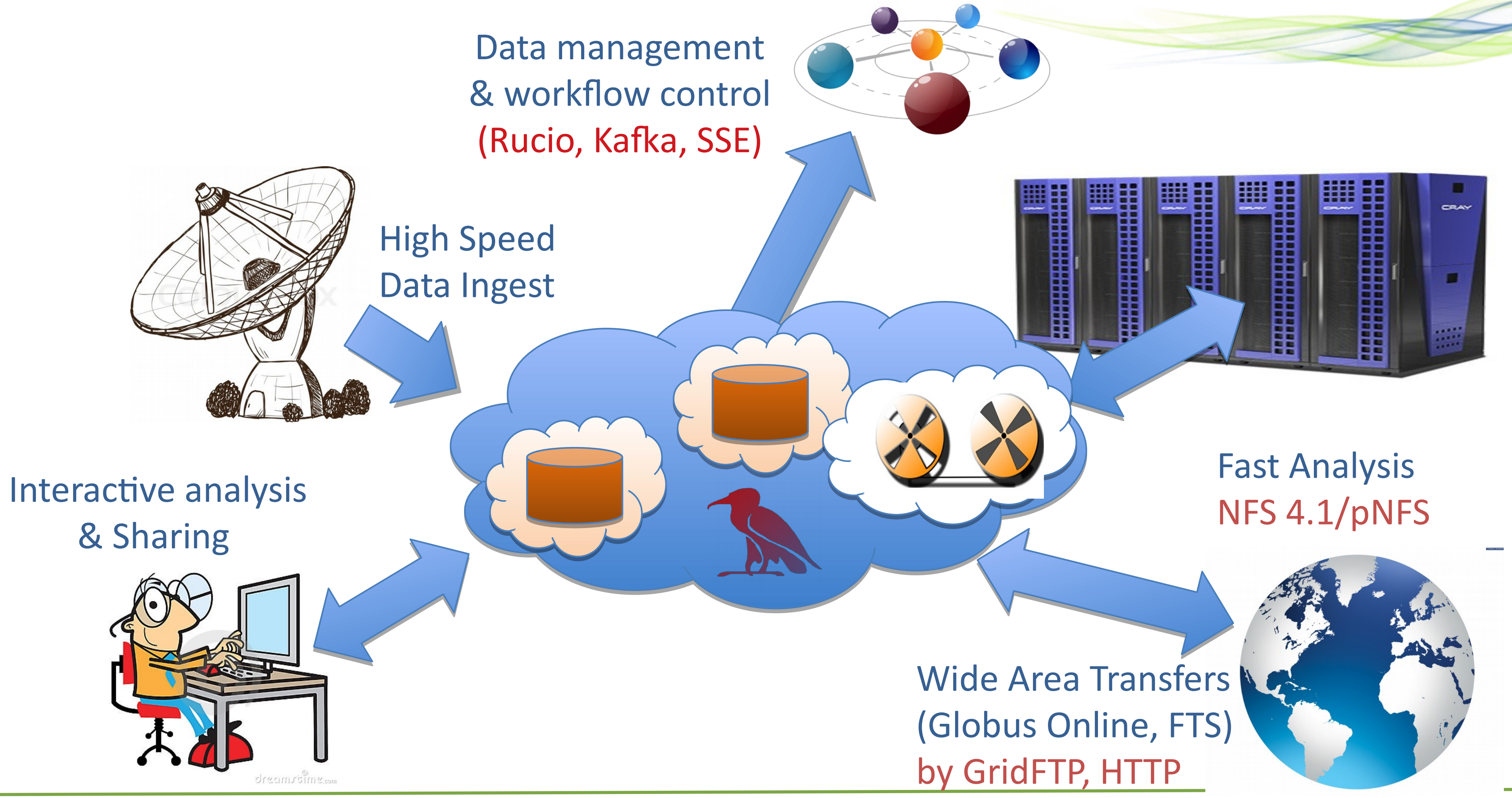
Fermilab

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Scientific data challenges

- Volume
- Fast ingest
- Chaotic Access
- Sharing
- Access Control
- Persistence & Long term archival
- Immutability





Data management & workflow control (Rucio, Kafka, SSE)

