Large-scale EXecution for Industry & Society

Deliverable D2.2

Key parts of LEXIS technology deployed on existing infrastructure and key technologies specification

Co-funded by the Horizon 2020 Framework Programme of the European Union
Grant Agreement Number 825532
ICT-11-2018-2019 (IA - Innovation Action)

DELIVERABLE ID | TITLE	D2.2 | Key parts of LEXIS technology deployed on existing infrastructure and key technologies specification
RESPONSIBLE AUTHOR	Marc Levrier (Bull/Atos)
WORKPACKAGE ID | TITLE	WP2 | Requirements definition & architecture design
WORKPACKAGE LEADER	Bull/Atos
DATE OF DELIVERY (CONTRACTUAL)	30/09/2019 (M09)
DATE OF DELIVERY (SUBMITTED)	04/10/2019 (M10)
VERSION | STATUS	V1.1| final
TYPE OF DELIVERABLE	R (Report)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Marc Levrier (Bull/Atos)
INTERNAL REVIEW	Sean Murphy (CYC); Stephan Hachinger (LRZ)

Project Coordinator: Dr. Jan Martinovič – IT4Innovations, VSB – Technical University of Ostrava
E-mail: jan.martinovic@vsb.cz, **Phone:** +420 597 329 598, **Web:** https://lexis-project.eu

DOCUMENT VERSION

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Creation	08/07/2019	Marc Levrier (Bull/Atos)
0.2	Changed first 2 sections' title for more clarity (following table of content review) and started the writing of section 2.	19/07/2019	Marc Levrier (Bull/Atos)
0.3	Added WP4 and WP8 related content and new logical architecture drawing from LINKS	30/07/2019	Marc Levrier (Bull/Atos) Carmine d'Amico (LINKS) Sean Murphy (CYC)
0.4	Added inputs about iRODS and AAI/OpenID, 3D remote visualization tests in LRZ and Burst Buffer integration	12/08/2019	Ruben-Jesus Garcia-Hernandez (LRZ) Martin Golasowski (IT4I) Cédric Koch-Hoffer (Bull/Atos)
0.5	Updated LRZ infra, description of EUDAT & iRODS services, federated AAI, use of FPGA	13/08/2019	Stephan Hachinger (LRZ), Frédéric Donnat (O24), Paolo Savio (LINKS), Huy-Nam Nguyen (Bull/Atos)
0.6	Reviewed and consolidated all partners' contributions from Aug 14 to 31.	02/09/2019	Sean Murphy (CYC), Martin Golasowski (IT4I), Ruben-Jesus Garcia-Hernandez & Stephan Hachinger (LRZ), Huy-Nam Nguyen (Bull/Atos)
0.7	Further review and consolidation on overall document, filling in missing LRZ information.	04/09/2019	Stephan Hachinger (LRZ)
0.8	Consolidation of Sections 2.3.2, 3.2.1	04/09/2019	Stephan Hachinger (LRZ)
0.8.1	Acceptance of a large amount of comments and updates and fixes from Stephan Hachinger (LRZ, in v0.7 & v0.8) and merged with extra FPGA-related information from Paolo Savio (LINKS) and orchestration-related updates from Laurent Ganne (Bull/Atos).	05/09/2019	Marc Levrier (Bull/Atos)
0.8.2	Removed redundant text and figure (already present in other deliverables) and made some text sections shorter.	06/09/2019	Marc Levrier (Bull/Atos)
0.8.3	Minor formatting fixes.	09/09/2019	Stephan Hachinger (LRZ), Marc Levrier (Bull/Atos)
0.8.4	Wording and reference related fixes. Additional explanations where needed.	30/09/2019	Sean Murphy (CYC), Jan Martinovič (IT4I), Marc Levrier (Bull/Atos)
1.0	Final version, all comments and requests from reviewers resolved.	03/10/2019	Marc Levrier (Bull/Atos)
1.1	Formal check from coordinator	03/10/2019	Kateřina Slaninová (IT4I)

GLOSSARY

ACRONYM	DESCRIPTION
AAI	Authentication & Authorization Infrastructure
AD	(Microsoft) Active Directory
BPMN	Business Process Model & Notation
CaaS	Container-as-a-Service
CFD	Computational Fluid Dynamics
CH, core.h	Core.hour
CLI	Command Line Interface
DDI	Distributed Data Infrastructure
ECMWF	European Centre for Medium-range Weather Forecast
GB	Gigabyte
Gbps	Gigabit per second
GHz	Gigahertz
GPFS	Global Parallel File System (from IBM)
HPC	High Performance Computing
IAM	Identity and Access Management
IT4I	IT for Innovation (Ostrava, Czech Republic)
IaaS	Infrastructure-as-a-Service
KPI	Key Performance Indicator
LDAP	Lightweight Directory Access Protocol
LRZ	Leibniz Supercomputing Center (in Garching, Munich area, Germany)
NFS	Network File System
NVME	Non-Volatile Memory Express (flash memory direct access from PCI Express)
OIDC	OpenID Connect protocol
PaaS	Platform-as-a-Service
PB	Petabyte
RAM	Random Access Memory
RoCE	RDMA over Converged Ethernet
SAML	Security Assertion Markup Language
SBB	Smart Burst Buffer
SSD	Solid State Drive

SME	Small & Medium Enterprises
TB	Terabyte
WCDA	Weather and Climate Data API
WP	Work Package
YORC	YORC

TABLE OF PARTNERS

ACRONYM	PARTNER
Avio Aero	GE AVIO SRL
AWI	ALFRED WEGENER INSTITUT HELMHOLTZ ZENTRUM FUR POLAR UND MEERESFORSCHUNG
BLABS	BAYNCORE LABS LIMITED
Bull/Atos	BULL SAS
CEA	COMMISSARIAT A L'ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES
CIMA	Centro Internazionale in Monitoraggio Ambientale - Fondazione CIMA
CYC	CYCLOPS LABS GMBH
ECMWF	EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
GFZ	HELMHOLTZ ZENTRUM POTSDAM DEUTSCHESGEOFORSCHUNGSZENTRUM GFZ
IT4I	VYSOKA SKOLA BANSKA - TECHNICKA UNIVERZITA OSTRAVA / IT4Innovations National Supercomputing Centre
ITHACA	ASSOCIAZIONE ITHACA
LINKS	FONDAZIONE LINKS / ISTITUTO SUPERIORE MARIO BOELLA ISMB
LRZ	BAYERISCHE AKADEMIE DER WISSENSCHAFTEN / Leibniz Rechenzentrum der BAdW
NUM	NUMTECH
O24	OUTPOST 24 FRANCE
TESEO	TESEO SPA TECNOLOGIE E SISTEMI ELETTRONICI ED OTTICI

TABLE OF CONTENTS

EXECUTIVE SUMMARY 6

1 INTRODUCTION 7

2 EXISTING KEY COMPONENTS 8

2.1 HIGH PERFORMANCE COMPUTING LAYER 8
2.1.1 IT4I 8
2.1.2 LRZ 9
2.2 CLOUD COMPUTING & ORCHESTRATION LAYER 9
2.2.1 IT4I 9
2.2.2 LRZ 13
2.3 DATA MANAGEMENT LAYER 15
2.3.1 IT4I 15
2.3.2 LRZ 15
2.4 IDENTITY & ACCESS MANAGEMENT (IAM) LAYER 16
2.4.1 IT4I 16
2.4.2 LRZ 17
2.5 WIDE AREA NETWORKS 18
2.5.1 IT4I 18
2.5.2 LRZ 18

3 KEY COMPONENTS BEING ADDED WITHIN LEXIS 19

3.1 FEDERATED IDENTITY & ACCESS MANAGEMENT (IAM) LAYER 20
3.1.1 Preamble 20
3.1.2 LEXIS IAM overview 20
3.2 FEDERATED DATA MANAGEMENT 21
3.2.1 DDI on iRODS/EUDAT basis (LEXIS Distributed Data Infrastructure) 21
3.2.2 WCDA (Weather & Climate Data API) 23
3.2.3 Site-local Burst buffers 24
3.3 USE OF FPGAS (AND GPUS) 25
3.3.1 Characteristics of FPGAs 25
3.3.2 General usage considerations for LEXIS: compression, encryption, conversion, SMP jobs 25
3.4 ORCHESTRATION 26
3.4.1 Workflows to orchestrate 26
3.4.2 Deployment of YORC 28
3.5 LEXIS USER EXPERIENCE AND PORTAL LAYER 29
3.5.1 Portal functionalities and roadmap 29
3.5.2 Overview of the portal software architecture 30
3.5.3 Implementation 32
3.6 REMOTE VISUALIZATION TECHNOLOGIES 32
3.6.1 IT4I 32
3.6.2 LRZ 33

