Co-funded by the Horizon 2020
Framework Programme of the European Union
Grant Agreement Number 825532

Large-scale EXecution for Industry & Society

LEXIS

www.lexis-project.eu

A TRANSNATIONAL DATA SYSTEM FOR HPC/CLOUD COMPUTING WORKFLOWS BASED ON IRODS/EUDAT

iRODS UGM

9 June, 2021

MARTIN GOLASOWSKI (IT4I, CZ)
MOHAMAD HAYEK (LRZ, DE)
RUBÉN J. GARCÍA-HERNÁNDEZ (LRZ, DE)

LEXIS Project Consortium

Large-scale EXecution for Industry & Society

{ At the confluence of HPC, Cloud Computing & Big Data }

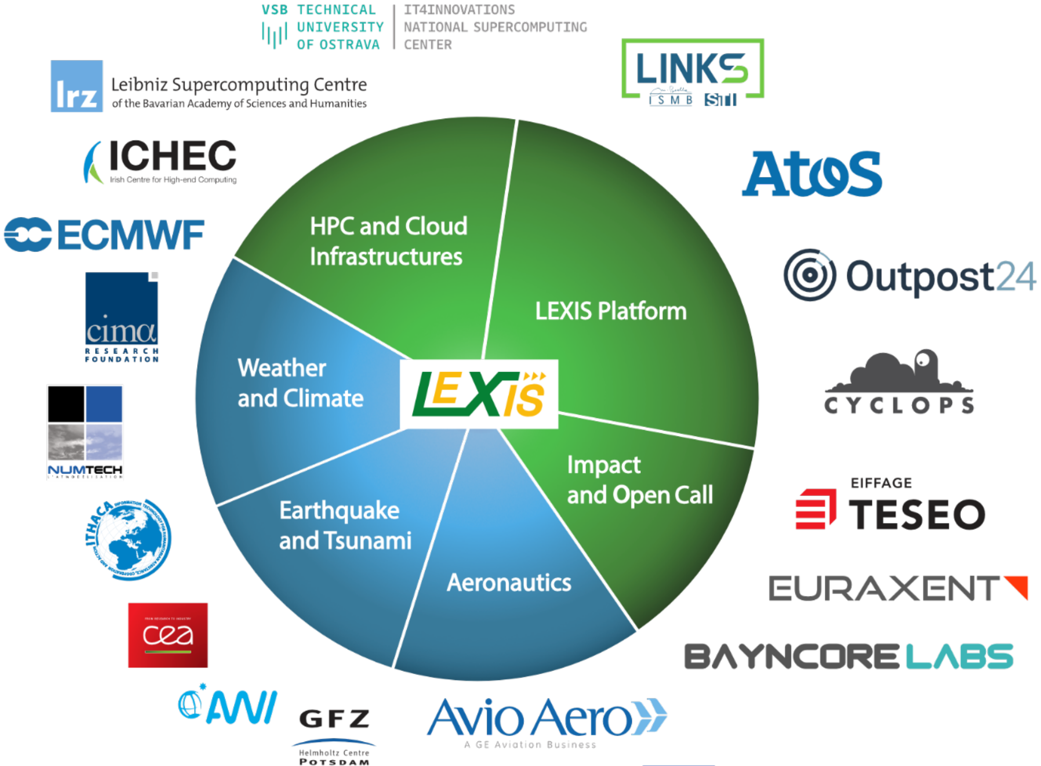
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825532

START DATE
January 1st, 2019

DURATION
36 Month

Funding / Overall Funding
€12.2 M / €14 M

- HPC & Cloud resource providers
- Scientific institutions
- Industrial companies & SMEs
- Information Technology providers



HPC and Cloud Infrastructures

LEXIS Platform

Impact and Open Call

Aeronautics

Earthquake and Tsunami

Weather and Climate

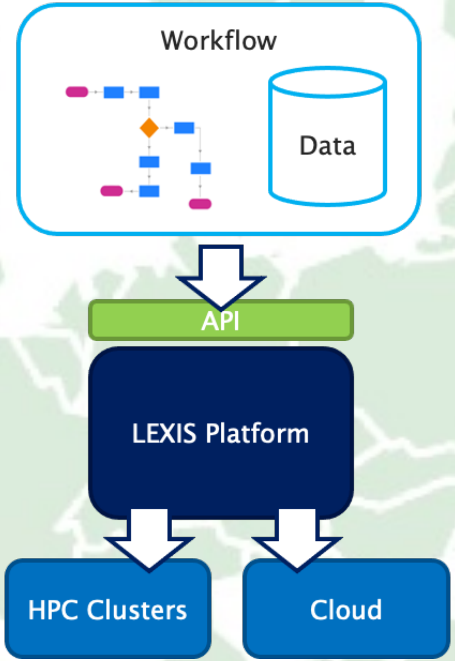
LEXIS

VSB TECHNICAL UNIVERSITY OF OSTRAVA
IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER
lrz Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities
LINKS ISMB SiTI
ICHEC Irish Centre for High-end Computing
Atos
ECMWF
Outpost24
cima RESEARCH FOUNDATION
CYCLOPS
NUMTECH
EIFFAGE TESEO
ITHACA
EURAXENT
cea
BAYNCORE LABS
AWI
GFZ Helmholtz Centre POTSDAM
Avio Aero A GE Aviation Business

This infrastructure is part of a project that has received funding from the European Union's Horizon 2020 research and innovationprogramme under grant agreement No 825532.

LEXIS project challenges

- **Dynamic, data-aware and complex workflows orchestration**
  - Execute complex workloads on Cloud and HPC
  - Easy access to state-of-the art compute resources
  - *REST-based APIs*
  - Federation of supercomputing centers
  - Real-time deadline-aware workflows over both Cloud and HPC
- **Cross-site data and metadata management solution**
  - Move data between various resources using single API
  - Distributed solution based on iRODS
  - Distributed data staging between resources
- **Data sharing between Cloud and HPC resources**
  - Accelerated by dedicated Burst Buffer nodes, high bandwidth network and FPGA cards for on-line processing (I/O acceleration)



LEXIS – What it is?

Advanced, distributed platform for HPC/Cloud/Big Data workflows, with Orchestration/Data solutions

Distributed Big Data & Simulations

Platform

Unified AAI

atform

Data Curation, MGMT, Sharing, Valorisation (EUDAT/iRODS)

