The XDC project: exascale data management for WLCG and other disciplines

eXtreme DataCloud

Data Management for extreme scale computing

Daniele Cesini – INFN-CNAF

info<at>extreme-datacloud.eu

eXtreme DataCloud is co-funded by the Horizon2020 Framework Program – Grant Agreement 777367

XDC Objectives

- The 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

XDC Foundations

- **XDC take the move from**
  - the INDIGO Data management activity
  - the experience of the project partners on data-management

- **Improve already existing, production quality, Federated Data Management services**
  - By adding missing functionalities requested by research communities
  - Must be coherently harmonized in the European e-Infrastructures
  - **TRL 6+ → TRL8** (as requested by the H2020 call)

XDC Consortium

ID	Partner	Country	Represented Community	Tools and system
1	INFN (Lead)	IT	HEP/WLCG	INDIGO-Orchestrator, INDIGO-CDMI(*)
2	DESY	DE	Research with Photons (XFEL)	dCache
3	CERN	CH	HEP/WLCG	EOS, DYNAFED, FTS
4	AGH	PL		ONEDATA
5	ECRIN	[ERIC]	Medical data	
6	UC	ES	Lifewatch	
7	CNRS	FR	Astro [CTA and LSST]	
8	EGI.eu	NL	EGI communities	

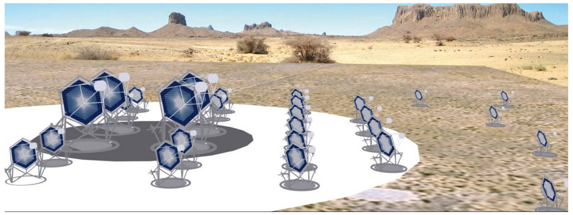
- 8 partners, 7 countries
- 7 research communities represented + EGI
- XDC Total Budget: 3.07Meuros
- XDC started on Nov 1st 2017 – will run for 27 months until Jan 31st 2020