High Speed Data Ingest

Interactive analysis & Sharing

Fast Analysis NFS 4.1/pNFS

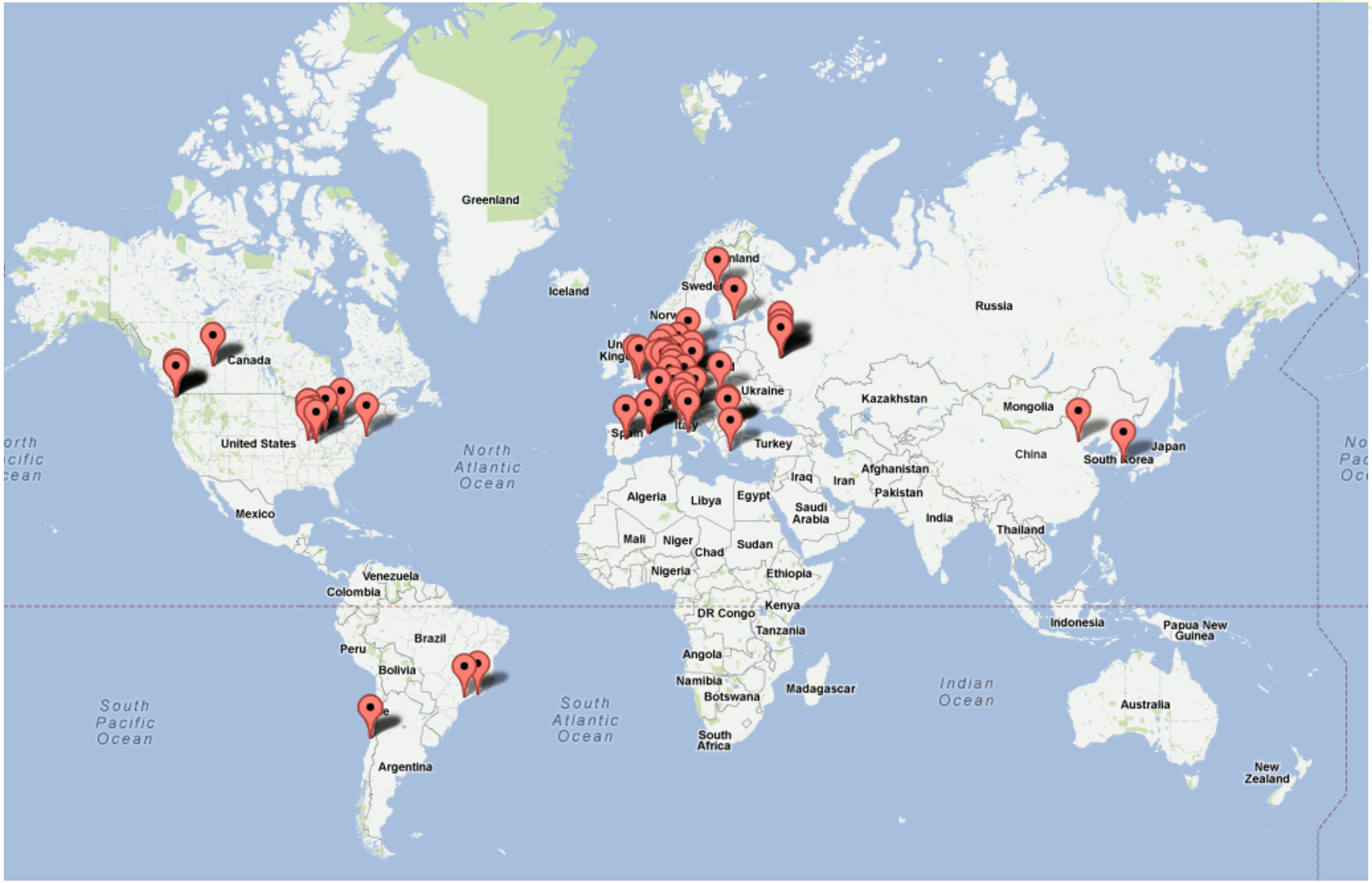
Wide Area Transfers (Globus Online, FTS) by GridFTP, HTTP

dCache 101: Motivation

- Data never fits into a single server
    - Multiple servers
    - Off-load to tape
- Growing number of client hosts
    - Main frame vs. Linux cluster
- Control over HW/OS selection
    - Better offers
    - Local expertise

