Data Management for extreme scale computing

The XDC project

Fernando Aguilar (IFCA UC-CSIC)

fernando.aguilar<at>extreme-datacloud.eu

Alessandro Costantini (INFN)

alessandro.costantini<at>extreme-datacloud.eu

Daniele Cesini (INFN - Project Coordinator)

daniele.cesini<at>extreme-datacloud.eu

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

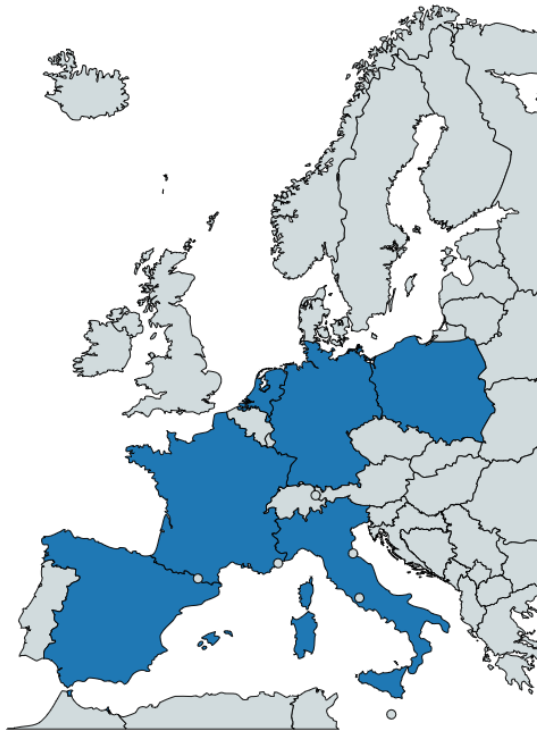
XDC

- 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
- Addresses the EINFRA-21-2017 (b)-2: “Computing e-infrastructure with extreme large datasets”
  - Deal with heterogeneous datasets
  - Bring to TRL8 and include in a unified service catalogue services and prototype at least at TRL6

XDC Consortium

ID	Partner	Country	Represented Community	Tools and system
1	INFN (Lead)	IT	HEP/WLCG	INDIGO-Orchestrator
2	DESY	DE	Research with Photons (XFEL)	dCache
3	CERN	CH	HEP/WLCG	EOS, DYNAFED, FTS, RUCIO
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
- 6 research communities represented + EGI
- XDC Total Budget: 3.07Meuros

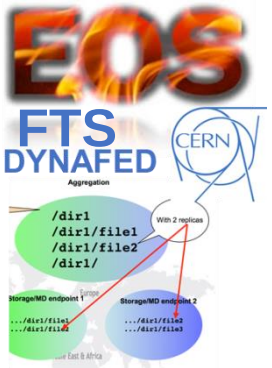


The Approach

- **Improve already existing, production quality Data Management services**
  - By adding missing functionalities requested by research communities
  - Based mainly on technologies provided by the partners and by the INDIGO-Datacloud project
  - Must be coherently harmonized in the European e-Infrastructures