Represented research communities

eXtreme DataCloud

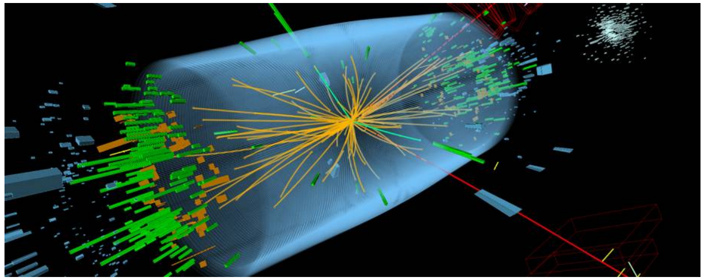
cta
cherenkov telescope array

LSST



European XFEL

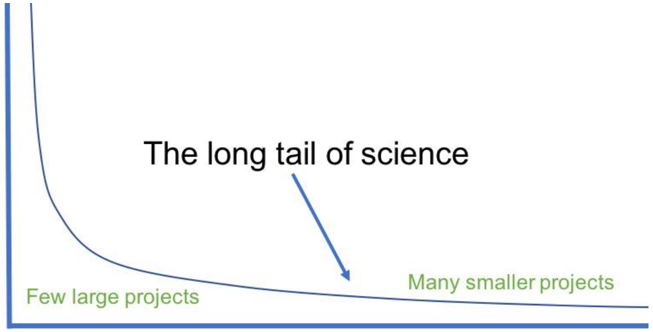
WLCG
Worldwide LHC Computing Grid



LifeWatch ERIC

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

ECRIN supports multinational clinical trials in Europe

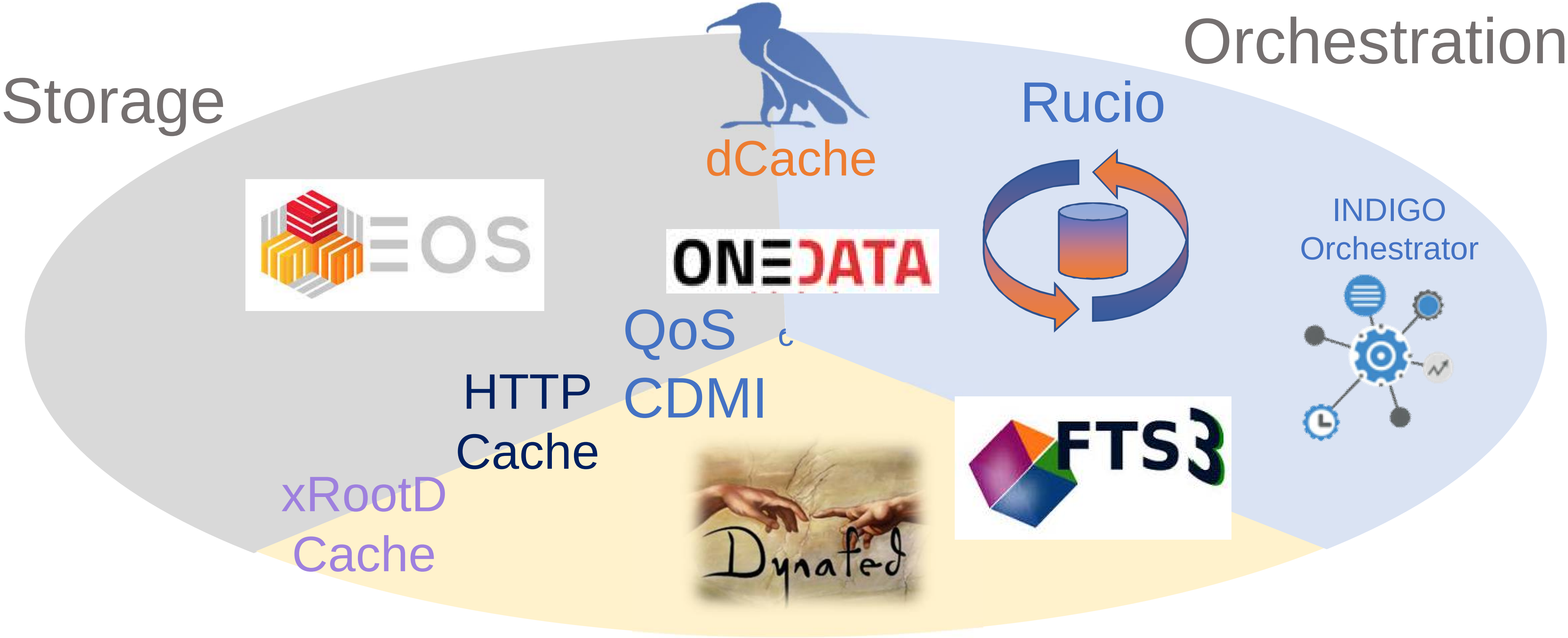


XDC Technical Topics

- Intelligent & Automated Dataset Distribution
  - Orchestration to realize a policy-driven data management
  - Data distribution policies based on Quality of Service (i.e. disks vs tape vs SSD) supporting geographical distributed resources (cross-sites)
  - Software lifecycle management
- Data management based on access patterns
  - Move to ‘glacier-like’ storage unused data, move to fast storage “hot” data
    - at infrastructure level
- Data pre-processing during ingestion
- Smart caching
  - Transparent access to remote data without the need of a-priori copy
- Metadata management
- Sensitive data handling
  - secure storage and encryption

