SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

eXtreme DataCloud: Providing Scalable Distributed Storage for the European Open Science Cloud

Patrick Fuhrmann

For the European Union H2020 eXtreme Data Cloud Project

Deutsches Elektronen Synchrotron

“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777367”.

Storyboard

EOSC, the European Open Science Cloud

- The European Open Science Cloud vision : “to give Europe a global lead in scientific data infrastructures and to ensure that European scientists reap the full benefits of data-driven science”. *European Cloud Initiative publication*
- Essentially all about FAIR data.
  - https://www.egi.eu/about/newsletters/what-is-the-european-open-science-cloud/

A slice of the EOSC Architecture

Jan 1, 2019

Infrastructure & Governance

EUDAT

EGI

EOSCpilot
The European Open Science Cloud for Research Pilot Project

Software & Integration

INDIGO - DataCloud
Better Software for Better Science

eXtreme DataCloud

Hybrid DataCloud

Pre-Procurement

HELIX NEBULA
THESCIENCECLOUD

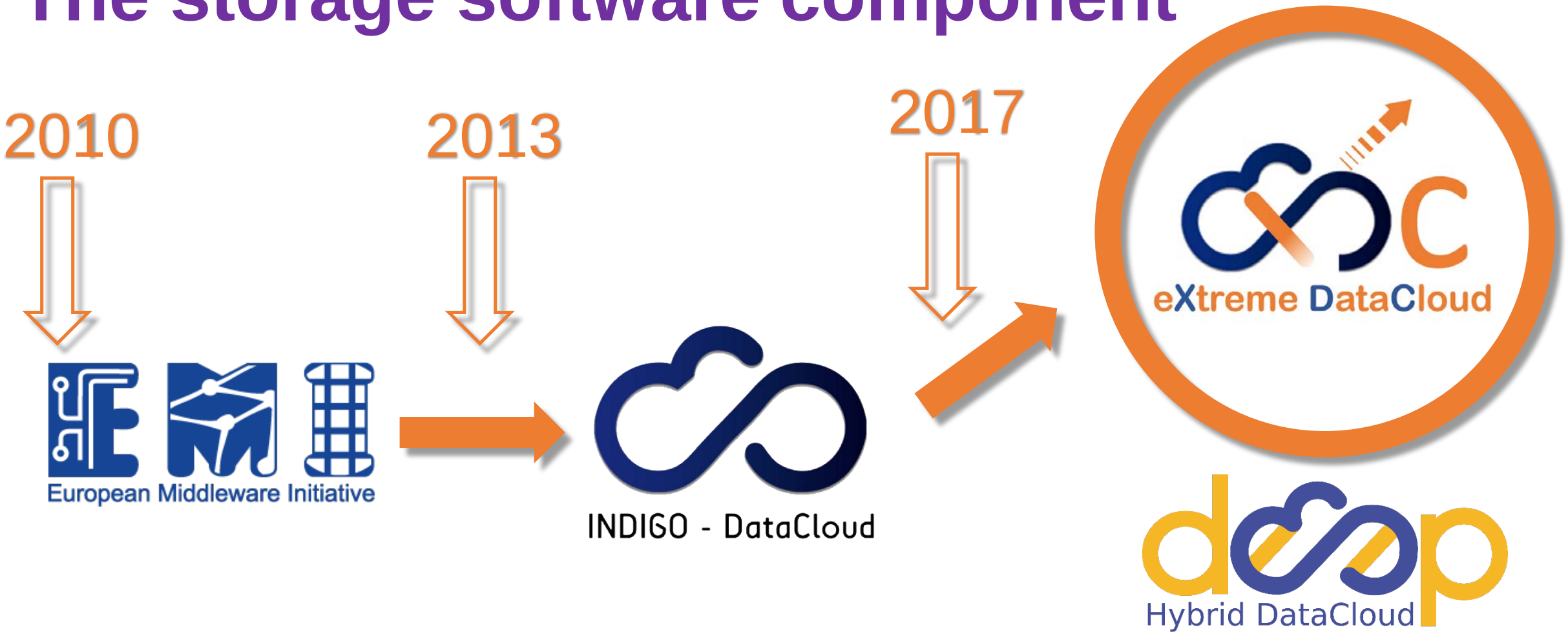
EOSC-hub

ESCAPE
(High Energy Physics)