Platform

Easy uptake by SMEs, Science, Industry „HPC…aaS“

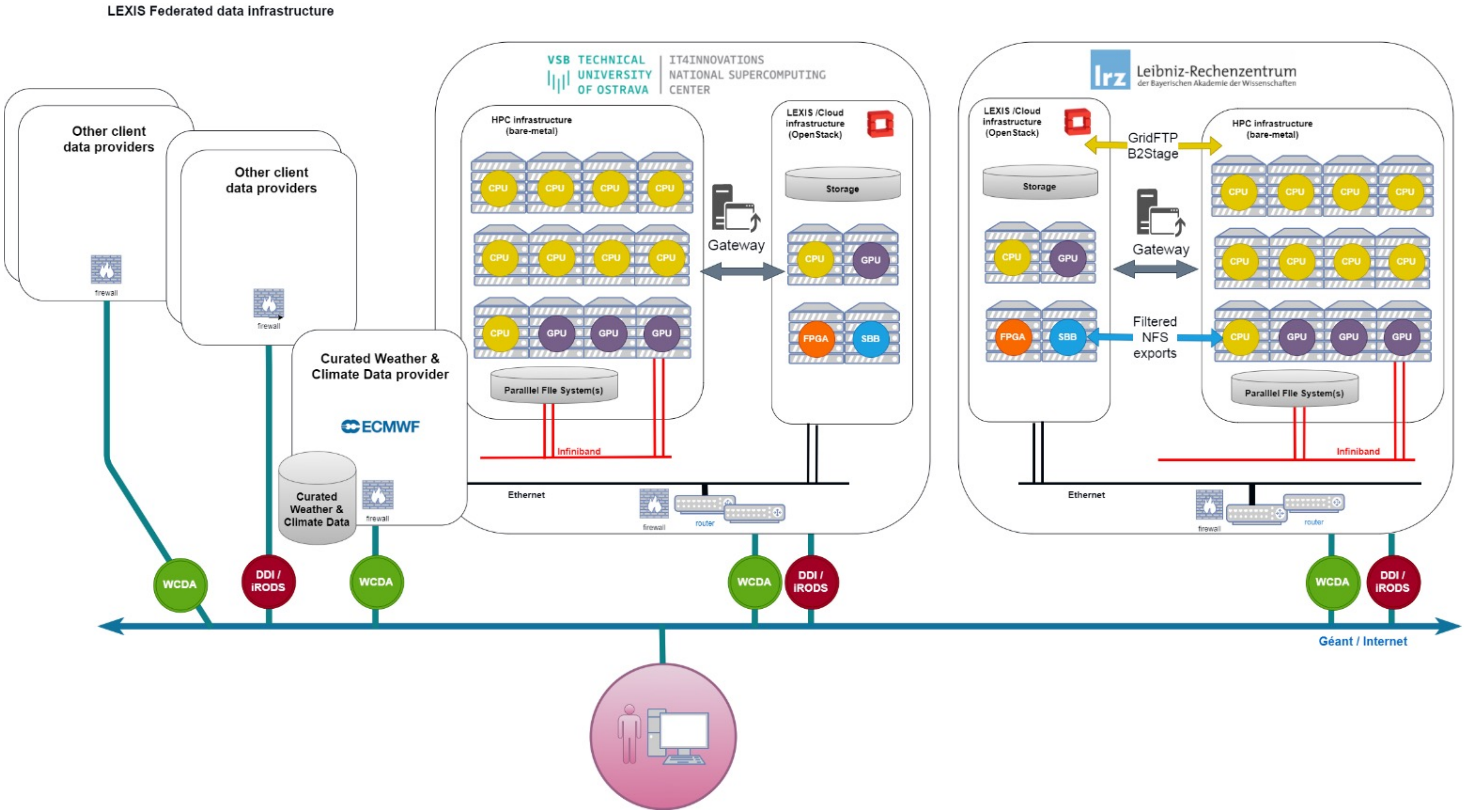
form

Fine-granular accounting/billing (CYCLOPS)

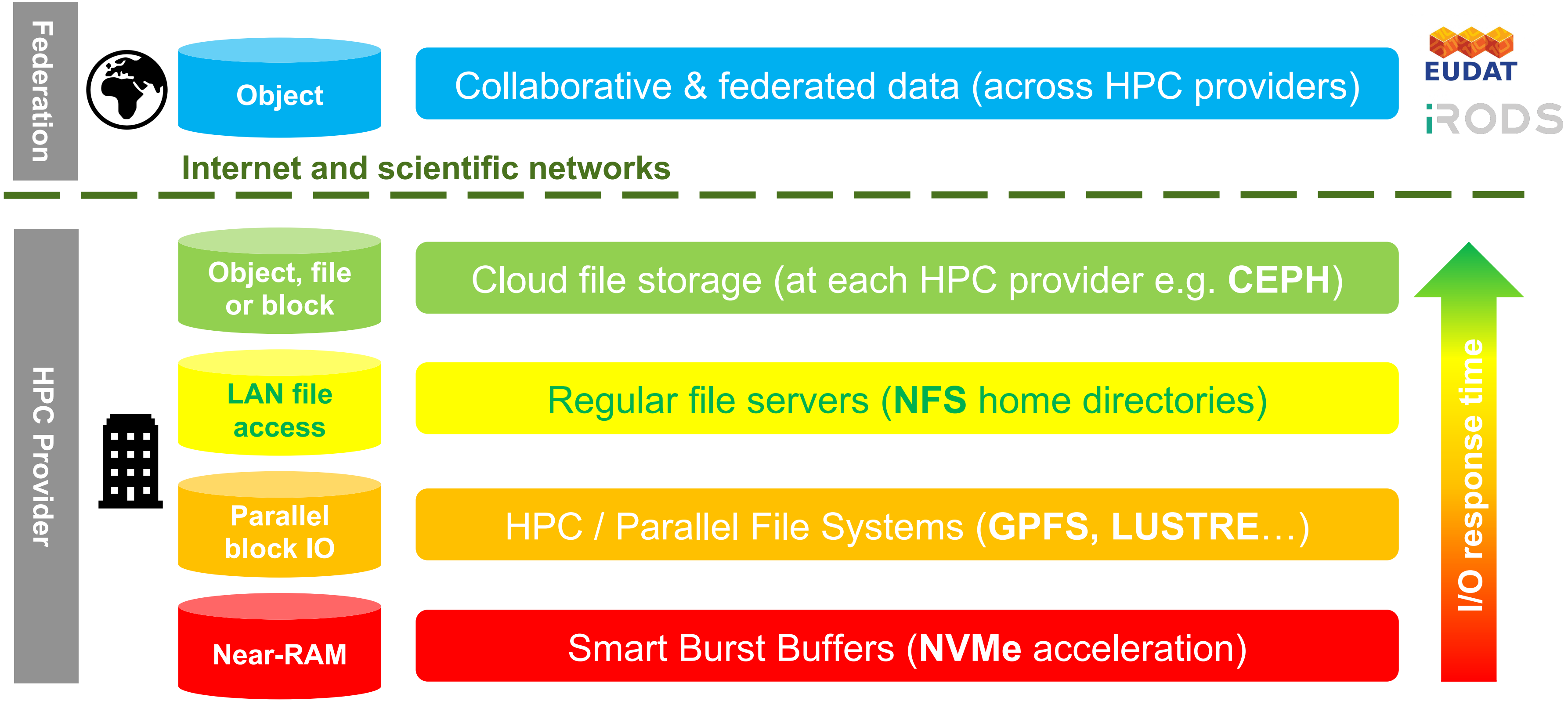
Cloud/HPC based workflows (YORC/HEAppE)

LEXIS Distributed Data Infrastructure (DDI)

High level view of the LEXIS HPC, Cloud & Big Data federation


LEXIS Federated data infrastructure
Other client data providers
firewall
Other client data providers
firewall
Curated Weather & Climate Data provider
ECMWF
Curated Weather & Climate Data
firewall
VSB TECHNICAL UNIVERSITY OF OSTRAVA
IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER
HPC infrastructure (bare-metal)
CPU
GPU
Parallel File System(s)
Infiniband
Gateway
LEXIS /Cloud infrastructure (OpenStack)
Storage
CPU
GPU
FPGA
SBB
Ethernet
firewall
router
lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften
LEXIS /Cloud infrastructure (OpenStack)
Storage
GridFTP B2Stage
Gateway
Filtered NFS exports
HPC infrastructure (bare-metal)
Parallel File System(s)
Infiniband
Ethernet
firewall
router
WCDA
DDI / iRODS
Géant / Internet

LEXIS DDI - Storage tiers



Federation

Object

Collaborative & federated data (across HPC providers)

EUDAT

iRODS

Internet and scientific networks

HPC Provider

Object, file or block

Cloud file storage (at each HPC provider e.g. **CEPH**)

LAN file access

Regular file servers (**NFS** home directories)

Parallel block IO

HPC / Parallel File Systems (**GPFS, LUSTRE…**)

Near-RAM

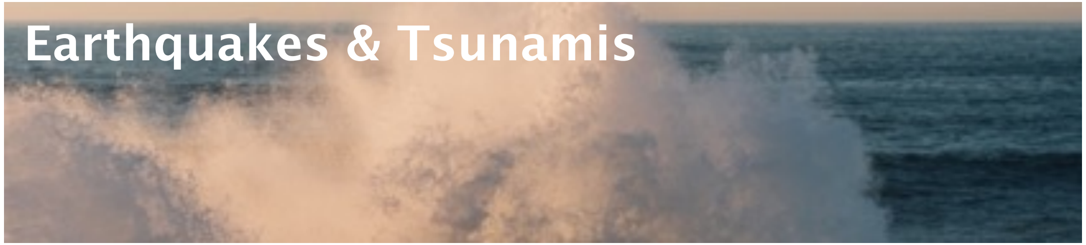
Smart Burst Buffers (**NVMe** acceleration)