The Toolbox

Production Level Components



Storage

Orchestration

Federation

dCache

Rucio

INDIGO Orchestrator

QoS

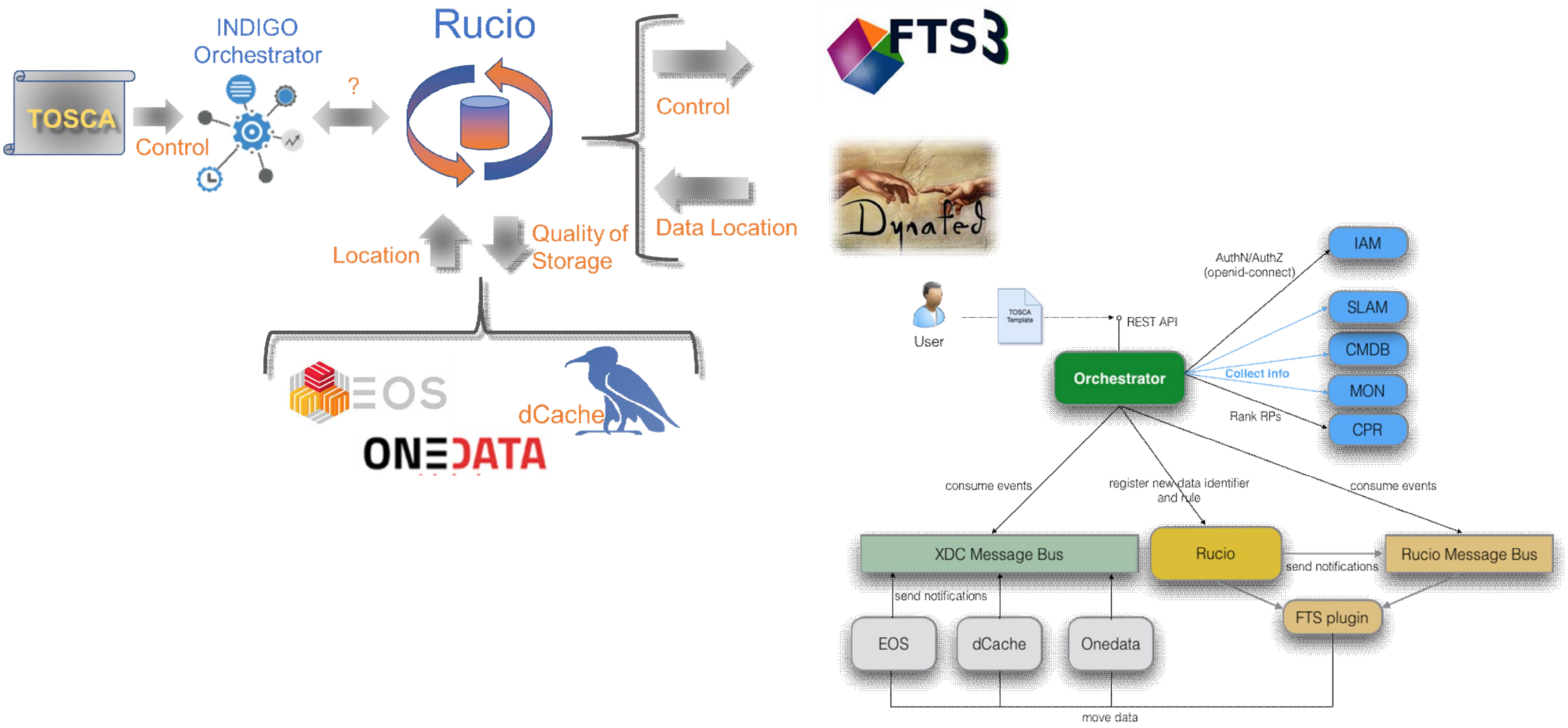
CDMI

HTTP Cache

xRootD Cache

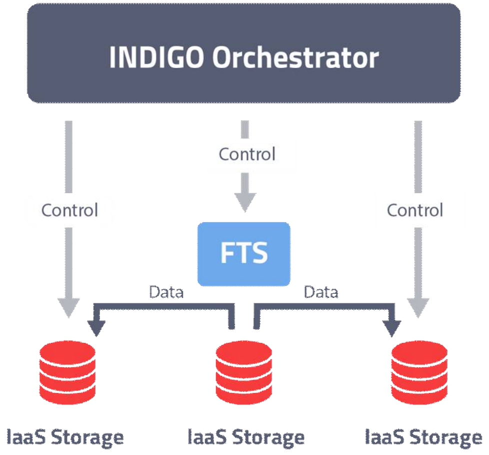
The Orchestration

Orchestration Control Flow



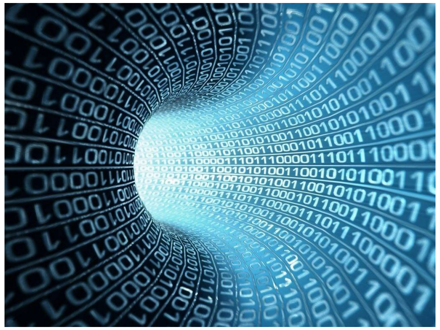
Policy driven Data Management

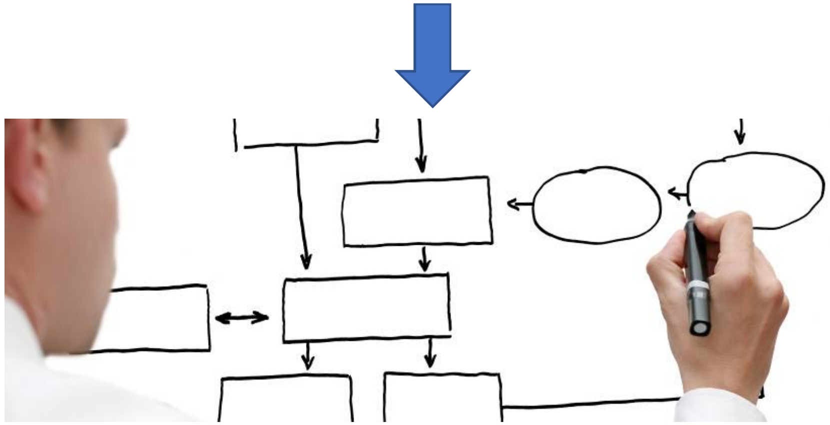
- Intelligent & Automated Dataset Distribution
  - A typical workflow
    - Initially the data will be stored on low latency devices for fast access
    - To ensure data safety, the data will be replicated to a second storage device and will be migrated to custodial systems, which might be tape or S3 appliances
    - Eligible users will get permission to restore archived data if necessary