INDIGO PaaS Orchestrator

INDIGO CDMI Server



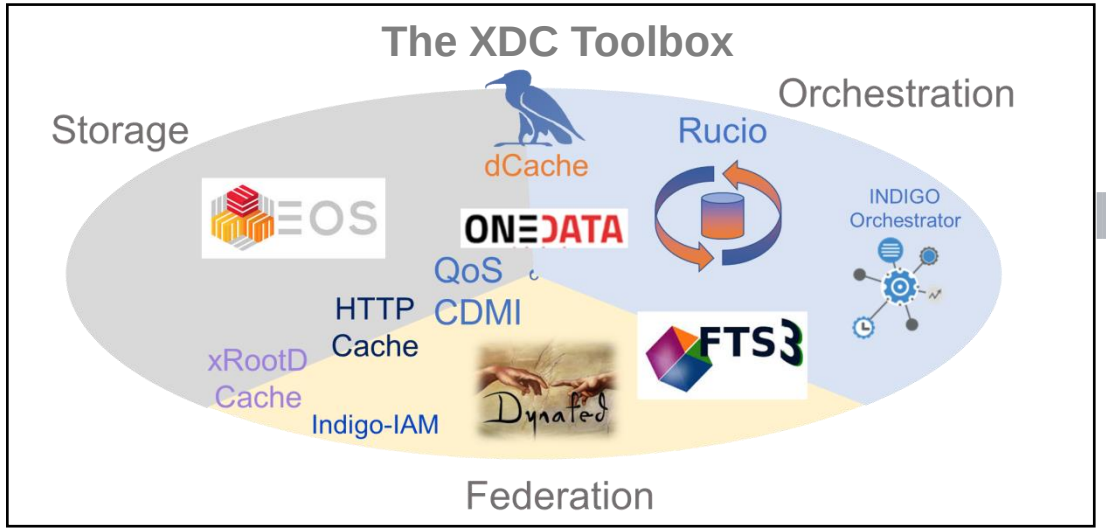
The New Functionalities

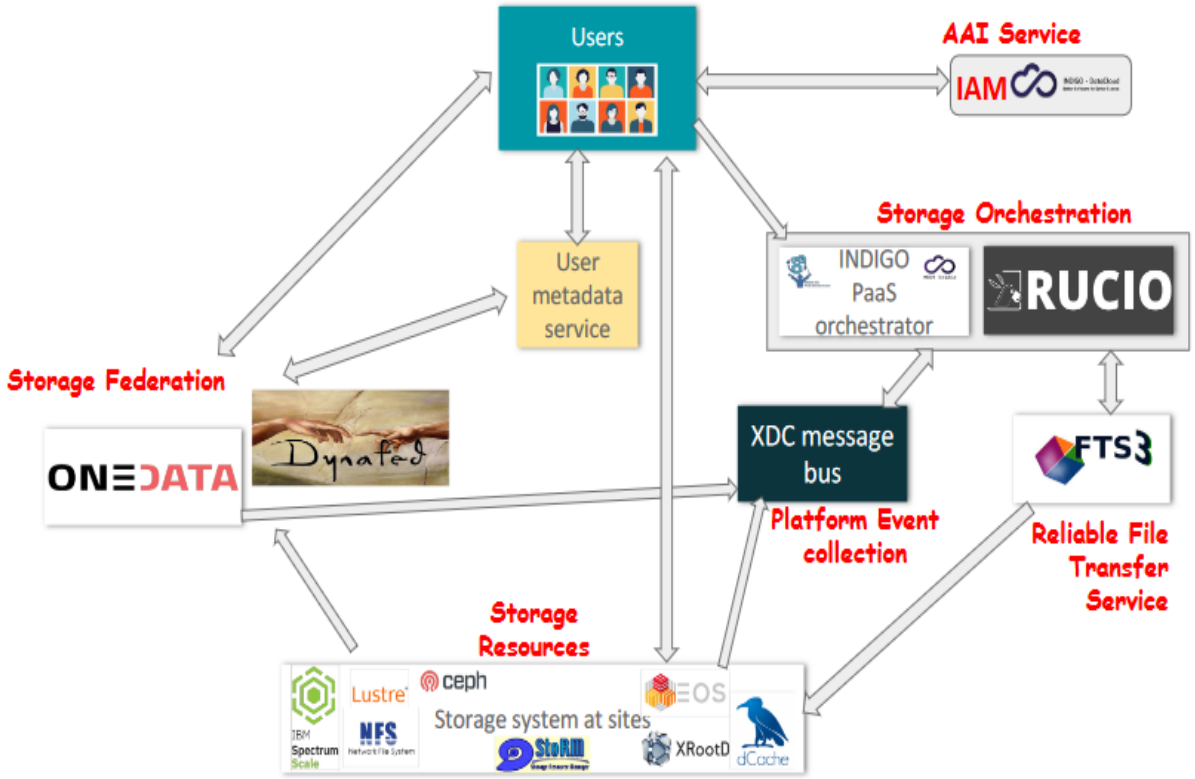
- Intelligent & Automated Dataset Distribution
  - Orchestration to realize a policy-driven data management
  - Data distribution policies based on Quality of Service (i.e. disks vs tape vs SSD) supporting geographical distributed resources (cross-sites)
  - Data lifecycle management
- Data pre-processing during ingestion
- Metadata management
- Data management based on storage events
- Smart caching
  - Transparent access to remote data without the need of a-priori copy
    - To support dynamic inclusion of diskless sites
    - To improve efficiency in multi-site storage systems and storage federations (i.e. Datalakes)
- Sensitive data handling
  - secure storage and encryption