4 SUMMARY 35

A REFERENCES 36

B APPENDIX: SAMPLE WIREFRAME DESIGNS OF LEXIS PORTAL 37

LIST OF TABLES

TABLE 1: IT4I HPC INFRASTRUCTURE TECHNOLOGY AND SIZING (SOURCE: IT4I) ... 8
TABLE 2: LRZ HPC INFRASTRUCTURE TECHNOLOGY AND SIZING ... 9
TABLE 3: IT4I CLOUD PLATFORM FOR LEXIS ... 10
TABLE 4: VIRTUAL MACHINE INSTANCES ... 11
TABLE 5: IT4I CLOUD INFRASTRUCTURE TECHNOLOGY AND SIZING (SOURCE: IT4I) ... 13
TABLE 6: LRZ CLOUD INFRASTRUCTURE TECHNOLOGY AND SIZING (SOURCE: LRZ) ... 14
TABLE 7: CAPABILITIES OF DIFFERENT RELEASES OF THE LEXIS PORTAL ... 30

LIST OF FIGURES

FIGURE 1: POSITION OF WP2 IN THE LEXIS PROJECT ... 6
FIGURE 2: LEXIS FEDERATED INFRASTRUCTURE OVERVIEW ... 7
FIGURE 3: CURRENT TEST ENVIRONMENT ALLOWING TO EVALUATE OPENSTACK KEYSTONE AND KEYCLOAK SERVICES ... 17
FIGURE 4: LRZ WAN CONNECTIVITY – MAIN ROUTES OVER DFN (GERMAN RESEARCH NETWORK) TOWARDS GÉANT AND IT4I ... 18
FIGURE 5: LEXIS ARCHITECTURE OVERVIEW – NEW SYSTEM COMPONENTS ... 19
FIGURE 6: DATA STAGING BETWEEN IRODS AND GRIDFTP SYSTEMS (HPC: LEFT PANEL, USER SYSTEM: RIGHT PANEL). SOURCE: EUDAT ... 23
FIGURE 7: WP6 (TSUNAMI/EARTHQUAKE) PILOT WORKFLOW CHART ... 26
FIGURE 8: WP7 (WEATHER/CLIMATE) PILOT WORKFLOW CHART ... 27
FIGURE 9: VIEW OF THE TEST KEYCLOAK IAM CONSOLE USED TO SECURE YORC REQUESTS ... 28
FIGURE 10: VIEW A WORKFLOW DEPLOYMENT IN YORC/ALIEN4CLOUD UI (ALSO EXISTS AS AN API) ... 29
FIGURE 11: LEXIS PORTAL SOFTWARE ARCHITECTURE FOR PORTAL RELEASE R1 ... 30
FIGURE 12: ENVISIONED ARCHITECTURE OF LEXIS PORTAL RELEASE R2 ... 31
FIGURE 13: REMOTE VISUALIZATION SESSION EXAMPLE APPLIED TO PARAVIEW ... 34
FIGURE 14: LEXIS PORTAL LOGIN PAGE ... 37
FIGURE 15: LEXIS PORTAL ENTRY PAGE ... 37
FIGURE 16: LEXIS PORTAL USER PROFILE VIEW ... 37
FIGURE 17: LEXIS AVAILABLE DATASETS VIEW PAGE ... 38
FIGURE 18: LEXIS SPECIFIC DATASET VIEW PAGE ... 38

EXECUTIVE SUMMARY

The LEXIS (Large-scale EXecution for Industry & Society) project is building an advanced Engineering platform at the confluence of HPC, Cloud and Big Data which will leverage large-scale geographically-distributed resources from existing HPC infrastructure, employ Big Data analytics solutions and augment them with Cloud services.

Existing assets in the LEXIS consortium as well as new required frameworks and technologies were described in D2.1 [1] issued in M04 (end of April 2019). The present D2.2 focuses on describing the new frameworks and technologies that are being added to the existing infrastructures and services.

Objectives of the deliverable

The detailed objectives of deliverable D2.2 are the following:

- Describe the technologies and frameworks being added to the LEXIS assets:
    - Why they have been selected,
    - Where they fit in the LEXIS system architecture,
    - How they interact with one another,
    - How they impact the pilots and related user experience.
- Update the overall LEXIS system architecture:
    - System architecture representations,
    - System workflow representations aiming at supporting the required pilots and used as main inputs for the WP4 and pilot teams to start implementing the related workflows in the orchestrator.

Position of the deliverable in the general context of the project

D2.2 is a product of WP2 Requirements Definition and Architecture Design in its Task 2.1 Infrastructure Evaluation and Key Technology Identification for LEXIS [1]. This report document is to be delivered at the end of M09.

As shown Figure 1, WP2 needs to set the foundations for business cases (WP5 to 7) to benefit from advanced techniques brought in by WP3 (Data management using NVME Burst Buffers) and WP4 (Orchestration and cloud services). An access portal (WP8) will also be part of the LEXIS architecture to provide easy access from Industries, SMEs and Academia looking for high Performance computing and Data Management.

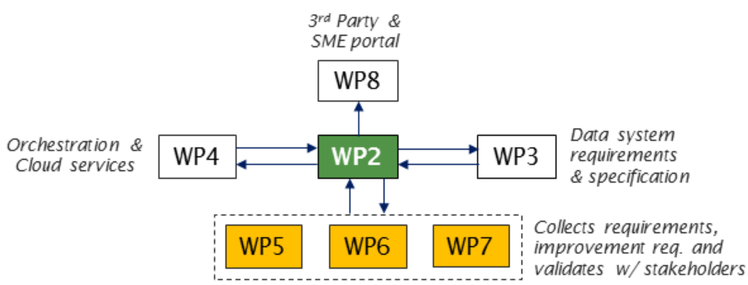

Figure 1: Position of WP2 in the LEXIS project

The main contributors to this deliverable are the "technical work packages", namely WP2, WP3, WP4 and WP8:

- Bull/Atos as the work package leader as well as for Burst-Buffers, Orchestration and FPGA related topics (WP2),
- IT4I as one federated HPC service provider and project lead,
- LRZ as the other federated HPC service provider and work package leader (WP3),
- LINKS as the coordinator of codesign tasks and work package leader (WP4),
- O24 for security related aspects and CYC for the portal related topics and work package leader (WP8).

1 INTRODUCTION

About the present deliverable (D2.2)

WP2 Requirements Definition and Architecture Design deals with both application use cases requirements and overall system design of underlying infrastructures and services.

Deliverable D2.1 [1] **was issued in M04**, which listed both existing Technology assets (hardware, software, middleware and skills) in the consortium and extra technology required to reach the project's objective. Figure 2 gives an overview of the LEXIS system architecture that was designed during this initial phase.

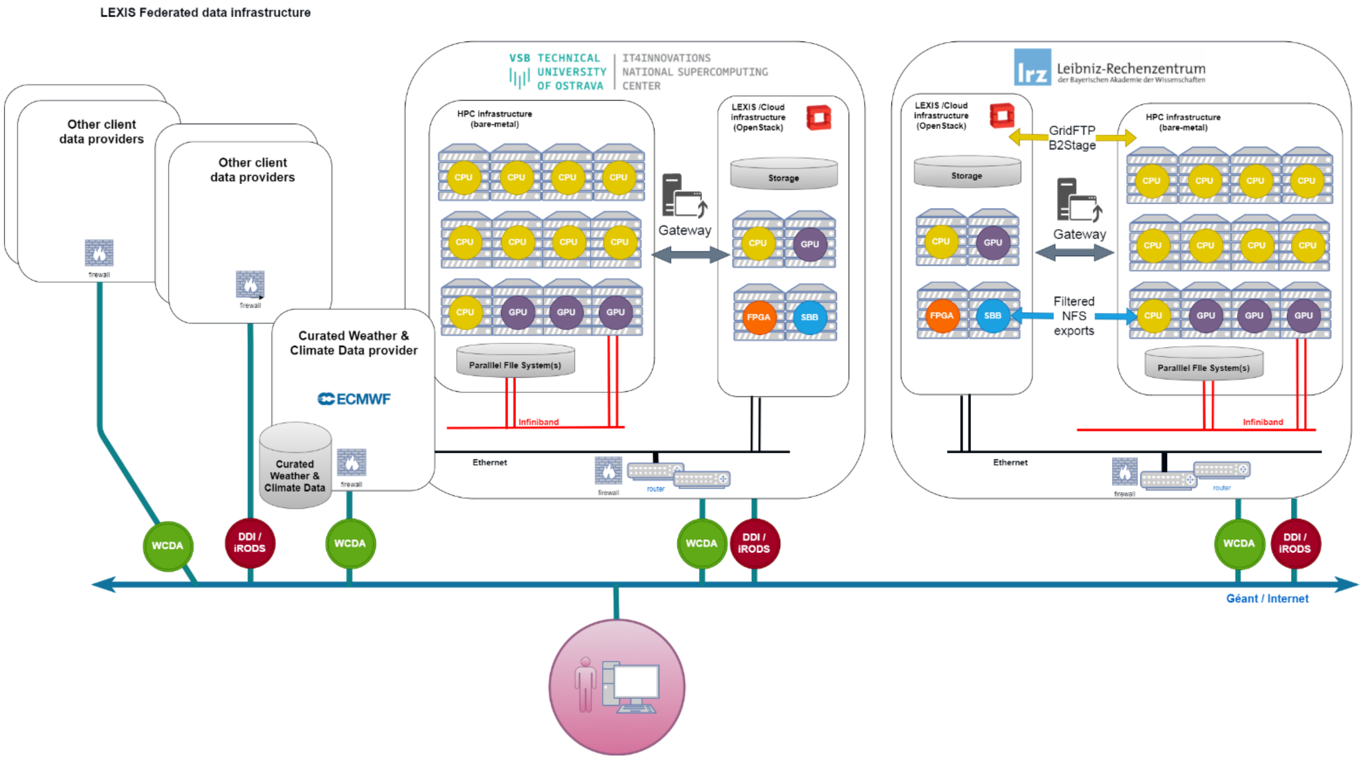

Figure 2: LEXIS federated infrastructure overview

D2.2 is the logical continuation of D2.1 [1], which means that the focus is now the new required key technologies: their identification, the **related** selection criteria, how they will work together as well as the resulting impact on pilots and users' experience.