    - After a grace period, Access Control will be changed from “private” to “open access”
  - Data management based on access pattern

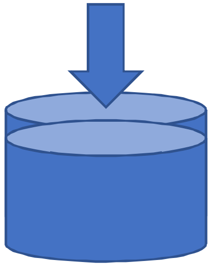


Data pre-processing

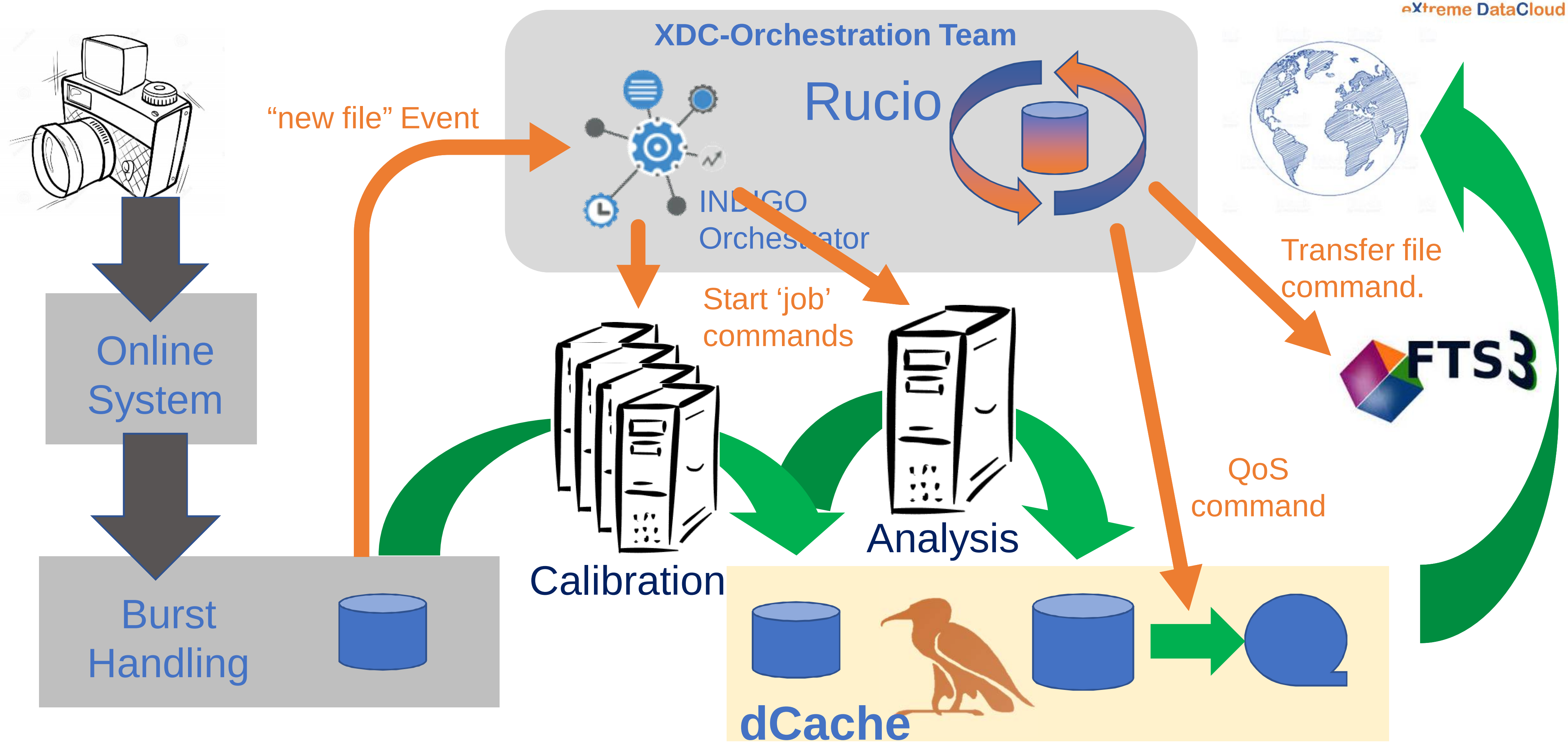
- Data pre-processing during ingestion
  - Automatically run user defined applications and workflows when data are uploaded
    - i.e. for Skimming, indexing, metadata extraction, consistency checks
  - Implement a solution to discover new data at specific locations
  - Create the functions to request the INDIGO PaaS Orchestrator to execute specific applications on the computing resources on the Infrastructure
  - Implement a high-level workflow engine, that will execute applications defined by the users
  - Implement the data mover to store the elaborated data in the final destination