I/O response time

LEXIS PILOT Use-Cases

Aeronautics

Computation Fluid Dynamics (CFD),
Rotating parts (gearboxes),
3D Visualization



Earthquakes & Tsunamis

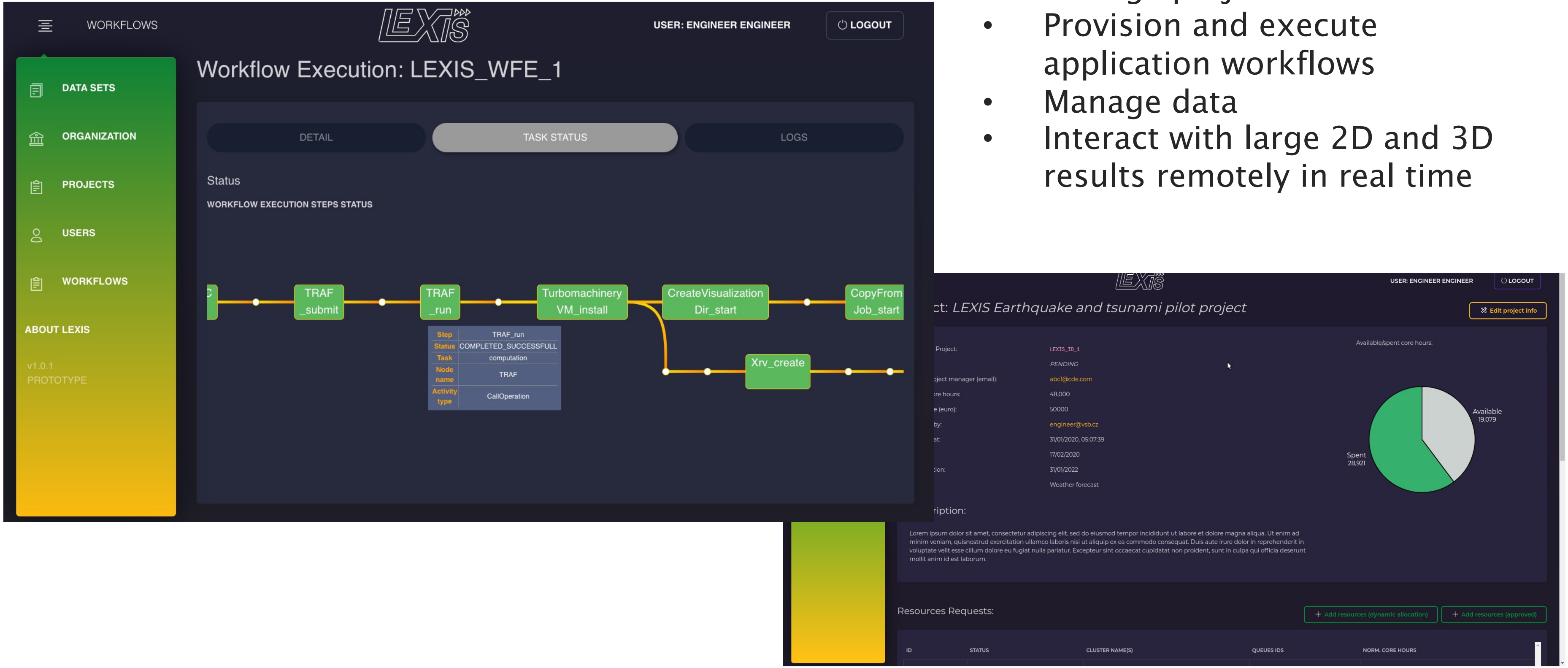
Earthquakes & Tsunami prediction models, geographic and urban databases, emergency organization, urgent computing

Weather & Climate

Weather & Climate models (WRF) and various post-processors for flood, wildfire & agriculture applications

User experience

LEXIS portal & 3D remote visualization



Workflow Execution: LEXIS_WFE_1

Status

WORKFLOW EXECUTION STEPS STATUS

Step	TRAF_run
Status	COMPLETED_SUCCESSFULL
Task	computation
Node name	TRAF
Activity type	CallOperation

ct: *LEXIS Earthquake and tsunami pilot project*

Project:	LEXIS_ID_1
	PENDING
oject manager (email):	abc1@cde.com
re hours:	48,000
e (euro):	50000
by:	engineer@vsb.cz
at:	31/01/2020, 05:07:39
	17/02/2020
ion:	31/01/2022
	Weather forecast

ription:

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quisnostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Resources Requests:

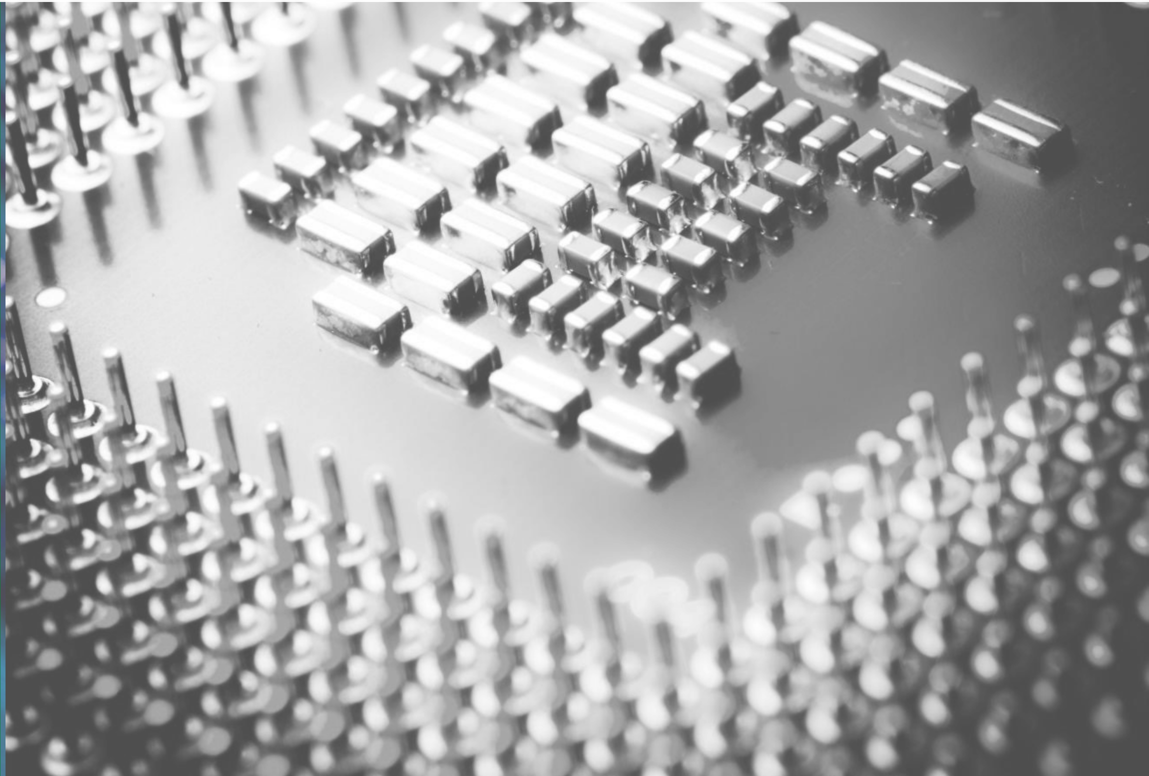
ID	STATUS	CLUSTER NAME(S)	QUEUES IDS	NORM. CORE HOURS

ALL-IN-ONE WEB INTERFACE

- Manage client organization
- Manage projects
- Provision and execute application workflows
- Manage data
- Interact with large 2D and 3D results remotely in real time