Our method for creating D2.2 consisted in making an inventory of all the necessary interactions between the significant number of LEXIS subsystems: hardware resources, cloud layers, data management systems, as well as various system services, HPC middleware, protocols and APIs. This was done with a focus on LEXIS sites' coherent federation, both in terms of security and resource management. To make sure these new frameworks and technologies were compatible, technical surveys and prototypes were used.

Structure of the document

- **Section 2** covers the existing assets like facilities, hardware, software and middleware present in both HPC service providers of the LEXIS consortium (IT4I and LRZ).
- **Section 3** which focuses on additional hardware, software and middleware components and related methodology required to implement the LEXIS platform.
- Finally a summary is provided in **Section 4**.

2 EXISTING KEY COMPONENTS

This section constitutes a detailed update regarding the status of the technical environments available in LEXIS, which already were briefly described in D2.1 [1]. It also gives information about the way these are key to reach LEXIS's objectives and how the consortium plans to operate them.

2.1 HIGH PERFORMANCE COMPUTING LAYER

One of the LEXIS platform's objectives is to orchestrate elaborate workflows dealing with very large data sets and containing HPC parts.

This section gives a brief reminder and update of existing HPC / bare-metal infrastructures and related activities on the primary LEXIS computing sites (LRZ and IT4I).

2.1.1 IT4I

Table 1 below inventories the four clusters available at IT4I, their peak power and main hardware components/architecture.

	Salomon cluster	Anselm cluster	Barbora cluster	DGX2
RPeak PFlops	2.00	0.9	0.84	0.13
#Nodes	1,008	209	198	1
Node types (CPU, GPU)	Bi-E5-2680V3-12c Bi-Phi 7120P-61c UV2000 SMP112c No local HD	180x2xE5-2470-8c 23xNvidia K20m 4xPhi 5110P	189x2xCL-18c 8x4xNvidia V100 1x4x12c fat node	2x24c w/ AVX-512 16x Nvidia V100
Interconnect	IB FDR56 7D enhanced HC	IB QDR40 non-blocking fat-tree	IB HDR200 fat-tree	NVLink (12xNVSwitch, 2.4Tbps/bissection) 8x100Gbps IB
Ethernet	4x10g WAN	4x10g WAN	Not mentioned	Not mentioned
Storage	1.7 PB Lustre 0.5 PB NFS Ramdisk on nodes	0.15 PB Lustre 0.3 PB Lustre (homes)	Burst buffer scratch 200 TB Home 25 TB	30 TB NVMe SSD
Software	Linux CentOS7 LMOD tools PBS Pro	Linux CentOS7 LMOD tools PBS Pro	Linux RHEL7 PBS Pro	Linux distro packaged by Nvidia

Table 1: IT4I HPC infrastructure technology and sizing (source: IT4I)

2.1.2 LRZ

Table 2 below inventories the clusters available at LRZ, their peak power and main hardware components/architecture [2] [3] [4].

	SuperMUC-NG cluster	SuperMUC cluster (remaining part "Phase 2", to be decomissioned end of 2019)	Linux Cluster
RPeak PFlops	26.9	3.58	c. 1
#Nodes	48 cores/node RAM per node: 6,336 Thin: 96 GB 144 Fat: 768 GB 32 Cloud GPU: 768GB 82 Cloud: 192 GB 1 Cloud Huge: 6TB	3,072 (86,016 cores, 194 TB RAM total)	198 RAM per node: CoolMuc2: 64 GB, CoolMuc3: 96 GB, Teramem: 6144 GB
Node types (CPU, GPU)	X86, Intel Skylake Xeon Platinum 8174	X86, Haswell Xeon Processor E5-2697 v3	c. 10000 KNL cores c. 10000 HSW cores 640 IVB cores 96 SMP cores
Interconnect	OmniPath, 100 Gbit/s.	Infiniband FDR14	Infiniband, OmniPath, ethernet
Ethernet	100 Gbit/s	100 Gbit/s	100 Gbit/s
Storage	50 PB @ 500 GB/s 20 PB @ 70 GB/s 256 TB (home)	15PB (scratch/work), 3.5PB (home)	GPFS, 1.4 TB NFS, several TB
Software	Linux SLES12 Slurm	Linux SLES11 IBM LSF	Linux SLES12 IBM LSF

Table 2: LRZ HPC infrastructure technology and sizing

2.2 CLOUD COMPUTING & ORCHESTRATION LAYER

This section describes existing cloud / virtual infrastructures available or planned for both locations. Both locations use OpenStack to manage and orchestrate their Infrastructure-as-a-Service & Platform-as-a-Service (IaaS & PaaS) offers but have their own system administration strategies and tools.

2.2.1 IT4I

The infrastructure at IT4I which is dedicated to testing and operation of the cloud part of the LEXIS platform is roughly divided in the following parts (Table 3):

- A set of 12 dedicated HPC nodes from the Anselm cluster,
- Dedicated/new LEXIS hardware (VI nodes, SBB, CEPH storage servers),
- Operational VMWare instance operated by IT4I supercomputing services.

Nodes	Purpose
cn7-cn12	OpenStack (Rocky)
cn13-cn15, cn18	CEPH storage (OSDs, monitors, managers, MDS)
cn16-cn17	VMWare vCenter 6

Table 3: IT4I cloud platform for LEXIS

Initial situation

Before the new hardware for LEXIS cloud servers arrives, dedicated HPC nodes from the Anselm cluster are currently being used for a proof-of-concept deployment of the software stack necessary for cloud operations and support of the LEXIS platform itself. The dedicated, new LEXIS hardware is now in procurement. Network topology and details were presented in [1]. Once installed, this infrastructure will be used for both operation and development of the LEXIS platform. The third resource (operational VMWare infrastructure) mentioned is used to create dedicated VMs used for proof-of-concept of services which must be accessible from public internet.

In this initial phase, OpenStack[1], CEPH[2] storage and VMWare vCenter are deployed on these dedicated HPC nodes from the Anselm cluster. The nodes are accessible only through IT4I VPN on a private subnet. Access to this subnet is allowed only per request authorized by the project coordinator. The purpose of this deployment is to get familiar with the operational and monitoring procedures of the individual components which will be part of the LEXIS Cloud infrastructure at IT4I.

Testing OpenStack deployment

The OpenStack deployment is realised using the *kolla-ansible* distribution, which provides automated deployment of necessary daemons running in Docker containers. This choice will provide us an easy way to scale the OpenStack deployment if necessary. The current deployment is based on the Openstack *Rocky* release, where:

- 4 nodes are used as KVM hypervisors,
- 1 node runs networking services,
- 1 node runs the management and monitoring services.

Testing CEPH deployment

The CEPH deployment is done by *ceph-ansible*[3] and is based on the *Mimic*[4] release. 1 TB of raw storage is available on 3 OSDs. Other services run in virtual machines in the proof of concept VMWare deployment. This CEPH instance is used as storage backend for the OpenStack deployment.

Testing VMWare deployment

The testing VMware deployment consists of two nodes running ESXi hypervisor and one vCenter appliance running as VM, having 64 vCPUs and 128 GB RAM available. This VMware instance only has HDDs available on the compute nodes available as datastores. The CEPH cluster can be connected to this VMWare cluster by iSCSI protocol, if necessary.

Testing iRODS, Keycloak and YORC deployments

[1] OpenStack: www.openstack.org
[2] CEPH: https://ceph.io
[3] Ceph-ancible: https://docs.ceph.com/ceph-ansible/master/
[4] Mimic release: https://docs.ceph.com/docs/master/releases/mimic/

iRODS[5] is hosted on two VMs running on the operational VMware instance at IT4I. Several configurations, including zone federation are being tested. A Keycloak[6] instance has been also deployed on the same infrastructure. Three VMs have been provided to Bull/Atos for testing the YORC[7] deployment. Table 4 below provides a summary of all VMs created for testing of the platform components. The intention for using this VMware instance and not the testing one deployed on the HPC nodes is simply to get more stability and to avoid potential interference caused by an unstable environment.