The simple X-FEL Use Case

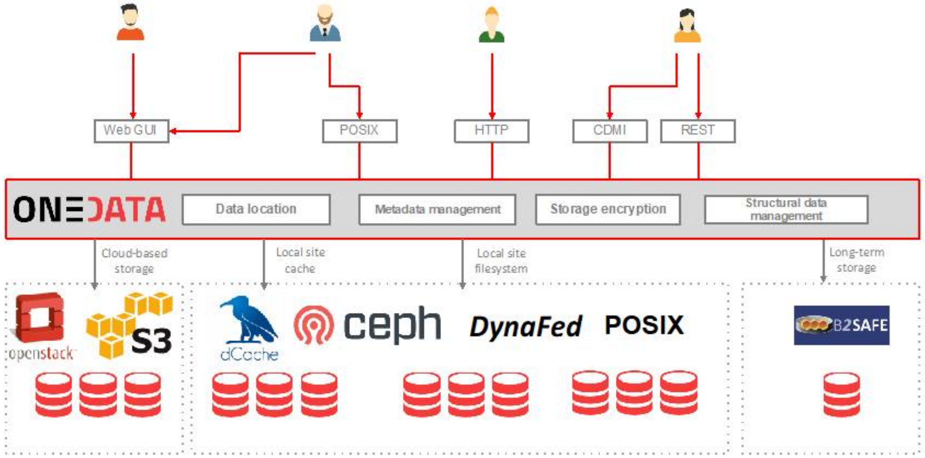


Orchestration
Metadata Management
Secure Storage

Onedata developments

- Unified data access platform at a PaaS level at the Exascale
- Advanced metadata management with no pre-defined schema
- Encryption Services and Secure Storage
- Sensitive data management and key storage within Onedata