PaN-OSC
(Photon Neutron)

ARCHIVER
(Pre-Procured Archiving Infrastructures)

The storage software component



The eXtreme DataCloud Cheat Sheet

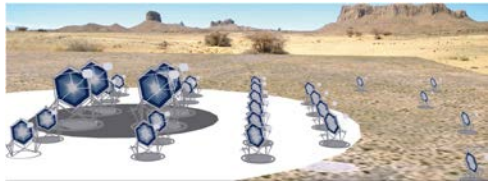
- 8 partners
- 7 countries (DE, IT, ES, PL, NL,UK, FR)
- 7 research communities represented + EGI
- XDC Total Budget: 3.07M Euros
- XDC (27 Months)
  - started Nov 1st 2017
  - until Jan 31st 2020

The eXtreme DataCloud Cheat Sheet (cont.)

- eXtreme DataCloud is a software development and integration project.
- Develops scalable technologies for federating storage resources and managing data in highly distributed computing environments.
  - Focus efficient, policy driven and Quality of Service based DM
- The targeted platforms are the current and next generation e-Infrastructures deployed in Europe.
  - European Open Science Cloud (EOSC)
  - The e-infrastructures used by the represented communities

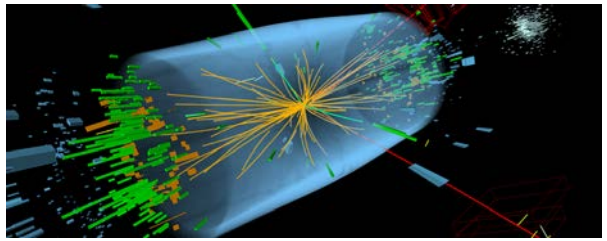
The XDC research communities

Astronomy



Photon Science

High Energy Physics



Bio Ecosystem

Medicine

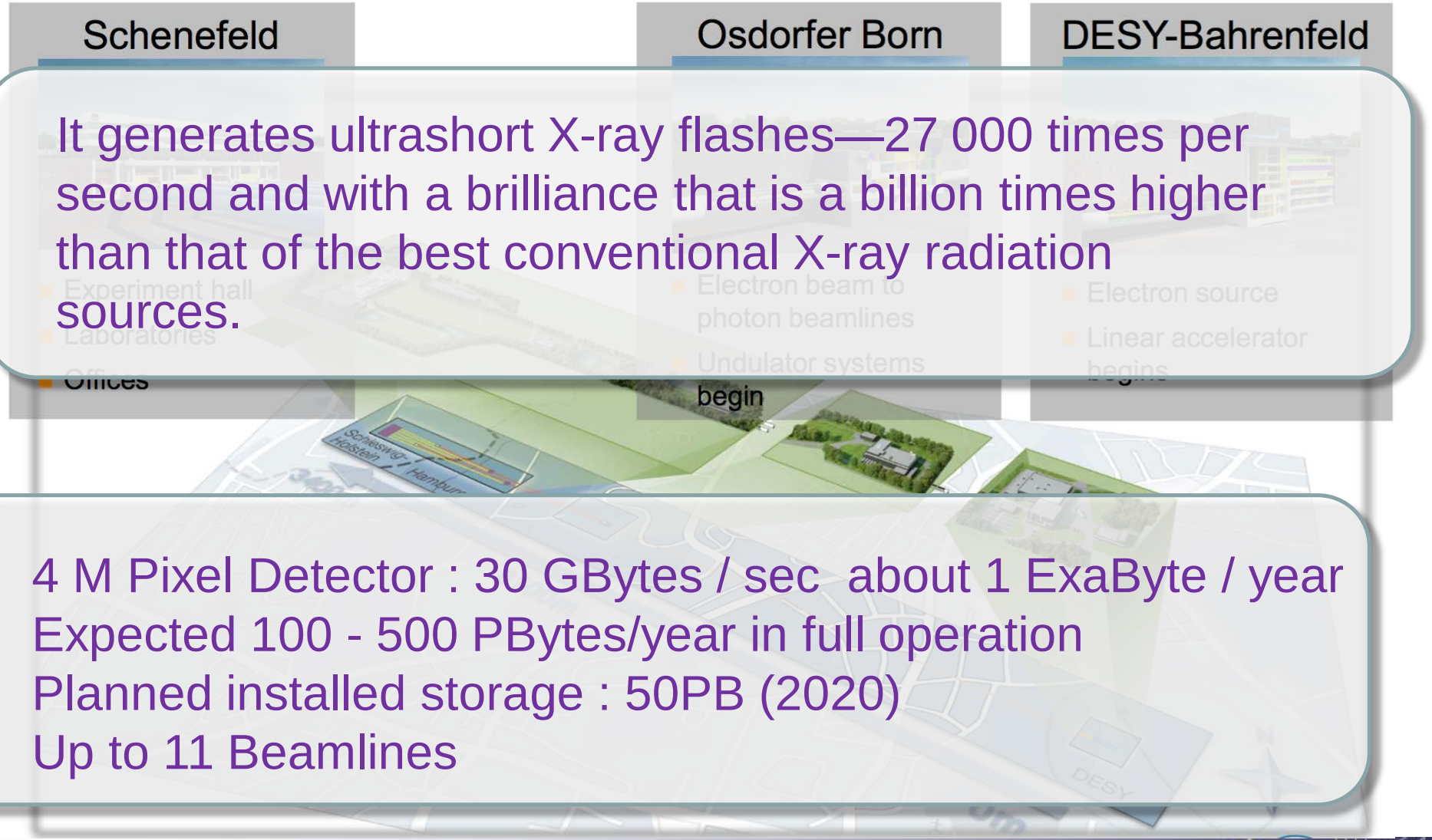
Joined Research Work Package 4 of XDC

- Implementing a configurable data workflow orchestration, in terms of data location and storage quality (QoS).
- Providing managed and unmanaged data caching services at all levels.
- Providing event based interfaces to external systems
  - Generating events to the XDC orchestration services when data is entering the XDC system.
  - Generating events to external compute clusters when data is ready to be processed.
- Federating heterogeneous data sources, building a virtual horizontal infrastructure-specific data space.

Starting with Storage Orchestration

Example : The European X-FEL

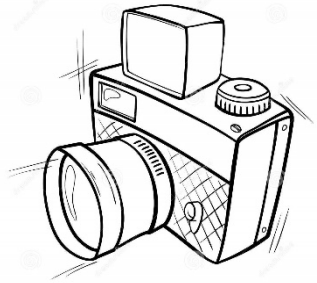
The European X-FEL Facility



It generates ultrashort X-ray flashes—27 000 times per second and with a brilliance that is a billion times higher than that of the best conventional X-ray radiation sources.

- 4 M Pixel Detector : 30 GBytes / sec about 1 ExaByte / year
- Expected 100 - 500 PBytes/year in full operation
- Planned installed storage : 50PB (2020)
- Up to 11 Beamlines