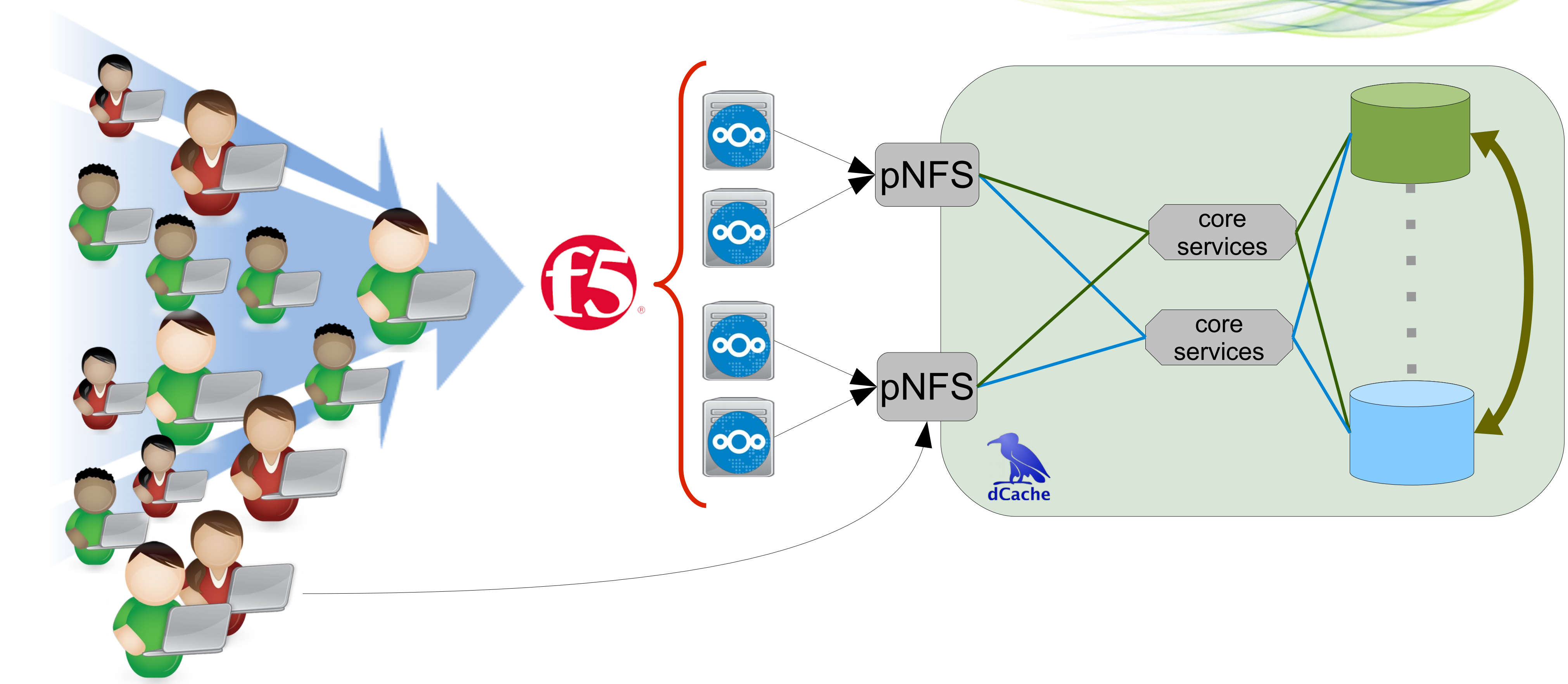
dCache 101: design

- Single-rooted namespace, distributed data
- Client talks to namespace for metadata operations only
- Bandwidth and performance grow with number of data servers
- Standard clients (OS native or experiment framework)
- Same data can be provided by any access protocol and security flavor



- HERA
- Tevatron
- WLCG
- Belle II
- LOFAR
- CTA
- IceCUBE
- EU-XFEL
- Petra3
- DUNE
- And many more ...

HA-nextCloud instance @ DESY

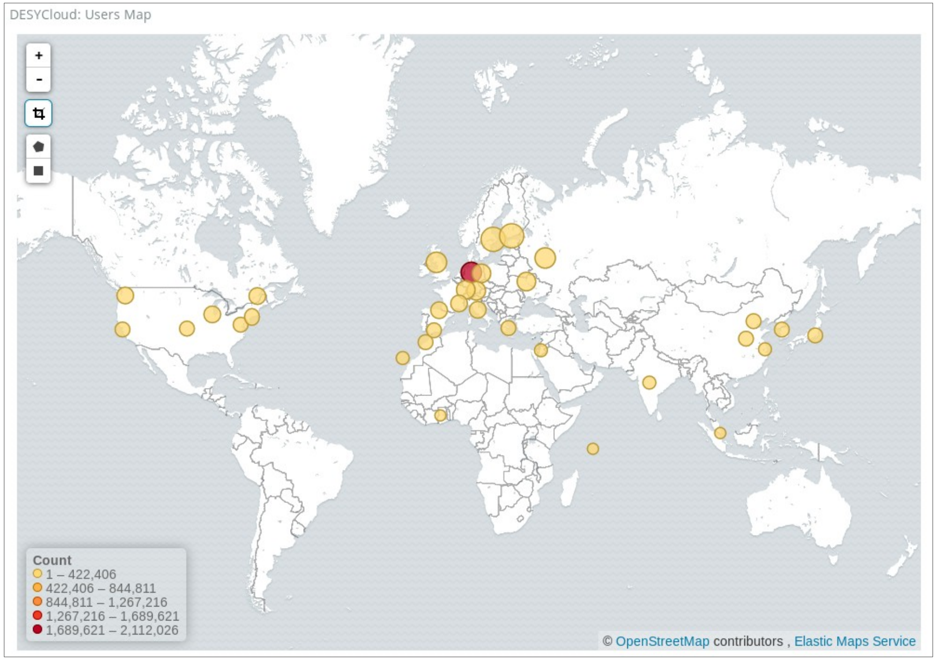


dCache as a storage backend

- PB-scale storage system
- No changes in nextCloud required
- Unique functionality
  - Tape integration
  - File ownership preservation
  - NFS export to selected users
  - Storage events
  - Data visible by all protocols and security flavors
- *Not standard dCache version (due to special configuration)*
  - WIP to make it a part of standard package