General Architecture Definition

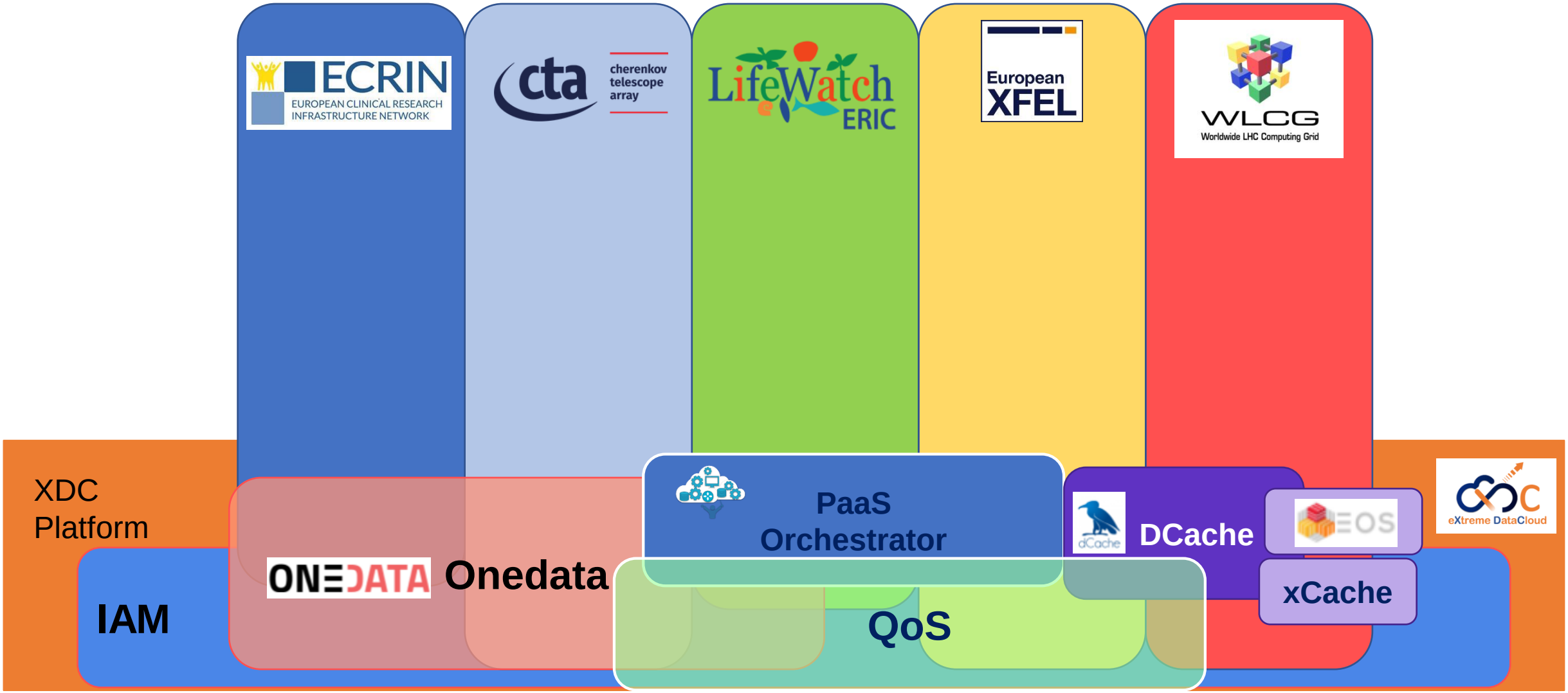
- XDC acts at all the e-infrastructure levels
  - Storage systems at sites
  - Federations of storage systems
    - regional and global
  - High level orchestration
  - User experience

- The “toolbox” was mapped in those levels to define the general architecture
  - Taking into account the user requirements





Use Cases – XDC Services



LifeWatch Use Case

- **Problem**: Data 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.
- **XDC Main Services**: Onedata (Metadata management and discovery, Digital Identifier minting, storage), PaaS Orchestrator (automatic preprocessing for data harmonization and model deployment).
- **Status**: Data Sources integrated, analysis with Orchestrator.