The European XFEL use case



TOSCA Template

Orchestrator

Online System

Burst Handling

Calibration

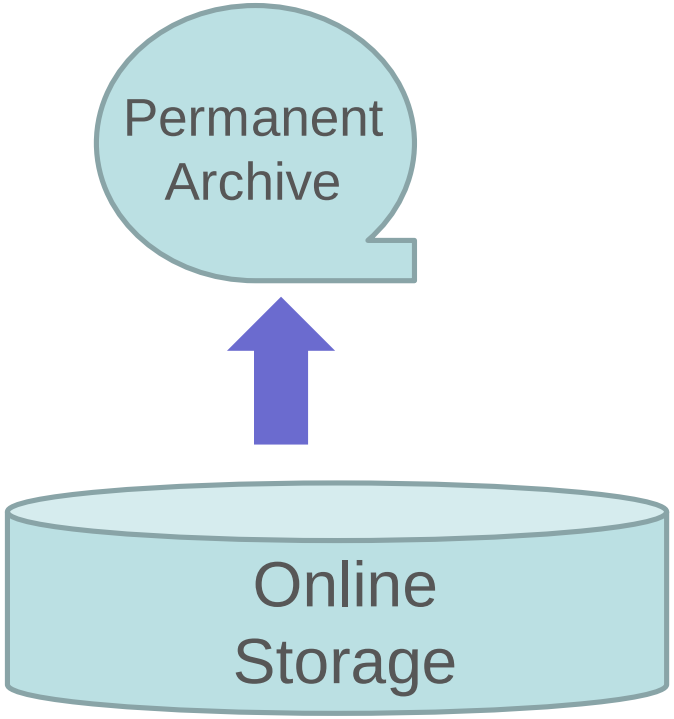
Permanent Archive

Individual Analysis

Online Storage

Introducing QoS in storage

- Technically : Copying data from a disk device to a custodial device (TAPE)
- Logically : Changing the Quality of the Storage to e.g. ‘super-durable’
- We come back to this in a moment.



Technologies used to make that work

- Storage Frontend : *IBM GPFS*
- Multi Tier Storage : *dCache*
- Storage Events : *Kafka and SSE*
- CMF and Containers : *Open Stack, Docker*
- Serverless (FaaS) : *OpenWhisk*
- Wide Area Transfers : *CERN File Transfer Service*
- Storage Federation : *CERN Dynafed*
- Orchestrator*: INDIGO Orchestrator and Rucio*

Now, coming back to Quality of Service in Storage

Example : The Worldwide LHC Computing Grid

QoS for the Worldwide LHC Compute Grid
(Status Quo)

- Worldwide distributed storage and compute resources.
  - Tier 1 : about 10 with disk and tape storage.
  - Tier 2 : about 250 with disk-only storage.
- WLCG can only afford a certain percentage of the total experiment data on disk.
- Mandatory: All data was available on tape at CERN plus two copies on tape at two Tier 1's.
- All other copies (disk or tape) were created and destroyed base on compute needs and monetary considerations.

QoS for WLCG: Disadvantages

- WLCG only knows two qualities: tape and disk.
- The quality description is tight to a storage technology, which is actually only indirectly associated to a quality.
- Central management of site-local storage is needed:
  - E.g. if you need to increase the ’retention policy’ at a site, a central service is requesting to copy the data from disk to tape.
- It’s very difficult to compare the actual costs of the different sites.
  - Sites are pledging the amount of disk and tape
  - There is no mentioning of the actual quality of the offer.
    - RAID versus JBOD
    - Or Universities can offer cheap disk storage, as it is maintained by students. However, those students only stay for a year, which makes the system extremely unreliable ’long term’.

QoS for WLCG: Better would be ...

- Defining a small set of qualities, independently of the storage technology used.
  - E.g. ‘super custodial’ could either be implemented as
    - Two tape copies in different storage system in different buildings.
    - Four disk copies on different disks arrays (different vendors) in different buildings.
- Each storage site would just publish the classes they support and the price of that class per unit. There would be no mentioning on how the infrastructure is implementing those classes.
- That would make accounting (billing) fair and transparent.
- On the technical level, if data needs to get a higher ‘retention’ or ‘access latency’ class, one only would have to change the ‘class’ of that data at that site (assuming the infrastructure supports that class)

What do we need to get this to work ?

- We need to agree on a first set of properties, like
  - Retention policy, access latency, locality
- We need a protocol or API to communicate those properties to the back-end storage.
- We need to provide a reference implementation for that protocol, supporting some storage back-ends.
- We need to build a demo infrastructure.

Engaging in the Research Data Alliance to find an appropriate vocabulary

- We are in the process to get an RDA working group created “Storage Definition WG”.
- We hope to collect realistic use cases for QoS in storage from a variety of science communities and infrastructures.
- We are using RDA to advertise our ideas.

Protocol to communicate QoS : CDMI

- Funding bodies (e.g. the European Commission) prefer to fund work based on standards.
- CDMI can be regarded a standard
- However, storage qualities were only available in CDMI in a rather rudimentary form.
- Transitions of QoS capabilities needed to be shoehorned into the protocol.

Protocol to communicate QoS : CDMI

- Consequently :
  - We became member of SNIA
  - We submitted our proposal for the extension
  - We called into the bi-weekly and discussed our proposed extensions.
  - We cleaned up the reference implementation
    - Now available in github (parallel to the original one)

The INDIGO CDMI Extension definitions

- Read capabilities objects using CDMI
  - e.g. the endpoint provides allowed capabilities
- CRUD of data objects or container objects using CDMI
  - e.g. Specifying the initial values of a file when writing the file.
- Change capabilities of data objects and container objects using CDMI
  - e.g. change the latency of objects within a container.
    - Can be used for “BRING ONLINE’ from custodial media.
  - cdmi_capabilities_target indicates: transition in progress

The INDIGO CDMI Extension definitions

What can you do for me ?