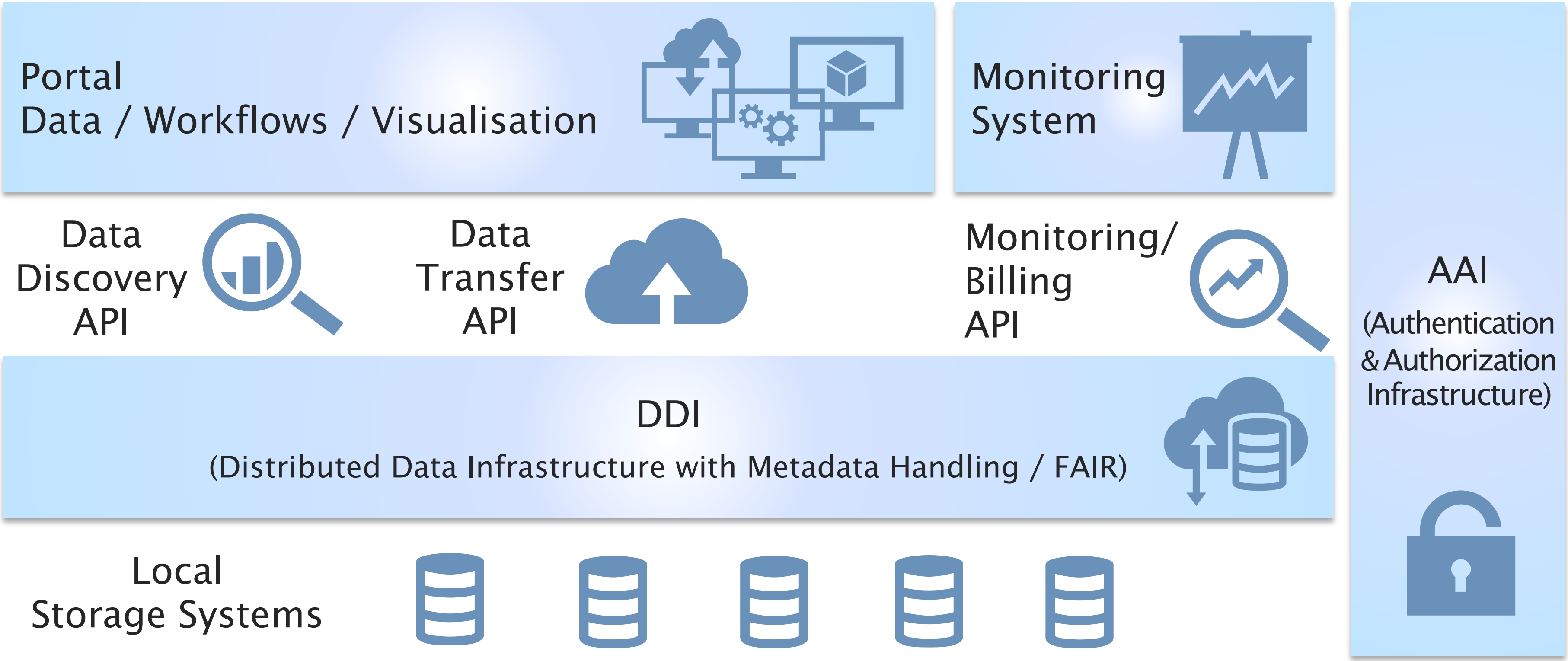
LEXIS DISTRIBUTED DATA INFRASTRUCTURE (DDI)

WORK PACKAGE 3



LEXIS DDI Integration

Distributed Data Infrastructure for the User – leveraging EUDAT components



Portal
Data / Workflows / Visualisation

Monitoring
System

Data
Discovery
API

Data
Transfer
API

Monitoring/
Billing
API

AAI

(Authentication
& Authorization
Infrastructure)

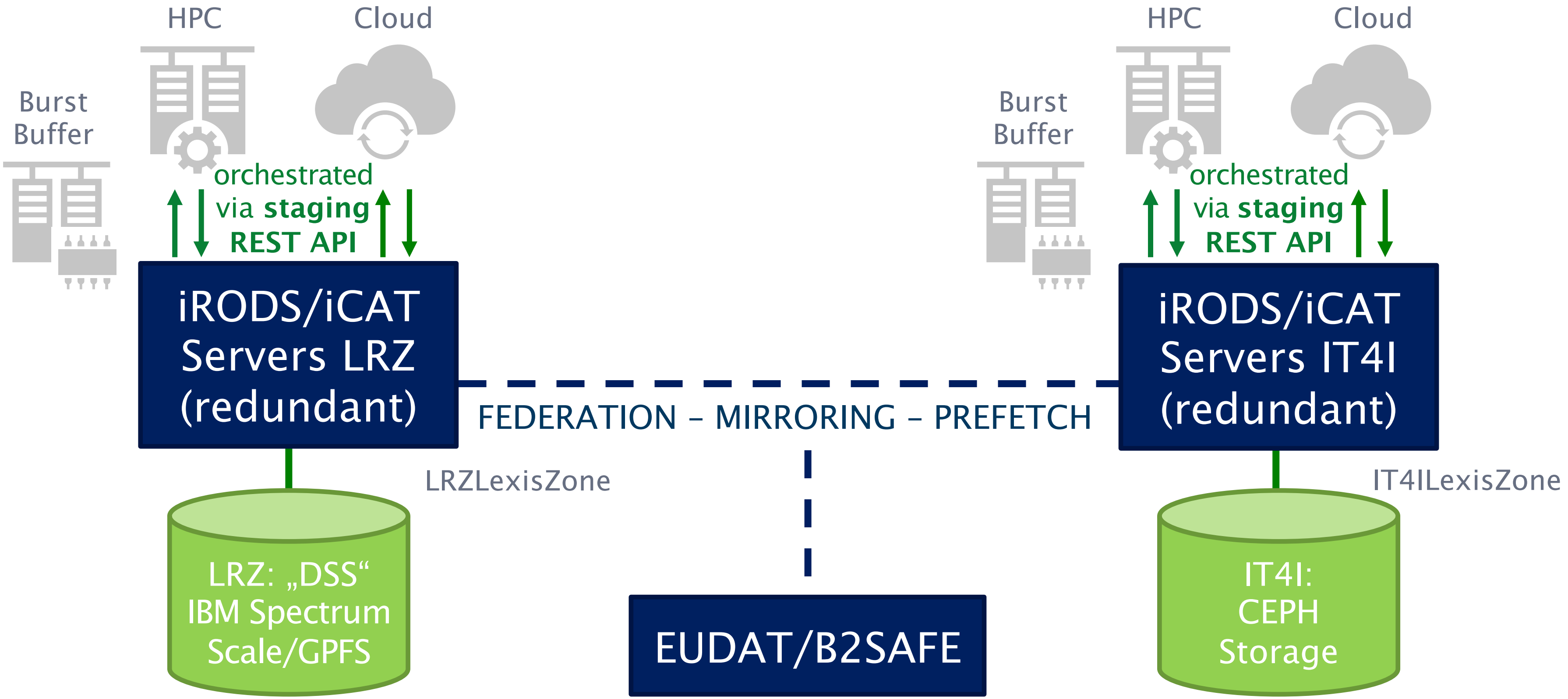
DDI

(Distributed Data Infrastructure with Metadata Handling / FAIR)