Monitoring

Basic monitoring services based on *Prometheus time series db*[8] and *Grafana*[9] have been deployed. CEPH metrics, OpenStack and hardware (NodeExporter) metrics are collected. Internal infrastructure such as Active Directory or Exchange servers have been tested in VMs in the VMware instance. This VMware instance also hosts two VMs to be used by CYC for the proof-of-concept deployment of their cloud billing platform.

OpenStack IAM

The OpenStack Horizon dashboard and Keystone identity service have been connected to the testing Keycloak instance deployed at IT4I ; the OpenID protocol is used for the authentication flow.

VM	Hostname	Purpose
1	YORC.it4i.cz	YORC main server
2	YORCwl1.it4i.cz	Dummy workload VM #1
3	YORCwl2.it4i.cz	Dummy workload VM #2
4	irods1.it4i.cz	iRODS Provider and iCAT
5	irods2.it4i.cz	iRODS Consumer
6	keycloak.it4i.cz	Keycloak testing

Table 4: Virtual machine instances

The proof-of-concept deployment will remain present until the planned LEXIS infrastructure will be in place and ready to use.

Work in progress

The planned procurement of the LEXIS (cloud) hardware will contain the following components:

- Virtualization and Cloud nodes,
- CEPH Storage servers,
- Network switches,
- Gateway servers,
- Two Burst buffer servers.

[5] iRODS: https://irods.org
[6] Keycloak: https://www.keycloak.org
[7] YORC: https://github.com/ystia/yorc
[8] Prometheus: https://prometheus.io
[9] Grafana: https://grafana.com

Technical description of the virtual infrastructure

The virtualization infrastructure will consist of 6 servers:

- 3 servers will be used as VMware ESXi hypervisors,
- Another 3 will be used to run the OpenStack cluster.

The VMware instance will be primarily used to host support services necessary for the platform operations (internal monitoring and security services, CEPH support services, etc.) and essential components of the LEXIS platform - AAI system, YORC, iRODS servers for DDI, Cyclops Cloud billing and monitoring system and other support services. These central services are permanent, critical and not going to evolve very much and will benefit from using the enterprise-grade VMWare stack, which is present on both sites, very mature and resilient.

The OpenStack instance, which is more adapted to dynamic workload management, will be used to run the computational workloads through the YORC. It can be also configured to support Kubernetes clusters or bare-metal workloads if necessary.

Both systems will be connected to the CEPH storage which will be deployed on 4 Object Storage Daemon (OSD) servers, each having 30 TB of storage and redundant 25 Gbps connection available. Depending on the level of redundancy required, the storage will have at least 60 TB available. The CEPH deployment will use the VMware instance to run its support services (MON, RGW, MDS and MGR servers). The CEPH will expose apart from its native interface an iSCSI target for serving block devices via iSCSI and RADOS gateway for serving objects, including S3 compatible objects.

The network used by this infrastructure will be based on 100G Ethernet provided by fully redundant Mellanox switches and ConnectX-5 NICs[10]. This network provides good flexibility, performance and compatibility with protocols which may be used later for deployment of the Smart Burst Buffers, like RDMA over Converged Ethernet[11], SR-IOV[12] for virtualization of PCIe resources and others. The network will be separated into several VLANs (storage, Cloud and VMware internal, public, etc.). Apart from this network, there will be also a management network based on 1G Ethernet, which will be used only for administrative purposes. WAN connection will be realized through internal IT4I infrastructure which currently offers 4 x 10 Gbps direct lines with upgrade to n x 100 Gbps in progress.

The gateway servers will provide a fully redundant connection to the internal fabric of Salomon and Barbora HPC clusters at IT4I. These servers will be also used to provide the NFS exports of SBB resources, if necessary. These servers will have additional NICs deployed to support the respective cluster fabrics (HDR and FDR InfiniBand).

The two burst buffer servers will provide 12.8 TB of fast SSD NVMe storage each. They will be connected to the internal 100G network of the infrastructure and will also have a dedicated NICs installed to support possible direct connection to the HPC cluster internal fabric. The servers will be also ready for installation of additional accelerator card in the PCIe x16 form factor with additional power available (GPU or FPGA).

Related system architecture diagrams are available in [1].

The entire infrastructure (except for the two Gateway servers) will be under administration of the IT4I LEXIS team to support various experimental configurations and modifications in a timely manner. Table 5 below provides an overview of the planned infrastructure.

[10] Mellanox switches and ConnectX-5 NICs: http://www.mellanox.com/related-docs/prod_adapter_cards/PB_ConnectX-5_EN_Card.pdf
[11] RoCE: https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet
[12] SR-IOV: https://blog.scottlowe.org/2009/12/02/what-is-sr-iov/

Equipment type	Planned procurement
#Servers	3 VMware, 3 OpenStack, 2 SBB, 2 GW, 4 OSD
	Not chosen yet + 2 burst buffer nodes (possibly w/ FPGA or GPU)
Interconnect	100G Ethernet (Mellanox)
Ethernet	100G fully redundant, 1G admin network
Storage	120 TB raw, 60 TB redundant (CEPH)
Software	Openstack, Kubernetes, VMWare ESXi or KVM, Yorc, HEAppE, LEXIS support services

Table 5: IT4I cloud infrastructure technology and sizing (source: IT4I)

The installation and deployment of the infrastructure will occur during Fall/Winter 2019, while experimental operational status should be reached in March 2020.

2.2.2 LRZ

In LRZ, the situation with the Cloud-Computing layer is much different: larger, already operational on-premises infrastructures will be used for LEXIS. Technologies are, however the same as in IT4I (OpenStack LRZ Compute Cloud + LRZ VMWare infrastructure). The Compute Cloud runs dynamic, user-created VMs on OpenStack while the VMWare infrastructure runs VMs which must be set up by LRZ administrators.

While the LRZ infrastructures are large enough to tackle heavy workloads, they are generally shared (at any moment) with other LRZ users and operated under fixed security regulations/concepts.

The challenge at LRZ is (or has been) merely to choose the right infrastructure for the workload, and to utilise it for LEXIS. As an important R&D aspect, at IT4I and LRZ a “LEXIS deployment” is thus tested and verified under very different conditions.

Testing iRODS, Keycloak and YORC deployments

Test deployments of these high-level components have been made at LRZ as in IT4I (see Section 2.2.1), with nearly all components running on the Compute Cloud.

Technical description of the virtual infrastructure

Table 6 below lists key characteristics of the virtual infrastructure (VMWare and OpenStack parts) at LRZ.

	VMWare IaaS/PaaS	OpenStack Compute Cloud
RPeak PFlops	in process of substantial extension	no benchmarking over whole cloud, aggregate between 0.1 and 1 (depending on GPU usage)
#Nodes	in process of substantial extension (currently about 1,000 VMs running)	114
Node types (CPU, GPU)	>500 x86_64 (Intel) CPU cores	about 4,000 recent Intel CPU cores, some nodes equipped with V100 GPUs
Interconnect	100GE	100GE
Ethernet	see interconnect (+ management network)	see interconnect (+ management network)
Storage	NAS-filer backend for VM images, recently being replaced	about 1PB CEPH backend (access to DSS storage via NFS)
Software	VMWare	OpenStack

Table 6: LRZ cloud infrastructure technology and sizing (source: LRZ)

Both infrastructures are operated by the LRZ group "ITS" and further documentation is available via doku.lrz.de. The group is also responsible for monitoring, which is continued overnight and weekends by LRZ night operators.

Network connectivity of VMs in both infrastructures is freely configurable, from (nearly) complete isolation to reachability under a public (internet-routable) IP address. Firewalling via *PFSense* (in massive use cases) and OpenStack Security Groups is available.

Due to the security concept, no hardware apart from the graphics card is usually passed through to the guest operating system, and the virtualised drivers must be used with corresponding restrictions.

Within LEXIS, burst buffer machines are deployed at LRZ. The concept for their placement within LRZ infrastructure and for their technical specification has been worked out with the project partners. The deployment will serve as a test on how these machines can be immersed in an existing, relatively fixed infrastructure. They are physically placed near the cloud machines, with firewalled/limited connectivity over the backbone (100GE GW) to the HPC clusters. This situation (including the fixed operational concept and security/firewalling regulations for HPC systems) is similar to IT4I. The Data-Node/Burst-Buffer hardware is expected to be delivered in M09.

The LRZ HPC, cloud ("VI Nodes") and storage infrastructure are the most relevant for LEXIS, including the placement of smart burst buffers. OpenNebula (Nebula VI nodes) are being decommissioned, while the OpenStack VI has meanwhile been upgraded to about 4,000 CPU cores. Related system architecture diagrams are available in [1].

The restrictions with the operational environments at LRZ are gracefully handled within the project, with IT4I being able to test somewhat more experimental low-level/hardware solutions and the focus of LRZ being somewhat more on the distributed data infrastructure and web portal.

2.3 DATA MANAGEMENT LAYER

This section describes existing data management infrastructures and related activities at IT4I and LRZ. For details about existing and planned local storage infrastructures at IT4I, LRZ and ECMWF the reader is referred to Deliverable D3.1 [5].