I support

'latency', 'location','retention'

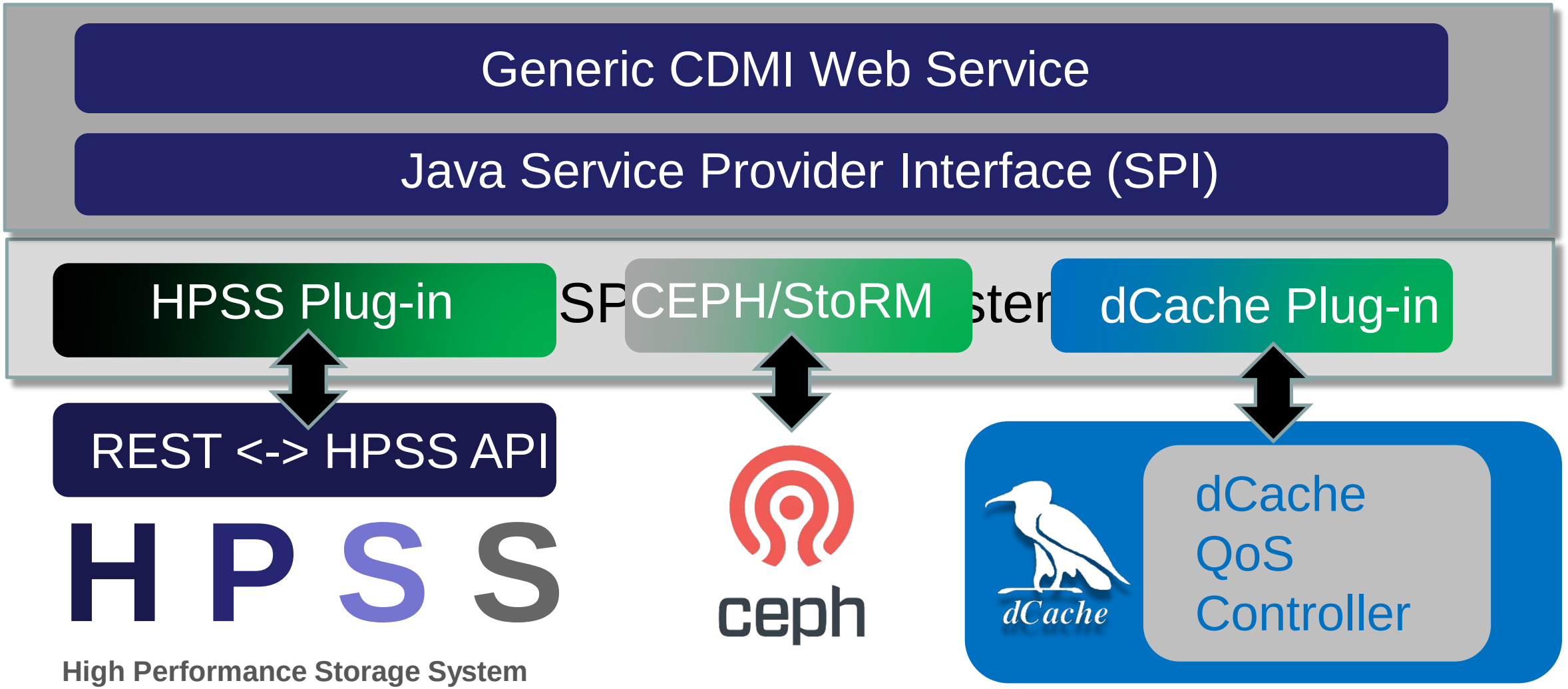
Please write my file with

Retention: custodial, Location: DE

Please change capability to 'online'

CDMI Extension Enabled Storage System

The CDMI Reference Implementation (INDIGO)