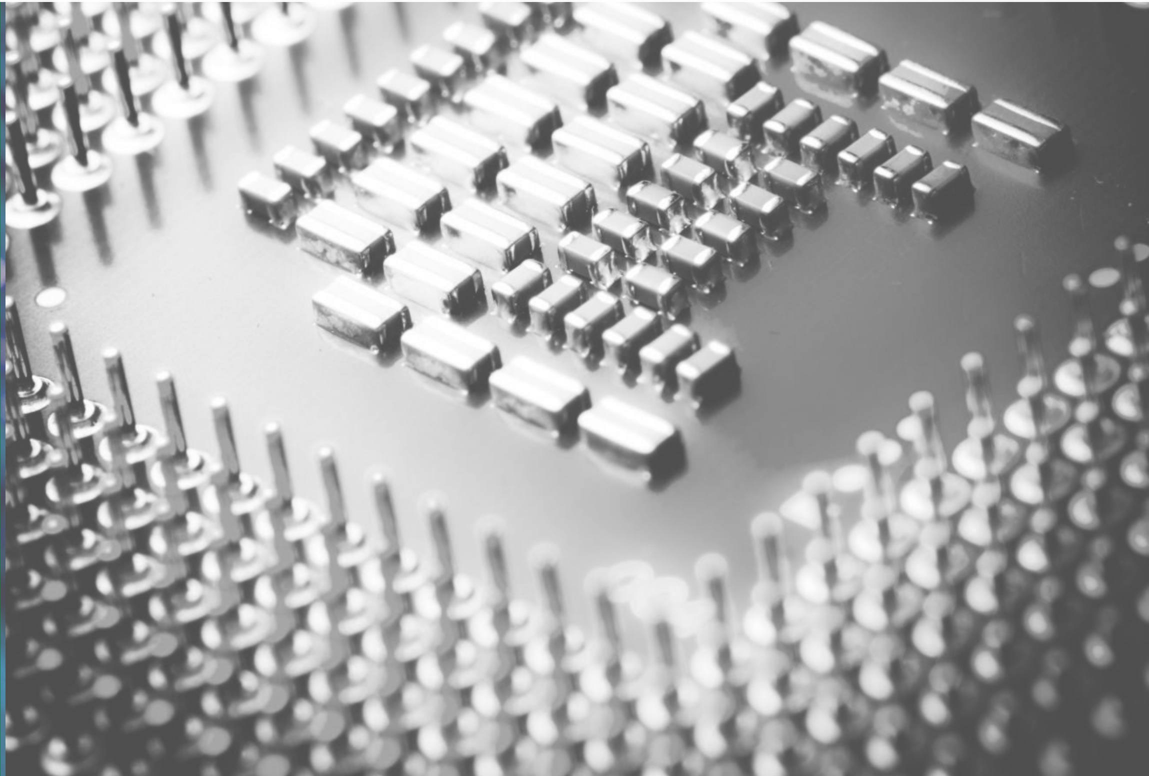
Local
Storage Systems

LEXIS DDI - Data federation

LEXIS WP3 (Lead: LRZ) – leveraging iRODS & EUDAT B2SAFE (and B2HANDLE, B2STAGE)


HPC
Cloud
Burst Buffer
orchestrated via staging REST API
iRODS/iCAT Servers LRZ (redundant)
FEDERATION – MIRRORING – PREFETCH
LRZLexisZone
LRZ: „DSS“ IBM Spectrum Scale/GPFS
EUDAT/B2SAFE
HPC
Cloud
Burst Buffer
orchestrated via staging REST API
iRODS/iCAT Servers IT4I (redundant)
IT4ILexisZone
IT4I: CEPH Storage

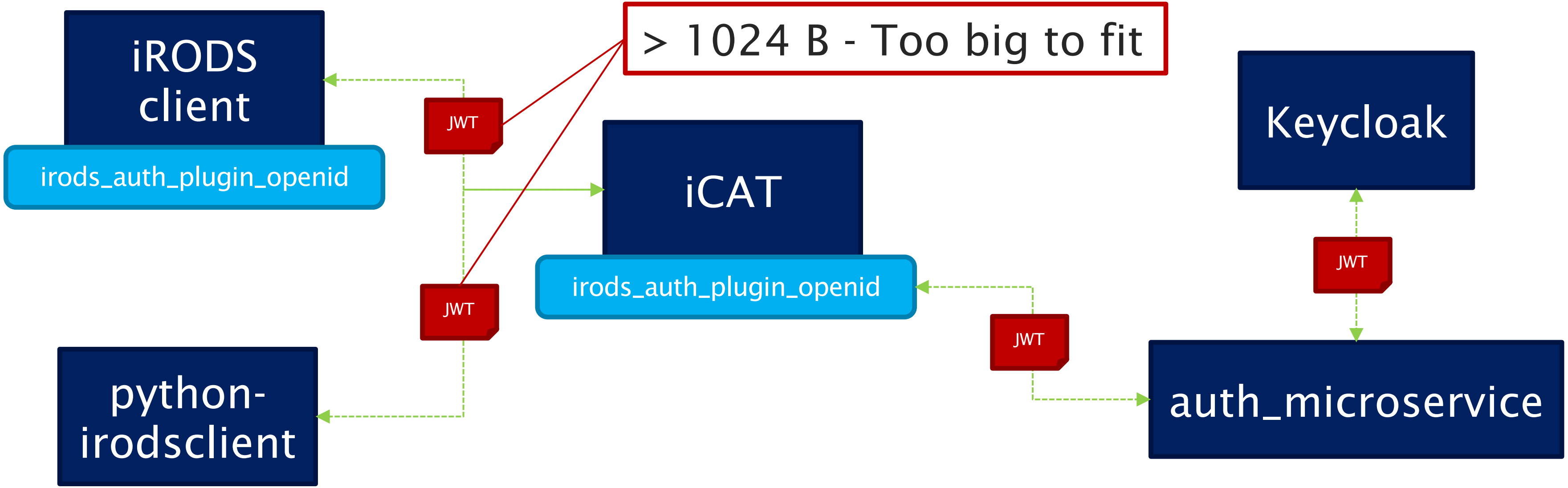
USING OPENID IN IRODS



iRODS OpenID integration

https://github.com/lexis-project/

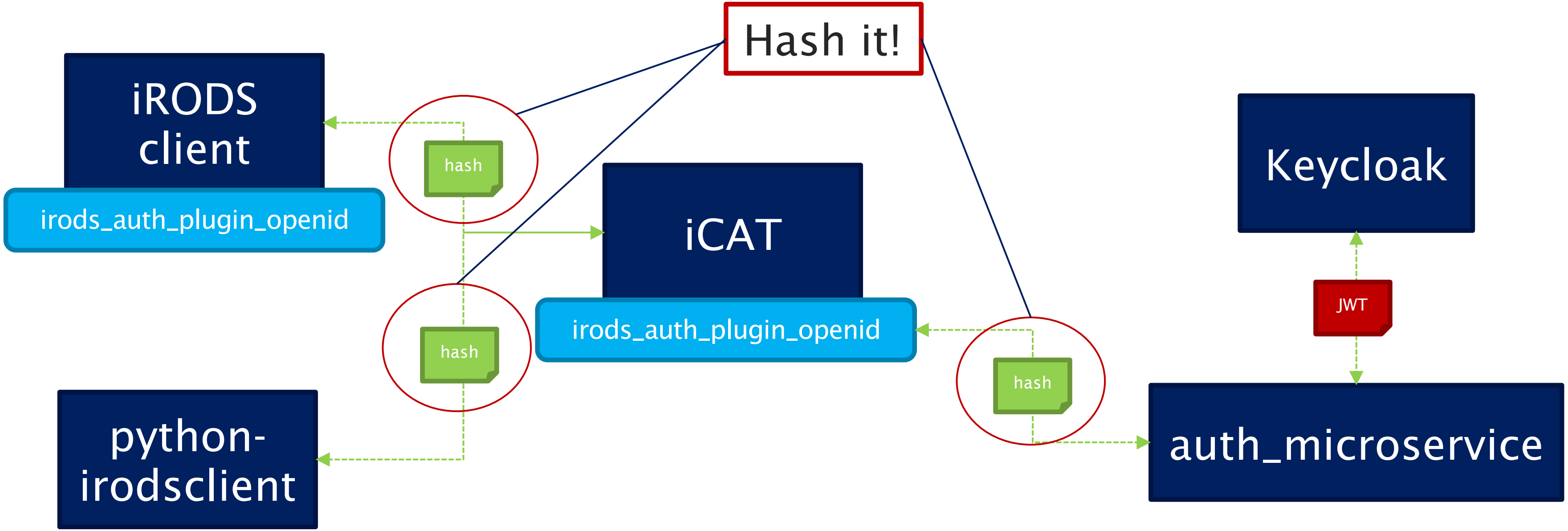
- OpenID
  - Web based authentization protocol – JWT tokens
- Keycloak
  - Open source Identity and Access Management solution – used in LEXIS as identity provider
  - Single-Sign On, Identity Brokering and Social Login, User Federation, Client Adapters



iRODS OpenID integration

https://github.com/lexis-project/

- Patches introduced by LEXIS project to auth_plugin_openid and python-irodsclient
  - Tokens larger than > 1024 B do not fit the username field in iRODS protocol
  - USER_PACKSTRUCT_INPUT_ERR: Use token hash instead of full JWT token
  - Other optimizations and extensions