2.3.1 IT4I

Each HPC cluster operated by IT4I has its own storage facilities usually based on a dedicated SAN solution. The detailed configuration is available online[13]. Essentially, each cluster has its own HOME storage which is intended for user data, it is usually backed up on tapes and it is realized through tiered NFS. SCRATCH storage is intended for large intermediate and ephemeral data and usually realized by LUSTRE[14] to provide sufficient bandwidth for parallel I/O operations. Quotas on total size and file count are imposed by default on all locations, however can be adjusted for special cases. File Access Control Lists are enabled by default on all locations.

Users can upload data through SCP/SFTP and rsync protocols through clusters‘ login nodes. The login nodes allow connection to externel HTTP(S) or FTP servers. GridFTP is also available upon request. Users can also use a GitLab instance[15] operated by the IT4I supercomputing servicesfor source code versioning.

Within the LEXIS context, two VMs (irods1-2.it4i.cz) with proof of concept iRODS instance have been deployed. There servers were used to test iRODS zone federation between LRZ and IT4I. Operational iRODS instance will be deployed in the VMware instance in the planned infrastructure. Currently a testing of AAI and CEPH integration with iRODS is in progress.

2.3.2 LRZ

At LRZ, HPC clusters have 3-layer storage system (HOME, WORK/PROJECT, SCRATCH), where the characteristics of HOME and SCRATCH management are very similar to the IT4I systems, even though back-ends are partially NAS filers and partially IBM Spectrum Scale (GPFS) appliances for massively-parallel I/O. At LRZ, SCRATCH features no user/group quotas but high-watermark deletion, making it an “unsafe” storage for temporary massive usage without administrative restrictions and without any guarantee on when data will be deleted. On WORK/PROJECT, data are permanent, and a group quota applies to each computational project.

Apart from these storage systems, LRZ offers the LRZ Data Science Storage (DSS), a GPFS system set up to ensure maximum performance as well as connectivity to practically all computational resources within LRZ. While the DSS is mounted via GPFS drivers on HPC systems, other systems can access it via NFS export, for which the requesting IP can be registered via a web API/user interface. DSS spaces are administrated as separate LRZ projects (cf. Section 2.4.2), whose master user can make (within quota restrictions) “storage containers”, to which any LRZ users can be given “invited” access via the DSS API and user interface. Due to its comprehensive and performant connectivity, DSS will play an important role as a storage backend for LEXIS at LRZ.

Details on LRZ’s HPC file systems and the DSS have been presented in D3.1 [5].

The data-transfer mechanisms of SCP and SFTP, as well as GridFTP for HPC systems, are offered at LRZ. In addition, LRZ’s GLOBUS endpoints play an increasing role in massive data transfer.

iRODS has been deployed at LRZ for a past project (AlpEnDAC.eu), reaching production state. For LEXIS, it is deployed on virtual machines (Compute Cloud for test instances, VMWare infrastructure for production instances). A redundant, production-ready iRODS deployment at LRZ is currently in progress (2 iCAT-enabled iRODS “provider”

[13] IT4I Infrastructure documentation: https://docs.it4i.cz
[14] LUSTRE: http://lustre.org/
[15] IT4I GitLab: https://code.it4i.cz

servers and 2 Postgres VMs being installed). AAI components tested at IT4I will be included in the system soon, as will be a larger storage back-end (iRODS "resource"), probably in the form of a DSS container.

2.4 IDENTITY & ACCESS MANAGEMENT (IAM) LAYER

This section describes existing identity and access management infrastructures and related activities on LEXIS's locations. All services provided by LEXIS (including our iRODS-based DDI) delegate their security operations to the LEXIS federated AAI. The AAI's main security components (authorization, authentication and identity services) are federated and abstracted behind the Keycloak tool[16], compliant with the Oauth2 framework.

Thanks to Keycloak (with possible other mechanisms as a supplement), we intend to unify the security management across all LEXIS functional services (data management, orchestration and other APIs) and across locations (IT4I, LRZ & ECMWF). Keycloak can handle realms (separate groups of users accounts) and locations in a flexible way, allowing to take into accounts providers' own constraints, like realms, directory services' schemas or backend types.

2.4.1 IT4I

One important point to mention is that IT4I teams have developed a middleware (named HEAppE[17]) that provides a secure encapsulation of HPC job management based on a mapping between an external end-user account and internal "technical" HPC accounts. HEAppE has also been deployed in LRZ in order to unify the security model between the cloud and HPC parts across the LEXIS facilities.

Access to HPC resources (via HEAppE)

- Pre-allocated HPC resources are approved for a specific computational project (e.g. 50,000 core.h).
- Service cluster accounts are bound to a specific computational project and therefore can be used only for the job submission under a specific project ID. They are the only ones allowed to run the users' workloads and consume the allocated core hours on behalf of this project.

Only the LEXIS users that are part of the approved LEXIS project group, with a valid computational project ID in IT4I or LRZ, can use HEAppE to submit jobs.

Access to cloud resources

Compared to the existing IAM environment in IT4I as presented in [1], the plan for the future OpenStack deployment on IT4I LEXIS infra is now evolving as follows:

- Keystone and Horizon services can be directly integrated with Keycloak using OpenID protocol.
- This deployment is for LEXIS purposes only, it is completely separated from IT4I HPC infrastructure.
- There are no external restrictions on the resource quotas now and internal decisions can be made according to the specific WP needs.
- LEXIS users can have optional direct access to the OpenStack Horizon dashboard and API (indirectly via the YORC API).
- The incoming LEXIS user from Keycloak is mapped to a user account stored in Keystone. If the account does not exist yet, it is created.
- During the login procedure, mapping rules are applied as described at OpenStack web[18]
- Mapping rules will respond to a few basic principles (Figure 3):
    - OpenStack user has same metadata as in Keycloak (username, full name, email, etc.).
    - Login is allowed only if a specific attribute is set in user profile in Keycloak (i.e. CloudUser=True, or similar).
    - LEXIS Computational Project is mapped to the same existing project in OpenStack.

[16] Keycloak: www.keycloak.org
[17] HEAppE: heappe.eu
[18] OpenStack mapping rules: https://docs.openstack.org/keystone/pike/advanced-topics/federation/mapping_combinations.html

- Creation of the computational project in OpenStack shall be triggered by the approval process, which also provides quotas on resource usage.

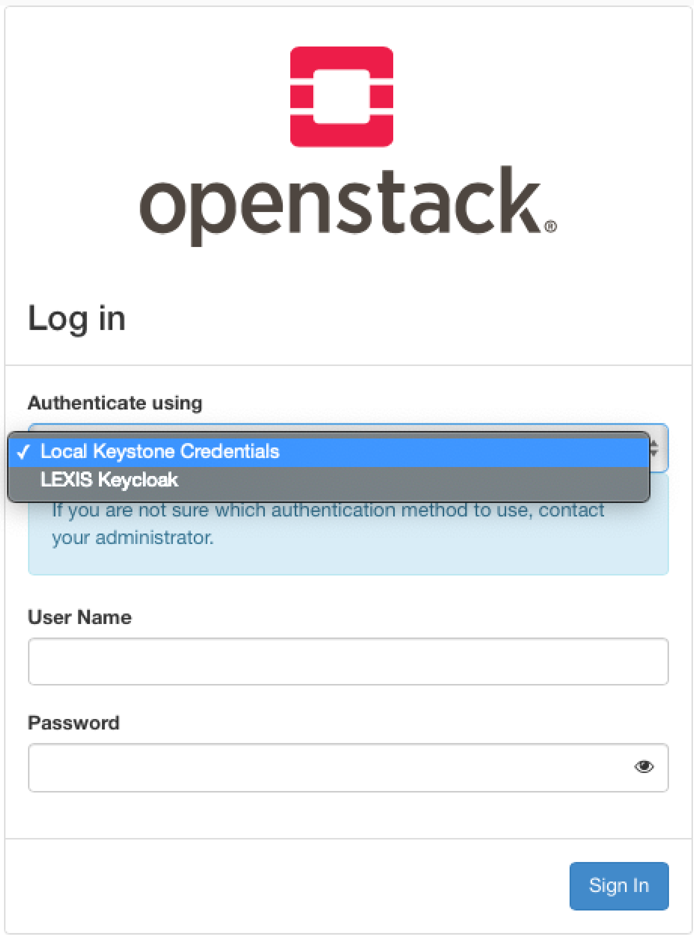

Figure 3: Current test environment allowing to evaluate OpenStack Keystone and Keycloak services

2.4.2 LRZ

The current identity and access management at LRZ (LRZ-SIM) is based on several LDAP services which centralize the information about users and their permissions as well as their organizational data and affiliation. The interface for manual changes to the LDAP data consists of an Identity Management Portal and some standalone webforms, e.g. for requesting HPC grants. Yet, there are many automated processes for user provisioning and deprovisioning, based on MicroFocus Identity Manager with "IDM drivers", custom programs and scripts (Perl, Ruby, and Python) running as daemons or cronjobs, CSV interfaces, and an increasing range of webservices, esp. REST-APIs (Ruby Sinatra on Apache Passenger).