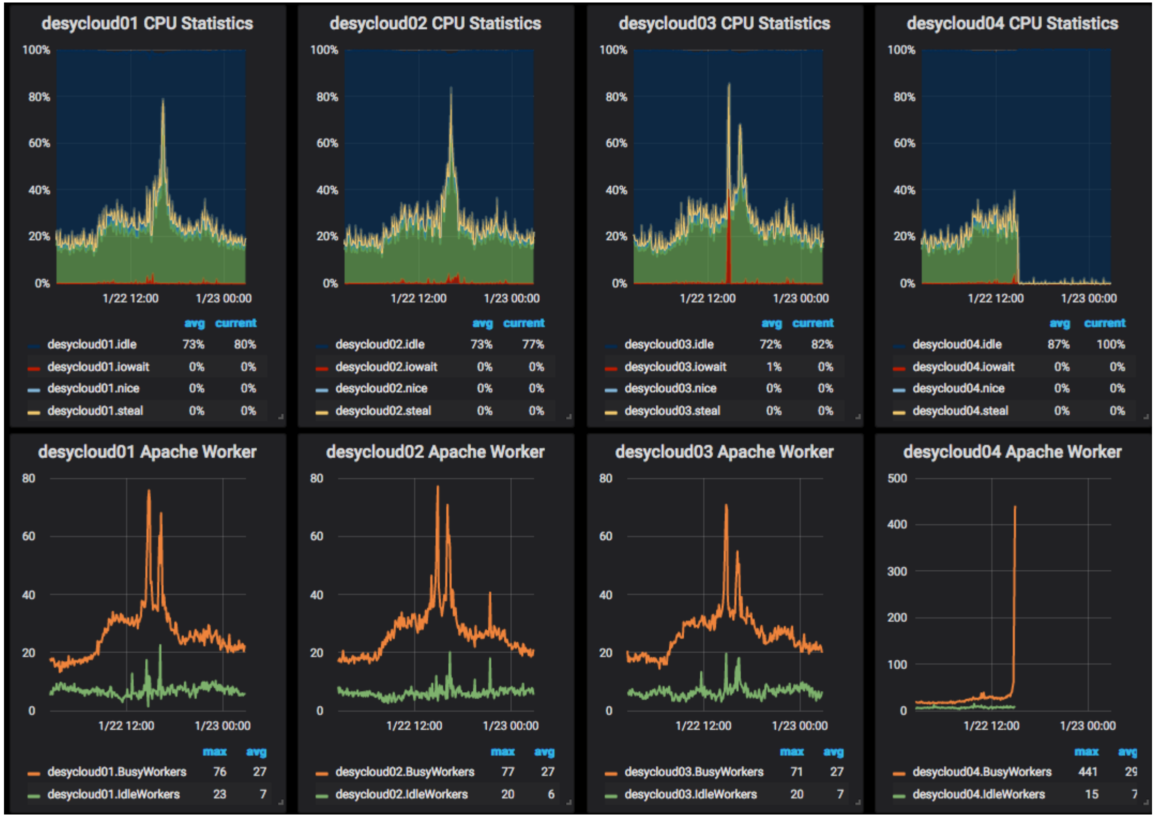
HA-nextCloud instance @ DESY

- x6 nextCloud front-ends
  - load-balanced with F5
  - two groups on different NFS servers
- x2 dCache-NFS servers
- x4 Physical data servers
  - 32 logical servers (dCache pools)
- x3 dCache core services
  - Hot stand-by namespace-DB replica
- 300TB installed capacity, 30TB used
  - ~ 53M stored files, x2 copies per file
  - ~ 95K new files per day (50% updates)
  - ~ 50K removed


DESYCloud: Users Map

HA-nextCloud instance @ DESY

- zero downtime
  - software updates
  - OS/HW updates
- Handles unexpected crashes



desycloud01 CPU Statistics

	avg	current
desycloud01.idle	73%	80%
desycloud01.iowait	0%	0%
desycloud01.nice	0%	0%
desycloud01.steal	0%	0%

desycloud02 CPU Statistics

	avg	current
desycloud02.idle	73%	77%
desycloud02.iowait	0%	0%
desycloud02.nice	0%	0%
desycloud02.steal	0%	0%

desycloud03 CPU Statistics

	avg	current
desycloud03.idle	72%	82%
desycloud03.iowait	1%	0%
desycloud03.nice	0%	0%
desycloud03.steal	0%	0%

desycloud04 CPU Statistics

	avg	current
desycloud04.idle	87%	100%
desycloud04.iowait	0%	0%
desycloud04.nice	0%	0%
desycloud04.steal	0%	0%