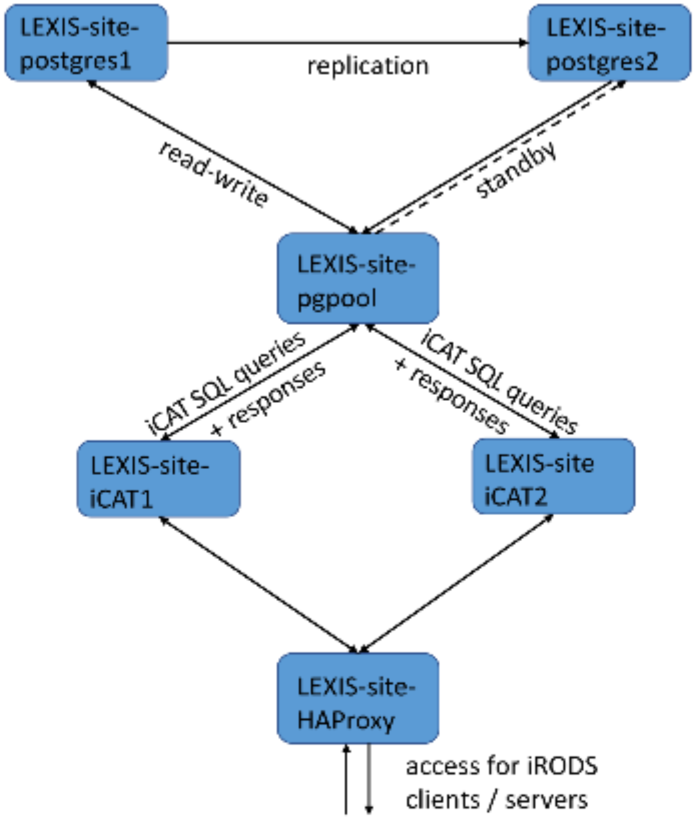


SETTING UP IRODS IN HIGH AVAILABILITY

HIGH AVAILABILITY SETUP

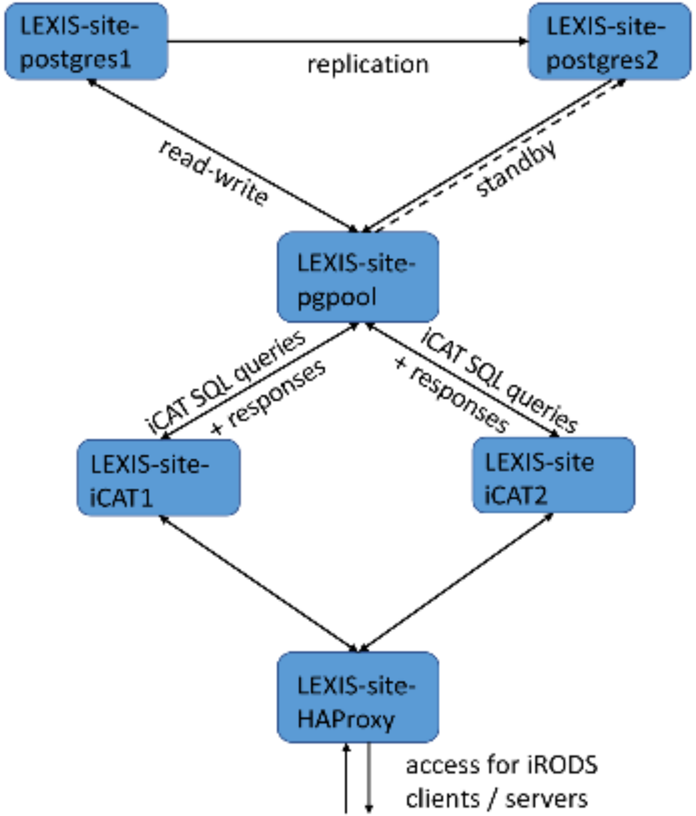
- A version of HAIRS (High-Availability iRODS System, cf. contributions of Kawai et al. to this meeting series) was deployed
  - Two instances of the ICAT server
  - A frontend instance containing HA Proxy
  - All three instances refer to themselves with the FQDN of the iRODS server

- Small problems(4.2.8):
  - Lots of error messages in the rodsServerLog
    - readWorkerTask - readStartupPack failed. -4000
  - Noisy logs causing the failure when executing some iRODS rules
    - Github issue #5471
    - readWorkerTask - readStartupPack failed. -4000



HIGH AVAILABILITY SETUP*(continued)*

- A redundant PostgreSQL database setup with repmgr and pgpool was deployed
  - Two instances of PostgreSQL containing the ICAT database
  - Replication between the two instances is enabled through repmgr
  - At a certain point in time, only one instance is set to primary and read/write access is allowed to the database
- Failover mechanism
  - Pgpool with an instance of PostgreSQL is deployed on a third machine.
  - Pgpool checks the status of the primary and the secondary databases.
  - When the primary database is down, pgpool triggers a failover mechanism



INTEGRATION WITH EUDAT

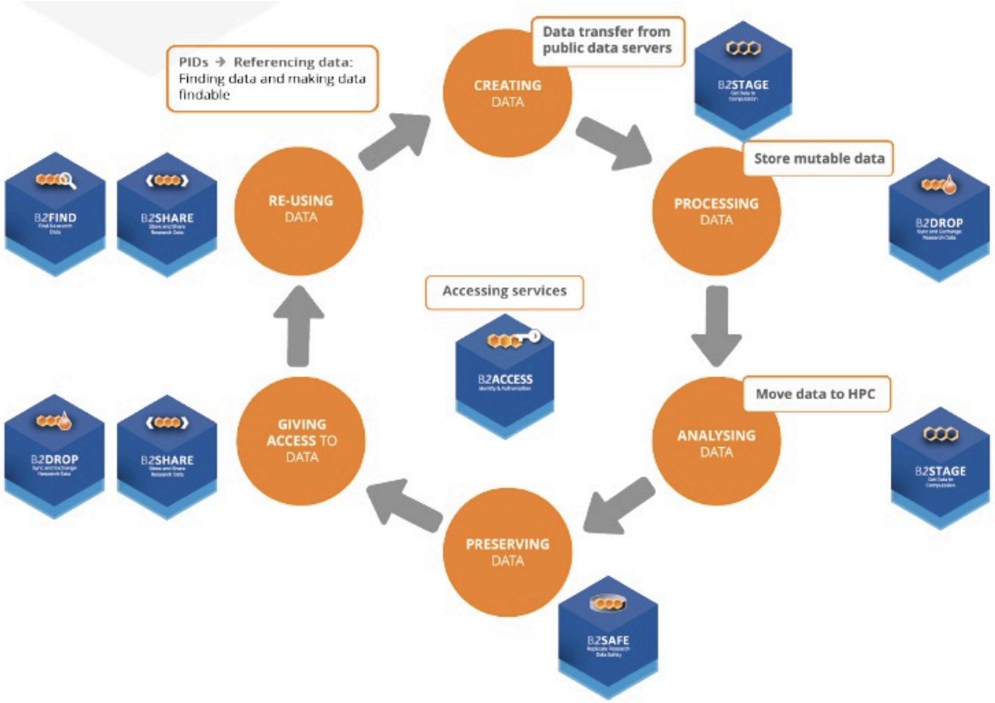
LEXIS DISTRIBUTED DATA INFRASTRUCTURE (DDI)

Immersion in European Data Management (EUDAT): EUDAT-B2* modules used

- **B2SAFE** – iRODS + add-on for policy-based data mirroring
- **B2HANDLE** – Persistent Identifier Provider
  → **FAIR** (Findable – Accessible – Interoperable – Reusable) Data
- **B2STAGE** – High Performance Data Movement