Groups of users are organized in LRZ projects. A user may have multiple usernames (variations of an initial username) in order to access project-restricted resources by using the username assigned to a given project. The LRZ is moving towards a single-username paradigm, and the new system (SuperMUC-NG) already implements it. Services (e.g. Cloud or Cluster usage, Storage, Backup, etc.) are enabled for each user separately.

The management of users and the services they are allowed to use is delegated by LRZ adminstration to one or multiple so-called „Master Users“ of each LRZ project. The master user asks LRZ administration for contingents to configure the project in a particular way. (maximum number of users in the project, maximum number of users he can give access to each specific LRZ service) when the project is proposed or when the project requriements change. He informs new users about their rights and duties. A LEXIS LRZ project has been set up and a read-only export of user data from this project in the form of a LDAP proxy for LEXIS has been set up. With this and possible further mechanisms, the LEXIS AAI system will be able to map LEXIS users onto LRZ accounts or, where needed, onto a LRZ functional account for shared usage.

2.5 WIDE AREA NETWORKS

2.5.1 IT4I

As illustrated on the CESNET network map presented in [1], the IT4I center is connected to the internet through 4 x 10 Gbit/s lines provided by the national academic network CESNET which is later connected to the international GÉANT network using n x 100 Gbit/s lines. The internet connection is shared among all the mentioned systems. The internal network infrastructure will be also upgraded to full 100 Gbit/s in the near future, which will increase the available connectivity to at least 100 Gbit/s to the outside world.

2.5.2 LRZ

As shown in Figure 4, the routing from LRZ to IT4I for LEXIS activities is provided through both:

- A direct 100 Gbit/s Ethernet route to Frankfurt to reach GEANT’s Frankfurt node, routed to GEANT’s Prague node in Czech Republic
- Another 100 Gbit/s Ethernet route via Erlangen in Germany (including a 2x100 Gbit segment)

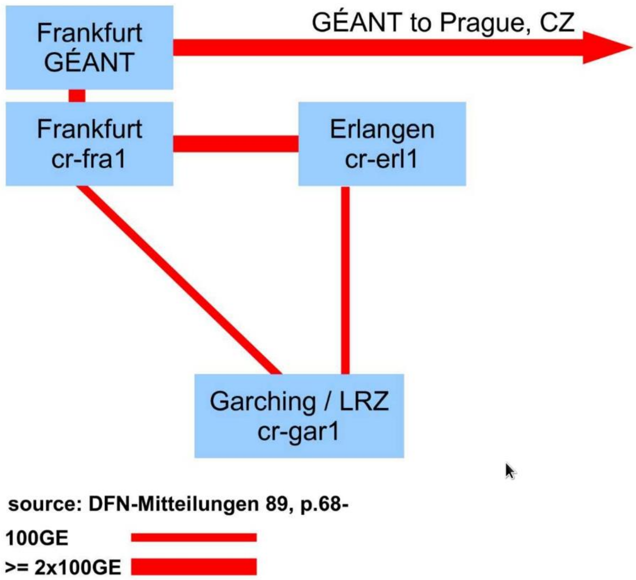

Figure 4: LRZ WAN connectivity – main routes over DFN (German Research Network) towards GÉANT and IT4I

3 KEY COMPONENTS BEING ADDED WITHIN LEXIS

This section discusses the key components being added within LEXIS:

- Federated, cross-provider AAI to provide LEXIS compute and data providers a common security model and reference implementation,
- Multi-layered data management:
  - Two high-level data management services and related APIs (one general-purpose, one for curated weather & climate data) for long- and mean-term data access and retention,
  - Temporary storage areas in both cloud and HPC environments,
  - Special Burst buffer (NVME-based) nodes for IO acceleration in the cloud part also equipped with GPUs and FPGAs for acceleration of various kinds of workloads.
- Orchestration services and related APIs (with YORC),
- LEXIS Portal,
- The attempt of using of FPGAs for both compute and I/O acceleration,
- 3D remote visualization technologies.

Figure 5 shows how the new components impact the overall LEXIS architecture.
Remark: In LRZ, an SBB infrastructure is also to be deployed, but its connectivity is still under discussion.

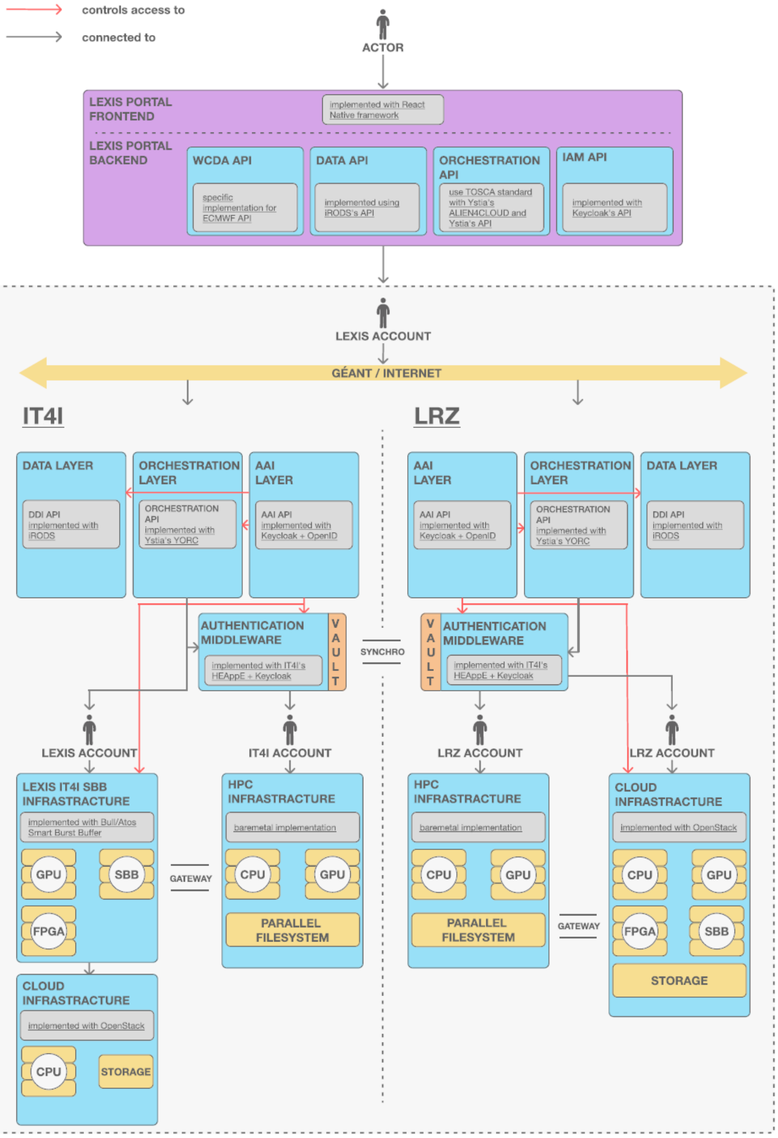

Figure 5: LEXIS architecture overview – new system components

3.1 FEDERATED IDENTITY & ACCESS MANAGEMENT (IAM) LAYER

3.1.1 Preamble

This section provides a short overview of the LEXIS "Authentication and Authorization Infrastructure" (AAI) that provides a federated IAM approach and design choices that were made since D2.1 delivered in M04. In this document it will be referenced as LEXIS IAM, which is the standard terminology matching this kind of concept and design from a technical perspective.

Security being a very large subject that has implications almost everywhere in the LEXIS overall system architecture, it implies coordination of most of the project's partners. It is meant to be a cross-WP topic. As security related activities are mostly assigned to WP4 and WP3 in LEXIS, we are only giving a short overview here.

Details about LEXIS AAI implementation choices are detailed in deliverable D4.1 [6] (due M09).

3.1.2 LEXIS IAM overview

While all the LEXIS service and data providers (IT4I, LRZ & ECMWF) have existing security environments and services in front of their HPC resources and data repositories, the LEXIS project requires to federate these under a single umbrella and make these differences invisible to the LEXIS users.

In addition, LEXIS adds cloud resources to the existing ones and this also impacts the target federated IAM model. This induces that state-of-the-art security standards must be applied, together with modern cloud computing best practises. The most visible and important of these standards are:

- OAuth2 token-based authorization framework based on:
    - Role-based access control (the RBAC approach),
    - OpenID Connect for identity management,
    - SAML (for the YORC orchestration services),
    - Single-sign-on authentication delegated to authentication and identity backend services.
- A centralized, single point of configuration API and web interface based on the Keycloak tool [5].

The goal is for LEXIS users to logon and authenticate once and be granted access to one or more LEXIS services depending on their role. Such a role will be managed by Keycloak, depending on user attributes or properties available in the user data base(s).