https://onedata.org



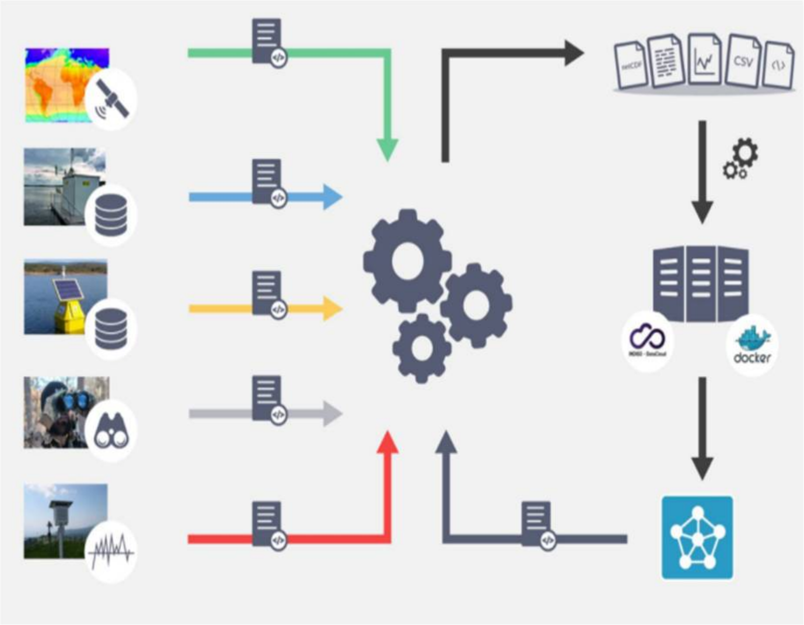
ONEDATA

GLOBAL DATA ACCESS SOLUTION FOR SCIENCE

Have the best of both worlds! Perform heavy computations on huge datasets. Access your data in a dropbox-like fashion regardless of its location. Publish and share your results with public or closed

LifeWatch Use Case

- **Problem**: Life Cycle Management of data related to **Water Quality** involving heterogeneous data sources
  - Satellite, Real-time monitoring, meteorological stations.
- **Goal**: Integrate data sources and different types of modelling tools to simulate freshwater masses in a FAIR data environment
  - Use of standards like EML (Ecological Metadata Language)
- **XDC Solution**:
  - Onedata
    - Metadata management and discovery, Digital Identifier minting, storage
  - PaaS Orchestrator
    - automatic preprocessing for data harmonization and model deployment