desycloud01 Apache Worker

	max	avg
desycloud01.BusyWorkers	76	27
desycloud01.IdleWorkers	23	7

desycloud02 Apache Worker

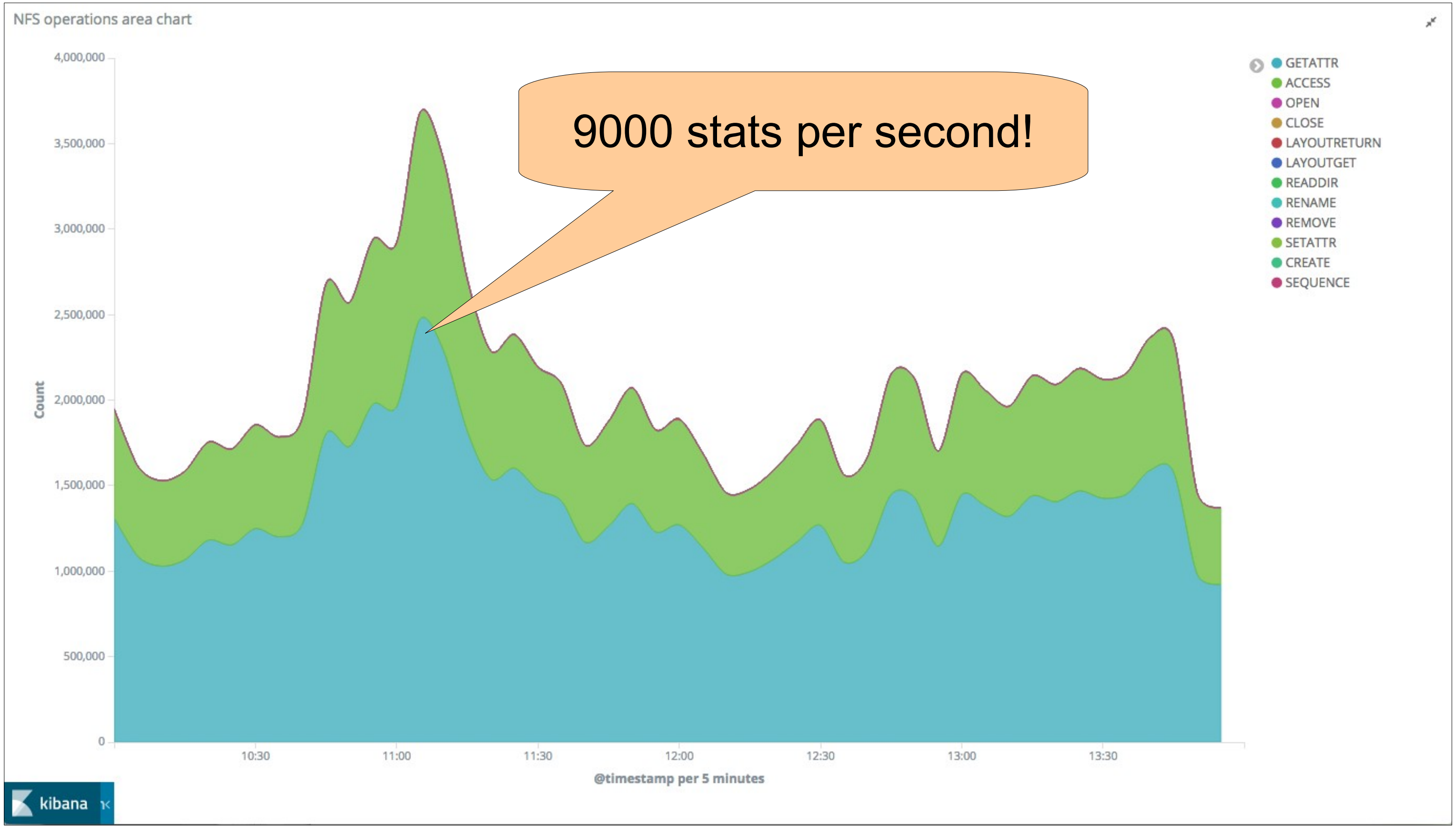
	max	avg
desycloud02.BusyWorkers	77	27
desycloud02.IdleWorkers	20	6

desycloud03 Apache Worker

	max	avg
desycloud03.BusyWorkers	71	27
desycloud03.IdleWorkers	20	7

desycloud04 Apache Worker

	max	avg
desycloud04.BusyWorkers	441	29
desycloud04.IdleWorkers	15	7



9000 stats per second!