The LEXIS services covered with this IAM API are the following:

- Data management APIs:
    - Weather & Climate Data API (WCDA),
    - General-purpose distributed data infrastructure API (DDI API).
- APIs on Orchestration system and for acquiring processing resources (CPU/GPU/FPGA nodes, etc.):
    - YORC API,
    - Alien4Cloud (A4C) API.
- Additional lower-level system services and security layers (OpenStack, HEAppE for external end-user to internal technical account mapping, iRODS, SLURM/PBS, …), detailed in D4.1 [6].

3.2 FEDERATED DATA MANAGEMENT

3.2.1 DDI on iRODS/EUDAT basis (LEXIS Distributed Data Infrastructure)

Rationale / Requirements

In LEXIS, a distributed data infrastructure (DDI) is used to

- Provide a unified view of data from multiple participating computing/data centres,
- Simplify data staging between Cloud, HPC and storage systems of these centres,
- Automate data synchronisation and staging such that data are ready for computation (i.e. available where the computation resource is),
- Publish results from the Pilot's computations (e.g. simulations), where possible and sensible, according to "FAIR" (Findable, Accessible, Interoperable, Reusable) Research Data Management,
- Integrate LEXIS into EUDAT as a European data federation.

These requirements have served as a guideline for choosing the DDI system in LEXIS.

Choice of DDI system

While an abundance of distributed data handling systems exists, the "integrated rule-oriented data system" (iRODS[19]) developed at RENCI and other iRODS partner systems turned out a very good match to the requirements mentioned above:

- iRODS provides a unified "pseudo file system structure" and allows flexible federation of (also geographically) distributed iRODS systems ("zones"). Files are then accessible via all federated servers, where the zone is the uppermost element of the path;
- With EUDAT B2STAGE, a data-staging module is available for iRODS, which allows it to connect to GridFTP servers in the HPC ecosystems of IT4I and LRZ; from cloud systems, iRODS can be accessed via various APIs and a FUSE (file system in user space) mount module;
- iRODS allows for flexible data-synchronisation rules (i.e. selective synchronisation of certain data collections between zones, etc.) which can be executed as event handlers (e.g. on file write) in near-real time;
- With EUDAT B2HANDLE, PIDs can be automatically assigned to data products in an iRODS system, and at LRZ a component is currently developed (independent of LEXIS) to automatically acquire DOIs for data products in an iRODS system; disseminating the data via landing pages and data-index exposition can then effectively make them FAIR;
- Using iRODS is currently a prerequisite to integrate with the EU-wide research data management effort EUDAT and its EUDAT-B2SAFE system.

iRODS, as a further *pro*, has already been used at one of LEXIS' participating institutions (LRZ), and expertise is thus already brought into the project with respect to this system. A major concern *against* choosing iRODs was the performance penalties using such a complex data-handling system. iRODS was thus tested against GlusterFS, as a bare distributed file system, with respect to I/O speed. While GlusterFS I/O speed is in some cases significantly higher (factors up to 10), iRODS I/O speed proved good enough for the application in LEXIS. Direct I/O on the DDI from HPC systems does not make sense in any case (as HPC systems efficiently write on the directly-fitted file systems, cf. Sec. 2.3), and I/O for data transfer and cloud tasks can well be handled by iRODS (I/O rates around 1GBit/s can currently be reached on LRZ/IT4I virtual machines).

[19] iRODS: www.irods.org

Usage of the DDI in LEXIS' Work Packages - APIs

The DDI interfaces with:

- The burst buffer/ data-node systems implemented by the co-design work packages of LEXIS (WP2-4), as these data nodes must have access to the DDI – they might be used to directly accelerate I/O onto/from the DDI,
- WP4's orchestrator solution, which has to make sure that input/output data for a workflow are available on the appropriate systems,
- WP5-7's workflows and the systems used in them (cloud, HPC, burst buffers),
- WP8's LEXIS portal with its user interfaces for data management.

iRODS provides a wealth of APIs and client libraries with binding to various programming languages. In the context of LEXIS, WP3 will provide the following APIs for the DDI (and possibly more):

- Data Search API – for WP8,
- Data Staging API – for WP4 and possibly WP8,
- Data Harmonisation/Fusion API – here, usage is still somewhat unclear,
- Monitoring API – for WP8 and billing/quota monitoring to Pilots,
- and probably the general, "official" iRODS REST API based on Java for various purposes.

WP5-7 can – besides usage via APIs (via WP4/8 orchestration and portal solutions or direct usage) – use the DDI by direct interfacing with the iRODS system (e.g. FUSE mounts). Data publication functionalities (FAIR Research Data Management) may require further APIs to be set up or at least an integration with the WP8 web portal, which might offer publication workflows.

Identification of key technologies

Besides iRODS itself and the APIs/client libraries offered in the iRODS ecosystem (goRODS, python bindings, iRODS REST API, icommands, irodsFs, etc.), a major part of key technologies for the DDI is coming from EUDAT. EUDAT aims at offering modules for a unified EU-wide Data Management, including components for FAIR Research Data Management such as handle acquisition (B2HANDLE) and data search (B2FIND). B2SAFE, a central EUDAT component, is, in turn, an add-on to iRODS, which can be used in LEXIS to federate the iRODS DDI with other data infrastructures and to define PID acquisition as well as data-replication rules. Finally, the EUDAT module B2STAGE is crucial for LEXIS as it interfaces iRODS with GridFTP, enabling data staging to LEXIS HPC systems.

Figure 6 next page shows the near-real-time inner workings of an EUDAT-enabled iRODS-based DDI when data are staged onto it using EUDAT B2STAGE ("iRODS-DSI").

The key technologies/ systems mentioned have been tested in M01-M09 of the LEXIS project and – with possible minor exceptions – proved to be useful and appropriate for the project.

Figure 6 next page shows data staging between iRODS and GridFTP with EUDAT's B2STAGE (iRODS-DSI module), and iRODS-replication workflows (EUDAT B2SAFE) as well as PID-acquisition workflows (EUDAT B2HANDLE) happening in the background in an EUDAT-enabled DDI.

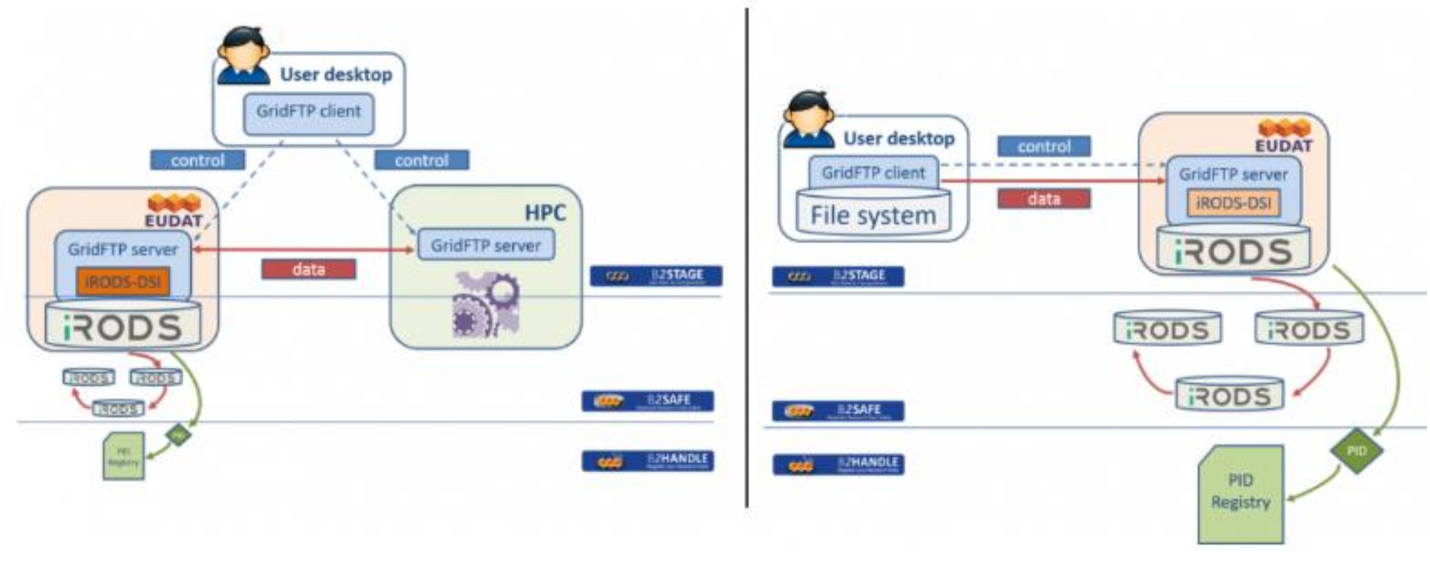

Figure 6: Data staging between iRODS and GridFTP systems (HPC: left panel, user system: right panel). Source: EUDAT

Integration with LEXIS AAI

A major technological challenge has been the integration of iRODS with the Keycloak-based LEXIS AAI. The iRODS-OpenID plugin is not fully mature. Owing to this and to a partial lack of integration of such an authentication mechanism into iRODS client libraries/API toolkits, significant development/bug-fixing efforts have been necessary to bring an iRODS-OpenID solution into reach.