European Testbed

curl://

CMDI
Extension

Storage
Broker
Web Service

European Infrastructure

Storage Backends

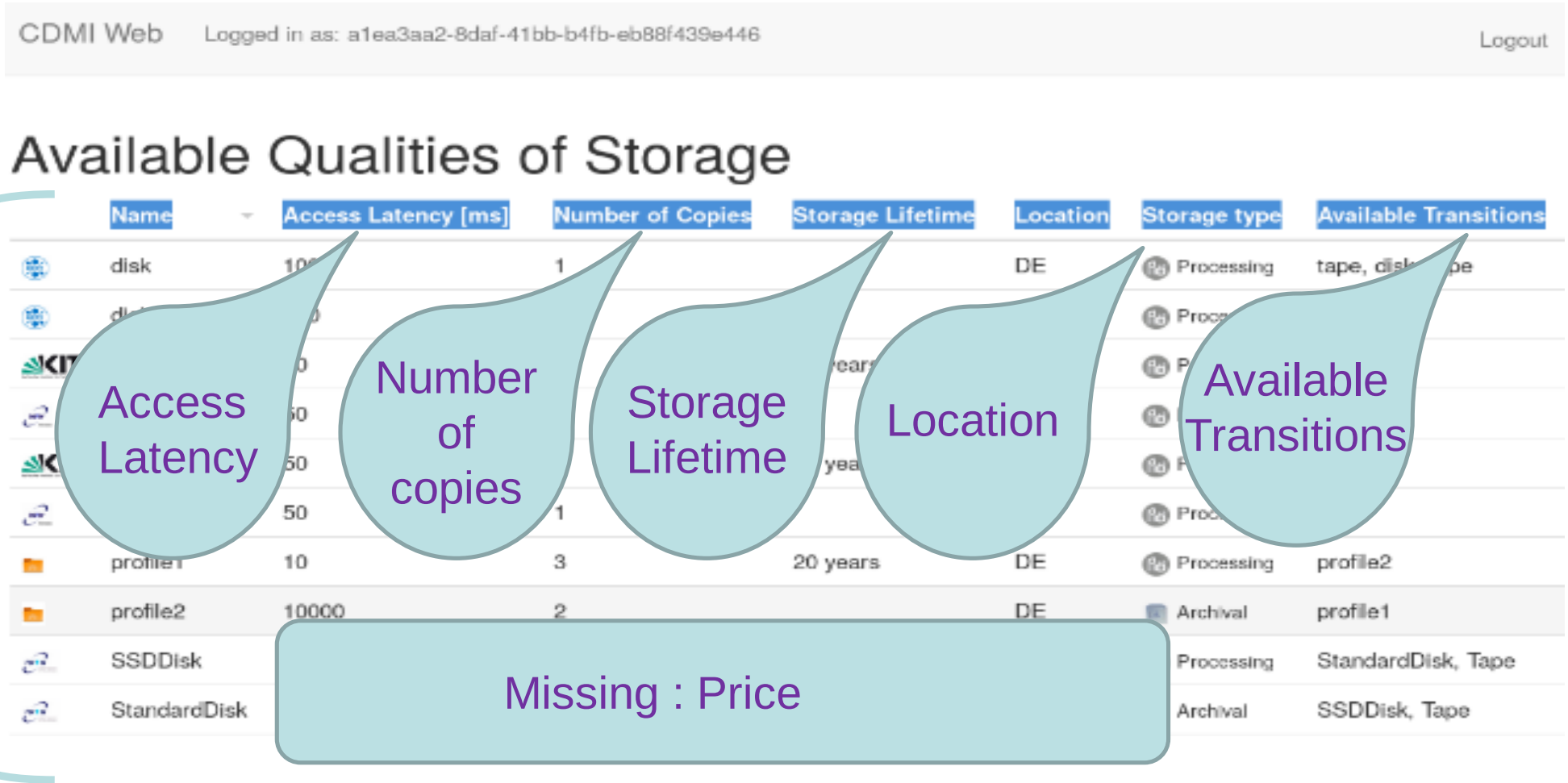
IBM
GPFS / HPFS

ceph

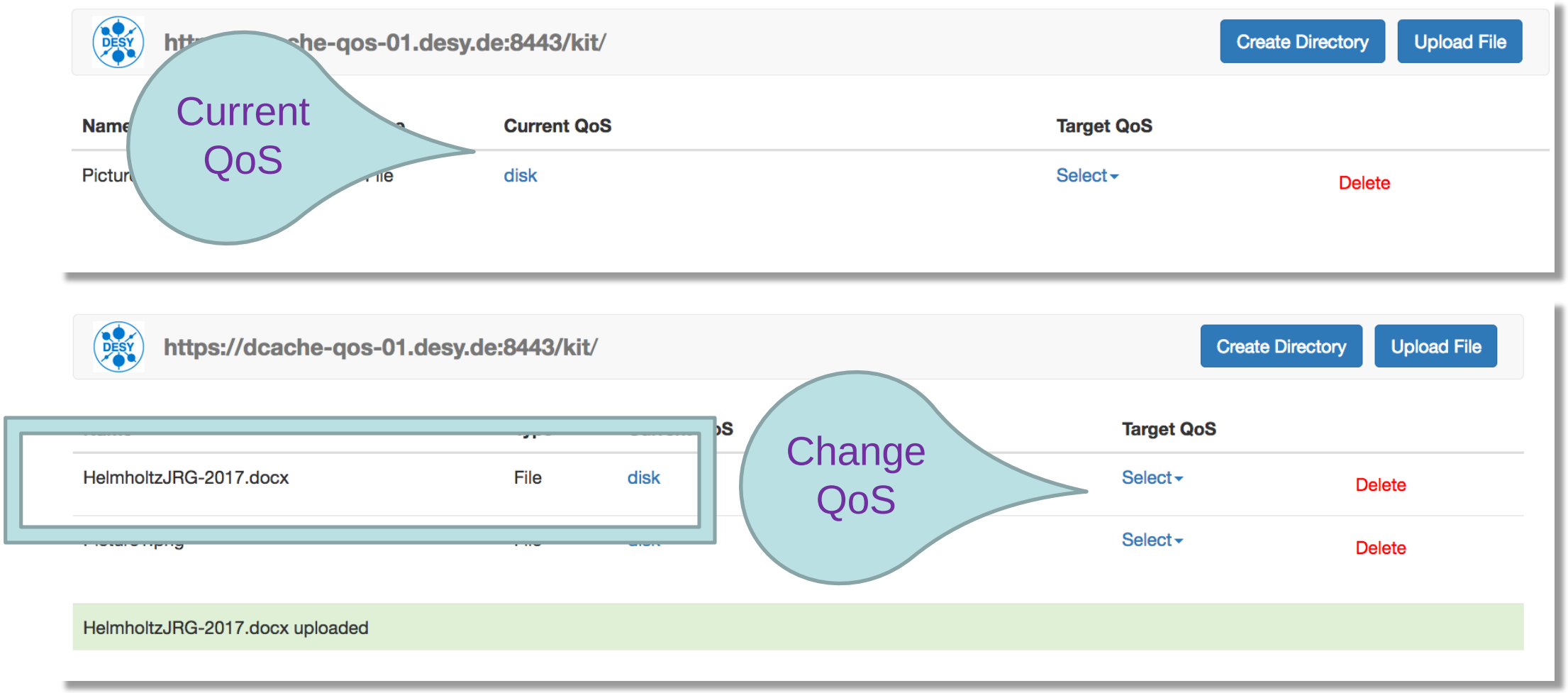
StoRM

dCache

Prototype Storage Broker via CDMI



Redirecting the client from the broker to the target infrastructure



Summary

- The eXtreme DataCloud project, as part of the European Open Science Cloud initiative, will provide software to orchestrate data analysis and data movements on the European level for large e-Infrastructures in the order of Exabytes.
- One of the main innovations is the consistent use of well defined QoS in storage, allowing the Cloud Platform layers to steer data access latency, retention policies, geographical locations and possible transitions from within the framework and w/o user or admin interactions.
- For now, XDC decided to use an extension of the CDMI protocol to negotiate QoS and QoS transitions between the PaaS layers and the storage providers.
  - The QoS vocabulary is defined with the Research Data Alliance
  - The CDMI protocol extensions with SNIA
- First results will be adopted by upcoming H2020 EU Projects and by infrastructures like WLCG and the European X-FEL.