ARE WE THERE YET !?!
MATT GROENING

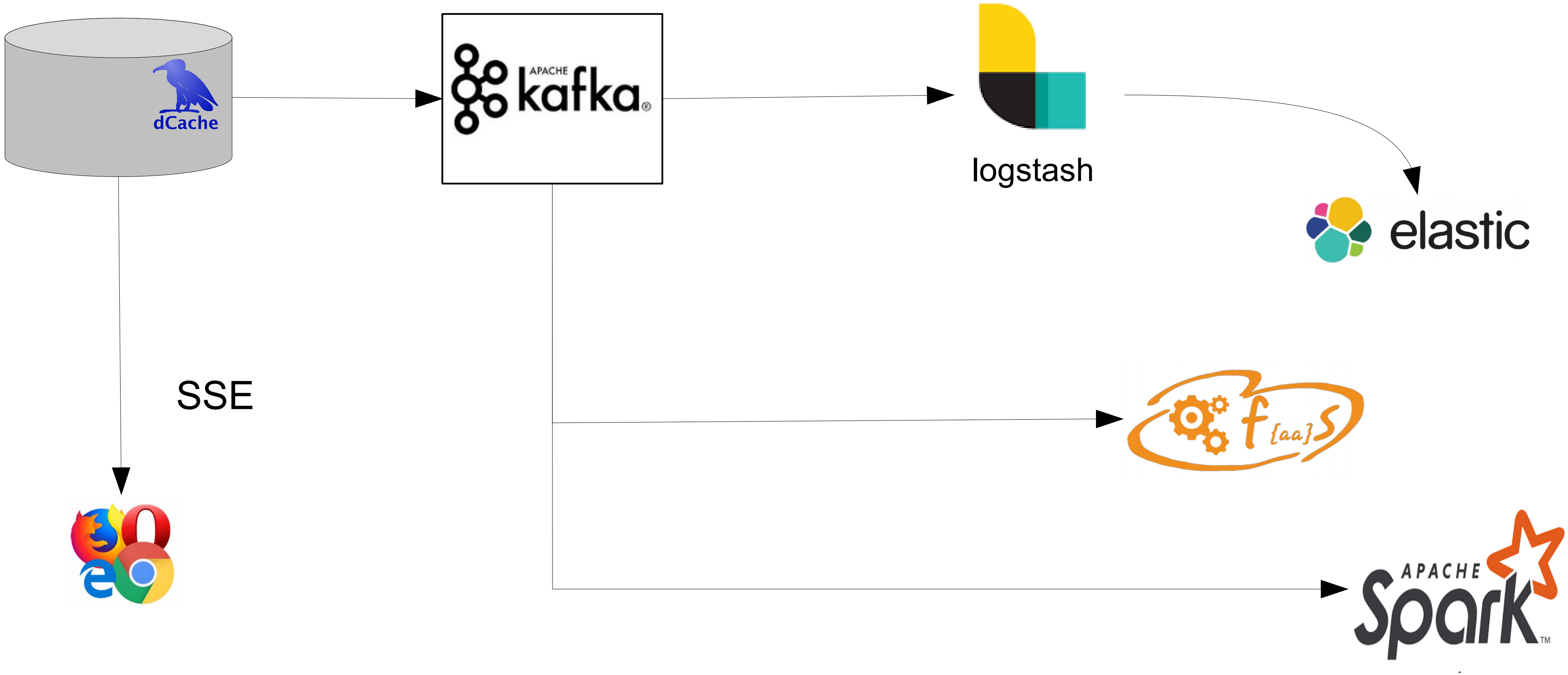
Storage events

- Storage system becomes a workflow engine
- Trigger actions on user activity
  - Stop polling, Please!
- System-global events
- Per user events (inotify)