CTA Use Case

- **Problem**: Cherenkov Telescope Array gathering data from the Cosmos (properties of the gamma ray). Complex and Big Data management in a distributed environment. Data quality Assurance.
- **Goal**: FAIR Data Management in the Cloud. Integration of tools to archive event/monitoring/calibration. Proprietary period management and user access control.
- **XDC Main Services**: Onedata (Metadata management and discovery, storage, intelligent dataset distribution), QoS (policy definition), PaaS Orchestrator (pre-processing).
- **Status**: Metadata extraction and storing/attachment in Onedata.

ECRIN Use Case

- **Problem**: Distributed information about clinical studies and data objects related to these studies across different registries and repositories. Metadata heterogeneity.
- **Goal**: Single environment to find data objects across repositories and registries, based on metadata.
- **XDC Main Services**: Onedata (Metadata management and discovery, Harvesting).
- **Status**: Different sources integration. Metadata harmonization.

WLCG Use Case

- **Problem**: Growing needs on storage space
  - up to 900 PB in 2027
  - Data ready to be used/exploited in a very distributed environment
- **Goal**: Reduce costs, resource aggregation, smart data allocation
  - Smart Caching technologies,Resource Federation
  - Multi-site storage - the “DataLake”.
  - Dynamic extension of sites to remote locations
  - Data management and QoS
- **XDC Main Services**: QoS (policies definition), Xcache (smart caching), EOS (data caching).
- **Status**: QoS, http caching.

XFEL: Use Case Description

- **Problem**: Complex Data management in a distributed and heterogeneous environment.
  - Quality assessment and refinement of the produced data
  - Data production and storage
  - Data analysis
- **Goal**: Data lifecycle management. Processing and analytics.
  - The system should be able to manage data of different qualities
  - Scientist Perspective : Precious (Raw), Scratch , for HPC
  - Technical Perspective: spinning disk, tape, cloud, SSD
  - Trigger “new file” events to orchestrate storage and processing. Orchestration, QoS
- **XDC Main Services**: QoS (policies), PaaS Orchestrator (Preprocessing), dCache (data replication, federation). Message BUS (events triggering).
- **Status**: Orchestration configured based on events.

First XDC Release

Involved tools

- CachingOnDemand
- dCache
- Dynafed
- EOS
- FTS,GFAL
- Onedata
- PaaS Orchestrator plugin
- TOSCA types & templates plugin

Key technical highlights

- OpenIDConnect support for token based authentication
- New QoS types integration and support in dCache, FTS, GFAL
- Orchestrator integration with other components
- Performance improvements in Onedata
- Support for groups and roles in Onedata
- EOS-dCache integration
- Caching systems instantiation
- Storage events notification in dCache
- EOS caching with XCache for geographic deployment
- EOS external storage adoption

XDC-1/Pulsar

https://releases.extreme-datacloud.eu/en/latest/releases/pulsar/index.html

Exploitation and sustainability path

- The Service Providers Board (SPB) goal is to link XDC with Service Provider within and outside of the project consortium in order to have a regular dialog among XDC and all the Service Providers
  - First meeting @EOSC-hub week 2019
- Making XDC products available through EGI distribution channels
  - UMD release already delivers dCache, FTS, GFAL, XRootd
- Identification and interaction with the EOSC-HUB Service Providers
  - To be included in the XDC SPB (if not already present)
- Pushing developments in the upstream repositories of all the services to ensure sustainability beyond the project
  - Double path
    - XDC repositories
    - Upstream repository
- XDC solutions included in other projects or actions.

Target on Communities.

- Dealing with BigData infrastructures
- Dealing with the DataManagement of extreme scale experiments