- **B2FIND** – Searchable Metadata Aggregator

- **B2ACCESS** – Authentication and Authorisation
- **B2DROP** – Data Workspace
- **B2SHARE** – Searchable Data Repository



Source: de Witt, S., "The Data Lifecycle" – presentation in EUDAT context https://eudat.eu/sites/default/files/Session1-EUDAT%20Services%20in%20the%20DLC-compressed.pdf

The FAIR side of LEXIS: Metadata, PIDs

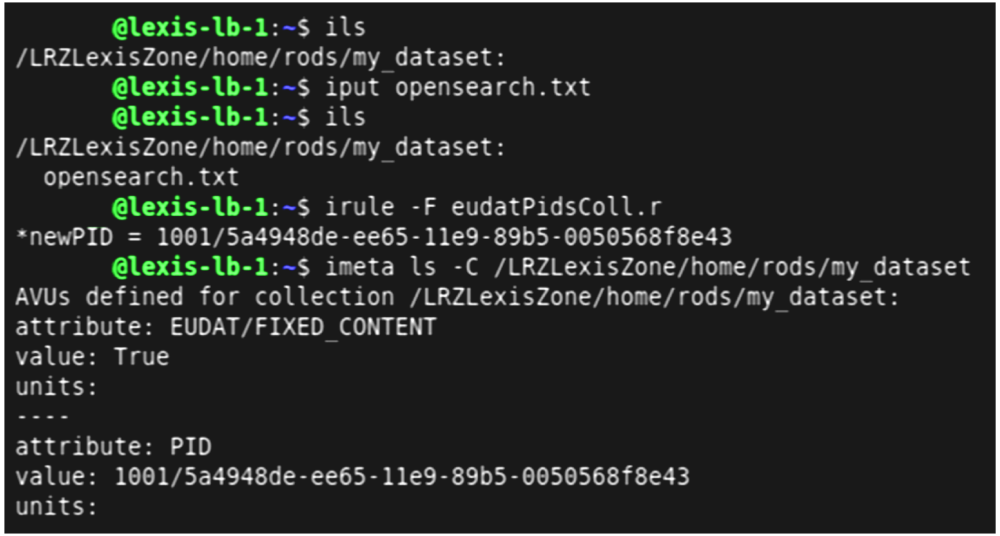
Findable, Accessible, Interoperable, Reuseable Research Data

- Most basic FAIR data requirements:
  - metadata
  - (world-)unique dataset identifier

- Metadata in LEXIS:
  - stored in iRODS Attribute-Value(-Unit) store for each data set
  - schema oriented at the basics from DataCite (schema.datacite.org)

- PIDs in LEXIS: B2HANDLE

- Aiming for findability of LEXIS public data sets via EUDAT-B2FIND



```
      @lexis-lb-1:~$ ils
/LRZLexisZone/home/rods/my_dataset:
      @lexis-lb-1:~$ iput opensearch.txt
      @lexis-lb-1:~$ ils
/LRZLexisZone/home/rods/my_dataset:
  opensearch.txt
      @lexis-lb-1:~$ irule -F eudatPidsColl.r
*newPID = 1001/5a4948de-ee65-11e9-89b5-0050568f8e43
      @lexis-lb-1:~$ imeta ls -C /LRZLexisZone/home/rods/my_dataset
AVUs defined for collection /LRZLexisZone/home/rods/my_dataset:
attribute: EUDAT/FIXED_CONTENT
value: True
units:
----
attribute: PID
value: 1001/5a4948de-ee65-11e9-89b5-0050568f8e43
units:
```

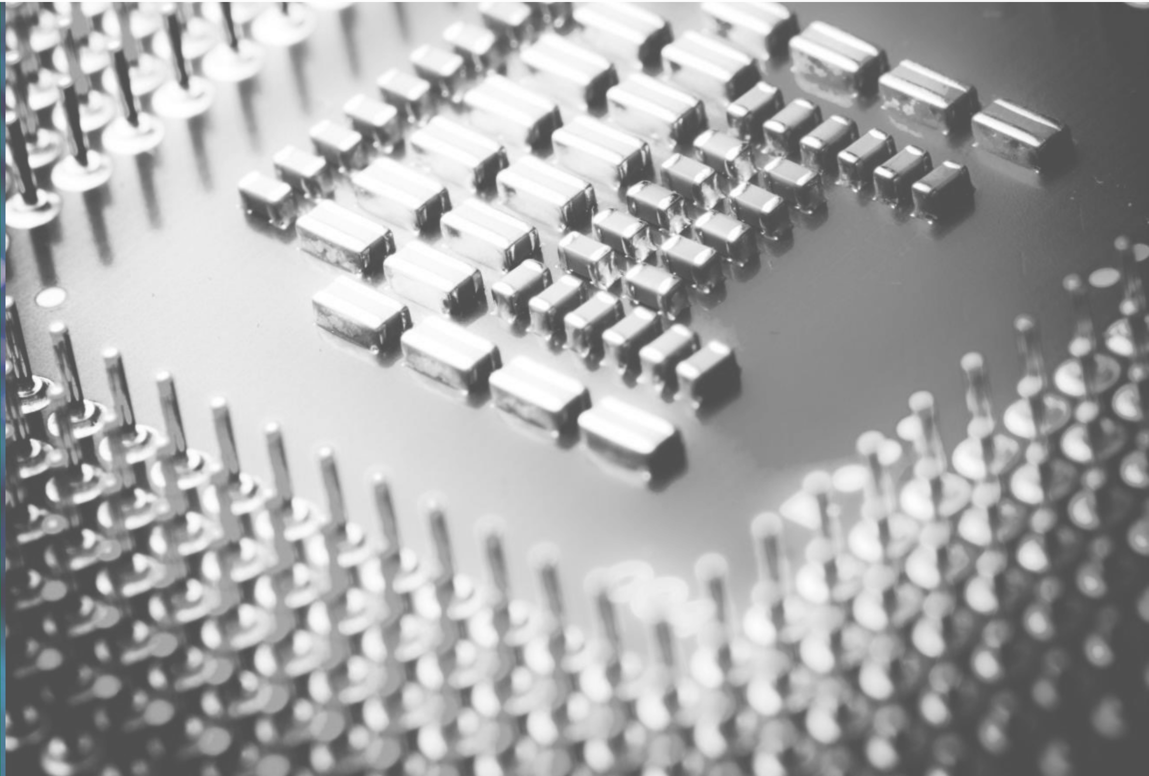
B2* SERVICES in LEXIS

- B2HANDLE
  - Based on the Handle System which offers a very reliable resolution service.
  - Adds metadata to an iRODS object/collection containing a unique PID and the PIDs of children objects/collections.

- B2SAFE
  - Adds a plugin on top of iRODS
  - Uses B2HANDLE and iRODS native rules to replicate data and keep track of children datasets

- B2STAGE
  - Adds a GridFTP server connection to iRODS.
  - Allows users to ingest data into iRODS through the reliable, high-performance GridFTP protocol