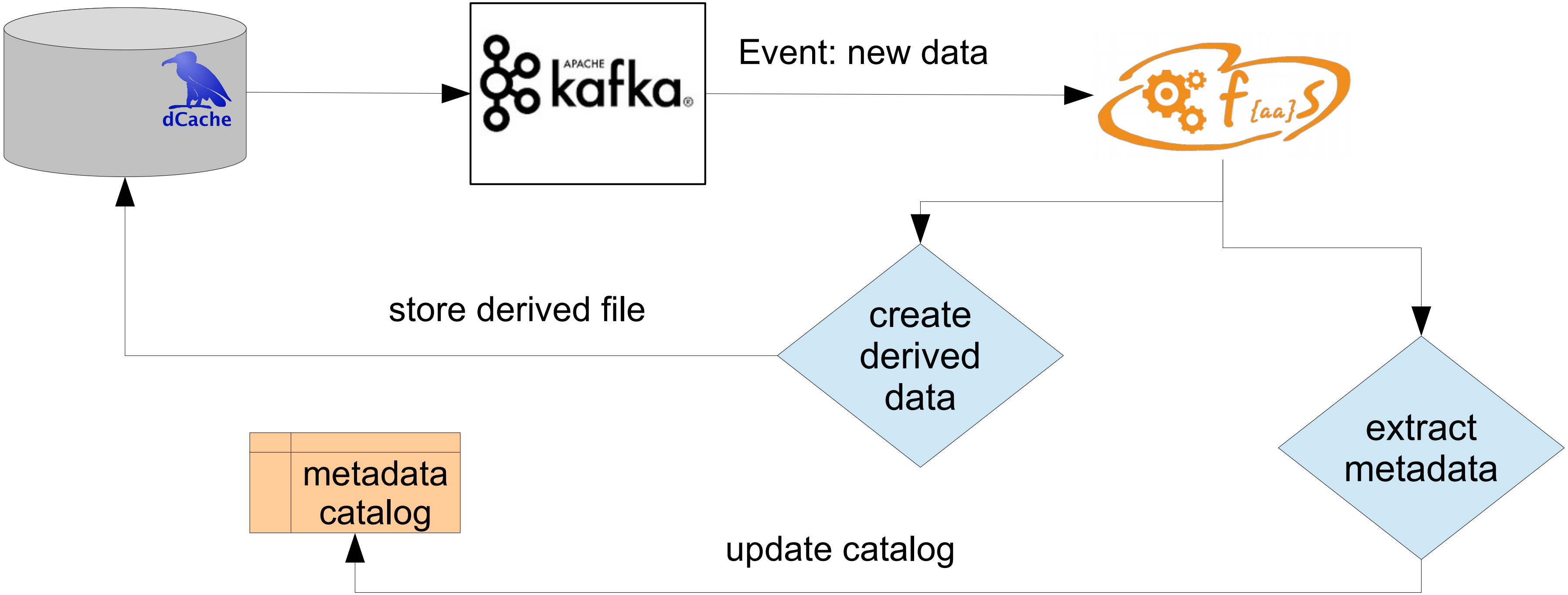
Storage events in dCache

- Kafka stream
  - Producer-consumer model
  - Kafka consumer is required
  - global events
  - Consumer keeps track of the last seen event
- Server-Send Events (SSE)
  - Producer-consumer model
  - HTTP connection “for receiving push notifications from a server“
  - User specific event stream
  - Client keeps track of the “Last-Event-ID“