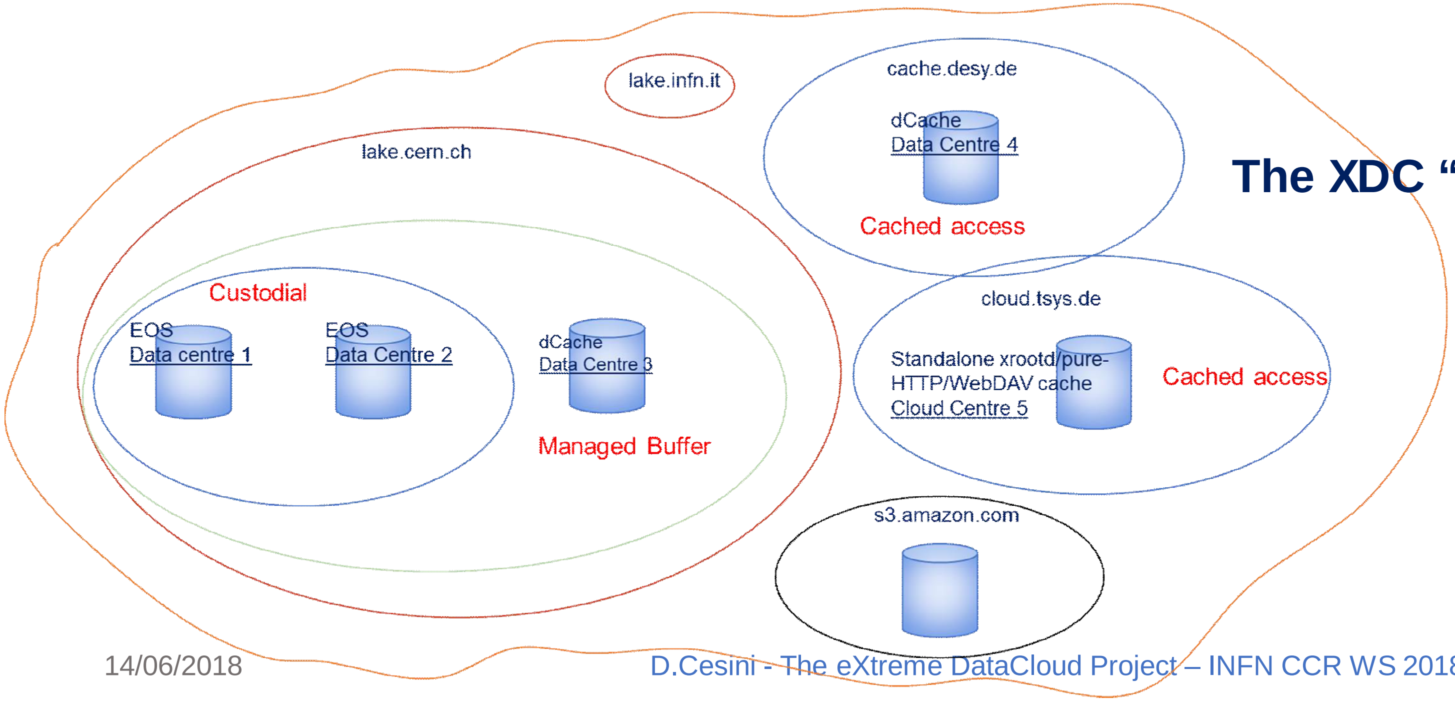
ECRIN Use Case

- **Problem**: Distributed files and data objects across different repositories. Metadata heterogeneity. Sensitive Data
- **Goal**: Single environment to make clinical trial data objects available for sharing with others. Sources are spread over
  - a variety of access mechanisms
  - several different locations
    - growing number of general and specialised data repositories
    - trial registries
    - Publications
    - the original researchers’ institutions
- **XDC Solution**: Onedata
  - Metadata management and discovery
  - Secure Storage

The Caching Part

Smart caching

- Smart caching
  - Develop a global caching infrastructure supporting the following building blocks:
    - dynamic integration of satellite sites by existing data centres
    - creation of standalone caches modelled on existing web solutions
    - federation of the above to create a large scale caching infrastructure



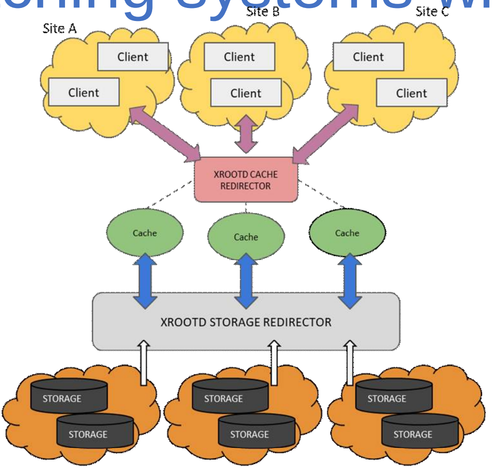
The XDC "Data Lake"

Project Status

- Started on Nov 1st 2017 – Kickoff meeting Jan 2018
- Detailed requirements collection from user communities completed
- Definition of the detailed architecture almost completed
- Creation of the Pilot Testbed started
  - Currently reserved for internal communities
  - Under discussion the possibility to open to external users
- Started the developments for the Orchestrator-Rucio integration
- Caching systems with XCache and HTTP

On the Testbed….

- Onedata release candidate 18.02.0-rc6
  - Improved stability and scalability
- dCache, EOS, RUCIO Orchestrator endpoints
  - + ancillary systems
- Caching systems with XROOTD and HTTP



Credits: XROOTD:D. Ciangottini, D.Spiga, T.Boccali – CMS and XDC
HTTP : A. Falabella

The Plan for the Next Months

- Architecture finalized - End of May 2018
- Pilot test bed in place – End of May 2018
- Event with User Communities – Jun 18-22 2018, Santander – joint with DEEP
- All Hands meeting @ DESY - Sept 2018
- XDC reference releases – 1 - Oct-Nov 2018
- ……
- XDC reference releases – 2 - Oct-Nov 2019
- Functionalities and scalability demonstrated - Jan 2020

XDC Contacts

- Website: www.extreme-datacloud.eu
- **@XtremeDataCloud on Twitter**
- **Mailing list: info<at>extreme-datacloud.eu**