CUSTOM APIS AND THE USE OF IRODS PYTHON CLIENT



LEXIS IRODS API

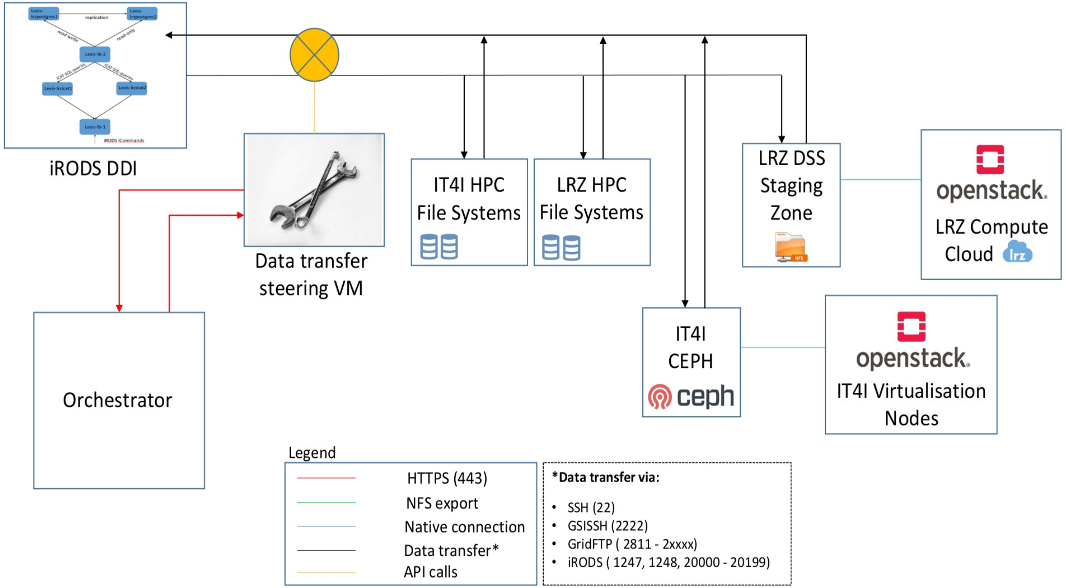
- The LEXIS iRODS API is used to:
  - Create and delete users across the federated iRODS zones
  - Create projects collections across the federated iRODS zones
  - Sets user’s ACLs based on project rights
  - Provides a token service that is used to connect to iRODS

- iRODS python client fork
  - The python client had to be forked to support openid authentication
  - https://github.com/lexis-project/python-irodsclient/tree/openid_20201105

LEXIS Staging API

Overview of the Staging API

- Django based RESTful API
- Scope: LEXIS orchestrator can move data by simple HTTP request
  - between iRODS,
  - Cloud, and
  - HPC resources at all LEXIS centers.
- Uses LEXIS AAI and the HEAppE middleware to authenticate the requests and the access to the resources
- Deploys a queuing system using Celery and RabbitMQ to allow asynchronous requests.



Legend

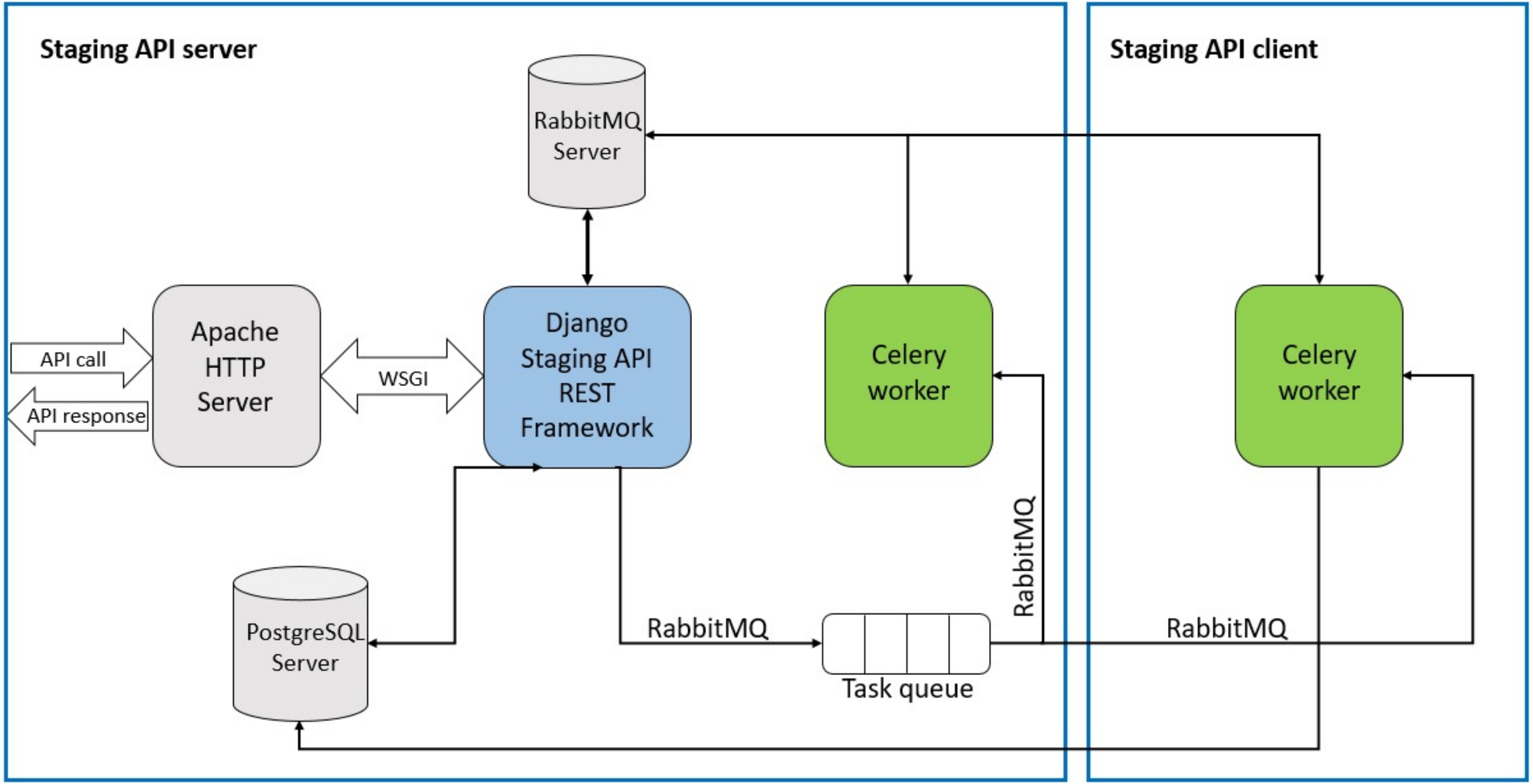
- HTTPS (443)
- NFS export
- Native connection
- Data transfer*
- API calls

***Data transfer via:**

- SSH (22)
- GSISSH (2222)
- GridFTP (2811 - 2xxxx)
- iRODS (1247, 1248, 20000 - 20199)

LEXIS Staging API *(continued)*

Architecture



Encryption and Compression API

- Django based RESTful API
- Deploys a queuing system using Celery and RabbitMQ to allow asynchronous requests.
- Allows user to encrypt and/or compress data before staging it to iRODS
- Encryption:
  - Uses aes-256-ctr
  - 1 encryption per project
  - Uses a dedicated machine with 64 VCPUs and NVME disk to perform the encryption
  - Available at each center
- Compression:
  - Staging large number of small files into iRODS results in a slow data transfer rate
  - Compressing the data before moving it to iRODS improves the transfer rate by up to x12
  - Uses a dedicated machine with 64 VCPUs and NVME disk to perform the compression
  - Available at each center

CONCLUSIONS AND OUTLOOK

- LEXIS **European Cloud-HPC Workflow Platform** (H2020) works with a **Distributed Data Infrastructure** based on **iRODS/EUDAT-B2SAFE**
- iRODS was chosen due to its ability to federate geographically distributed data sources
- Different setups of iRODS were tested. The HAIRS deployment with redundant PostgreSQL setup, provided highly available access to the federated data infrastructure.
- EUDAT services provided us with the means to achieve the DATA FAIR principles
- The iRODS Python client has been crucial for developing interfaces to other LEXIS components.
- The iRODS OpenID connection provided an obstacle when trying to connect the LEXIS AAI to iRODS. Although we found a workaround, it would be interesting to see a native iRODS implementation in iRODS 4.3.X

CONTACT

Martin Golasowski
LEXIS Task 3.2 & 3.4
martin.golasowski@vsb.cz

Mohamad Hayek
LEXIS Task 3.3 lead
mohamad.hayek@lrz.de

Website & further contacts:

www.lexis-project.eu

Large-scale EXecution for Industry & Society

LEXIS

CONSORTIUM

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

AWI

cea

Outpost24

Avio Aero
A GE Aviation Business

Atos

lrz Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

LINKS
ISMB SiTI

ITHACA

cima RESEARCH FOUNDATION

CYCLOPS

GFZ
Helmholtz Centre Potsdam

EIFFAGE TESEO

NUMTECH

ECMWF

BAYNCORE LABS