Storage events



Workflow control



Event: new data

store derived file

create derived data

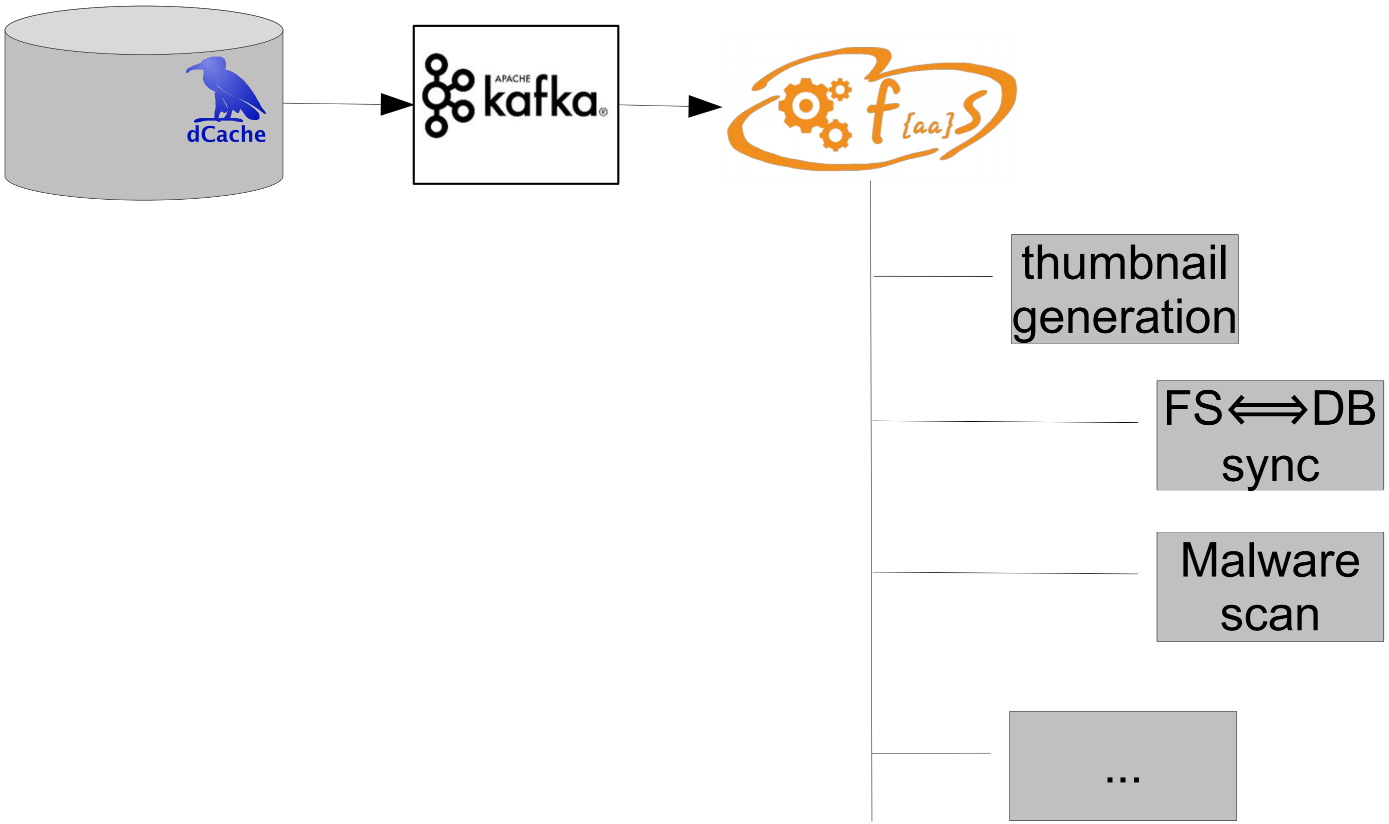
extract metadata

metadata catalog

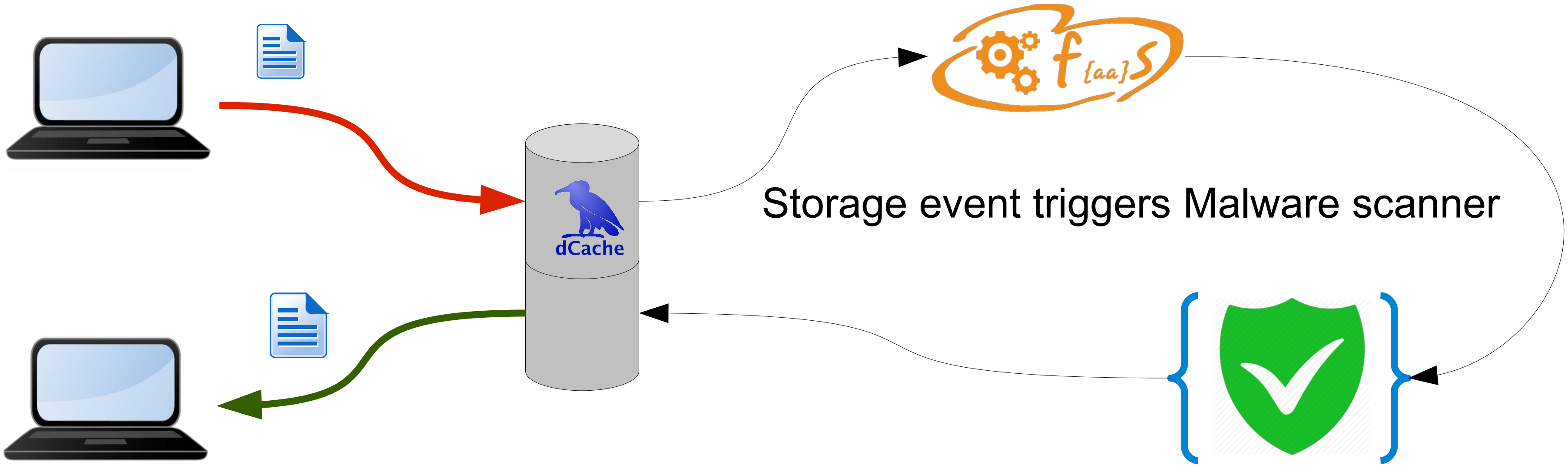
update catalog

by Michael Schuh (XFEL Data Ingesting and Processing in the EOSC)

Event Processing with FaaS



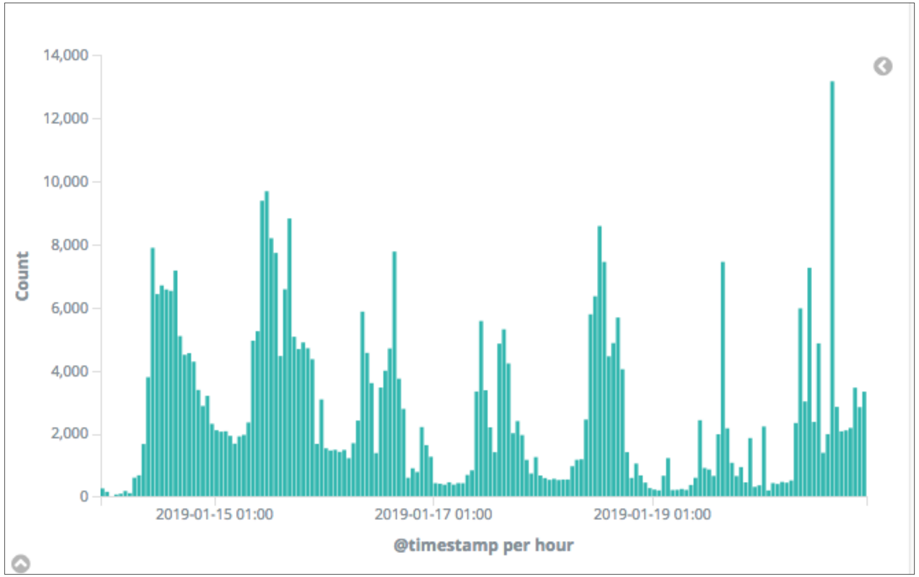
Processing with FaaS (WIP)



Storage event triggers Malware scanner

FaaS (Server less)

- State less
  - persistence in the storage
- Scales on demand
  - make idle resources available to others
- Available with public clouds
  - AWS: S3 + Amazon Lambda
  - Azure: S3 + Azure functions



Summary

- dCache is a widely used storage system for scientific data.
- HA setup and unique features make is attractive as a backend for sync'n'share services.
- Storage events allow workflow integration with cloud.
- Non scientific data has new requirements
  - extended reliability
  - end-to-end encryption
  - additional data safety
- Better integration into sync'n'share software required to expose full potential and reduce functionality duplication.

13‘th dCache user workshop May 21-22, Madrid

More info: https://www.dcache.org