Meanwhile, the efforts mentioned have succeeded such that an OpenID-based authentication in iRODS seems possible. This would integrate iRODS/EUDAT-B2SAFE fully with the LEXIS AAI, which (as opposed to possible work-around solutions) is our preferred strategy.

3.2.2 WCDA (Weather & Climate Data API)

Besides the DDI system, special storage libraries for certain use cases or institutes are being developed and deployed in LEXIS. Most importantly, a Weather and Climate Data API (WCDA) will be developed by WP7 in collaboration with WP3. This API will enable users to store, retrieve, access and manage curated weather & climate data. The WCDA is not a general-purpose storage; it only deals with curated weather & climate data (for which metadata and content have been checked and accepted).

The WCDA is designed as a distributed system, facilitating data transfer between components of weather & climate workflows. The WCDA storage backend uses a custom high-performance database, an instance of which will be deployed at the LEXIS computing sites each, besides one instance at EMCWF. As a backend file system, systems implementing the POSIX API are suitable (at LRZ, for example the „LEXIS experimental storage system“, cf. D3.1 [5]). The repositories at LRZ and IT4I can, with respect to the instance at ECMWF, be configured as „caches“ or as partial or full data mirrors in order to „bring the data to the computing power. Currently, it is being discussed how far the WCDA performs data transfer between these sites on its own, or if this would rather be a task for the orchestrator (acting on the WCDA or moving data itself). As a result, however, it is clear that the the WCDA must mostly serve data to applications irrespective of where they are run.

Details about the WCDA concept, related API as well as technical architecture figures are available in deliverable D7.1 [8].

3.2.3 Site-local Burst buffers

3 possible approaches

The main difficulty with the Burst Buffers in the LEXIS project is that they will be located in the cloud/open part even though they are also required to accelerate workflows HPC executed in the HPC/closed part.

In terms of system integration of the Burst Buffers in the LEXIS architecture, 3 use cases have been considered since the release of [1] in M04 (end April 2019):

- Cloud-to-HPC accelerated transfers (based on an NFS gateway),
- Scratch volume for cloud jobs,
- Burst buffer accelerator for cloud jobs.

Approach #1: Cloud-to-HPC accelerated transfers (based on an NFS gateway)

Bull/Atos is currently adding the support of NFS (in addition to Lustre) with their BB. The technical proposal is the following:

- Data on the Lustre- and GPFS-based HPC file systems at LRZ/IT4I is exported via NFS (on Ethernet) to the BB,
- The BB in turn exports the data to the cloud compute node via NFS too.

Approach #2: Scratch volume for cloud jobs

In this approach, the BB is planned to be used as a bunch of flash. A storage volume is exported, thanks to NVMe-oF (NVMe over Fabric) protocol, to a node hosting a GIS DB server. This protocol allows to fulfil the strong storage access constraints (i.e. low latency, full POSIX-IO semantic, etc.) needed to the backend storage of a GIS DB server.

However, the NVMe-oF protocol needs a network supporting RDMA technology and an exported volume can only be bound to one physical node of the Cloud cluster infrastructure. Technical checks are currently on-going about the ability for IT4i and LRZ network equipments to support RDMA over Converged Ethernet (RoCE).

If not, less efficient alternatives can be investigated:

- Replacement of RoCE with internet Wide Area RDMA (iWARP) Protocol,
- Replacement of NVMe-oF with Internet Small Computer System Interface (iSCSI) protocol.

Approach #3: BB accelerator for cloud jobs

A last possible use case is the use of Burst Buffers to accelerate data access to the CEPH storage. However, CEPH storage is currently used as block devices managed by the Cinder OpenStack module. Bull Burst Buffers can probably not be used in such a configuration. Indeed, a Burst Buffer is not exposed as a new block device but is meant to intercept glibc IO calls (e.g. read, write, etc.).

Currently, there is no known LEXIS application planning to access the CEPH storage directly as a POSIX and/or Object file storage. Indeed, LEXIS applications will mainly need some fast scratch storage spaces. This demand would be covered by the previous use case, making approach #3 less relevant for the moment.

Placement of Burst Buffer(s) in the IT/Network infrastructure at IT4I/LRZ

At IT4I, the HPC systems and LEXIS Cloud infrastructure are operated in different network segments and are only sharing the boundary network infrastructure. The Burst Buffer has connectivity to the Cloud infrastructure; access to the internal high-speed network fabric of the HPC clusters is only possible via a dedicated gateway which possibly restricts outwards traffic.

At LRZ, the situation is relatively similar to IT4I; while the Compute Cloud has a relatively open connectivity to e.g. the internet, the HPC clusters (SuperMUC-NG / SuperMUC / Linux Cluster) and especially their compute nodes and storage systems have strictly separated network segments with limited external connectivity. The Burst Buffer, situated in the cloud segment, will however have access to a GPFS system (Data Science Store) exported via NFS and attached to LRZ HPC systems. This motivates the development of NFS-as-a-backend functionality for the burst buffer.

3.3 USE OF FPGAS (AND GPUS)

3.3.1 Characteristics of FPGAs

FPGAs can perform some computing functions very efficiently and very fast but there are limitations to what they can do efficiently. The sweet spot for FPGA processing is on-the-fly or in-line processing, functions that can be performed on a stream of data without knowing what the rest of the data looks like. This includes Fourier transforms, matrix filters and more generally anything that can be done on a small section of data or spanned into a global pipeline scheme that can be very useful to handle high data rates.

FPGAs are weak when they must deal with a large amount of strongly coupled data, since FPGAs generally have little memory compared to CPUs, they don't have banks of registers to use for addressing, and they can't juggle data around. The implication is that running algorithms that require random access to the data are not practical; also functions where iterative operations must be performed again are not ideal for FPGA's.

3.3.2 General usage considerations for LEXIS: compression, encryption, conversion, SMP jobs

The Burst Buffers mentioned in the previous section are – as a LEXIS speciality – equipped with GPU and FPGA cards, allowing for accelerated data processing tasks (e.g. on-the-fly data compression) to be directly executed on them. While GPU usage will be used where appropriate, employing standard toolkits and acceleration techniques, the FPGA is a more ambitious and innovative point in LEXIS as described in detail below.

According to the previous analysis, FPGAs used for data conversion is wide-spread and generally unseen by the user but when they are brought to the forefront of processing, they can offload processing power form the CPU and can enable extremely high bandwidths to be managed. For this purpose, we intend to introduce this acceleration technology to both data nodes and compute nodes which respective tasks of handling workflows and processing suit well with the FPGA criterions.

For the investigation of this acceleration technology, we have selected the state-of-the-art 520N PCIe board featuring an Intel Stratix 10 FPGA[20], along with four banks of DDR4 external memory, four network ports to enable dramatic FPGA-to-FPGA scaling independent of the PCIe bus, plus support for an array of serial I/O protocols operating up at 10/25/40/100GbE. In addition, the support of Intel FPGA OpenCL Software Development Kit provides the following functionalities for software development:

- OpenCL support for software-orientated customers,
- Abstraction for faster development,
- Push-button flow for FPGA executable, driver, and API,
- Add optimized HDL IP cores to OpenCL designs as libraries,
- Therefore, both traditional HDL and higher abstraction C, C++ and OpenCL-based tool flows will be supported.

The proposal here is to accelerate encryption, encoding and/or compression for data transfer. However, BB nodes could also fit small SMP jobs (shared memory parallelism, not using Infiniband), like Tsunawi or possibly post-processing/Paraview/viz jobs. LEXIS could then take advantage of data proximity on burst buffer to efficiently process them with an FPGA accelerator.

[20] Intel Stratix 10 FPGA: www.bittware.com/fpga/520n/

On top of this, we could expose a few YORC orchestrator scenario(s), allowing to annotate a job or task as requesting FPGA acceleration (visible nice-to-have).

It could at least be interesting to perform following tasks with an FPGA accelerator on a data node:

- Encryption encoding,
- Compression encoding,
- Data format conversion,
- Data restructuration (temporal realignment),
- Weather data curation (i.e. WCDA conversion)(?)

At least the first two points are the classic benchmark for HLS tools, so the main point would be to bring the generic benchmark code to a "production quality" implementation that can be used on field.

The question was important because LEXIS's primary focus is not speedup. LEXIS should rather be understood as a demonstrator of complex HPC and large data-centric workflows with orchestration.

3.4 ORCHESTRATION

3.4.1 Workflows to orchestrate

This section describes the workflows that each pilot plans to configure in the YORC. The key point here is to automate the provisioning of each part of the application workflow on the adequate parts of the LEXIS infrastructure, taking several hard or soft constraints at the same time (preferred platform, the data proximity or available resources) while staying under full control of the LEXIS federated AAI, wherever data is to be accessed, and software is to be executed.

The two following Figure 7 and Figure 8 are representations of the WP6 and WP7 pilots' target application workflows, which were extensively described in [1]. Similar workflow representation for WP5 was already proposed in [1].

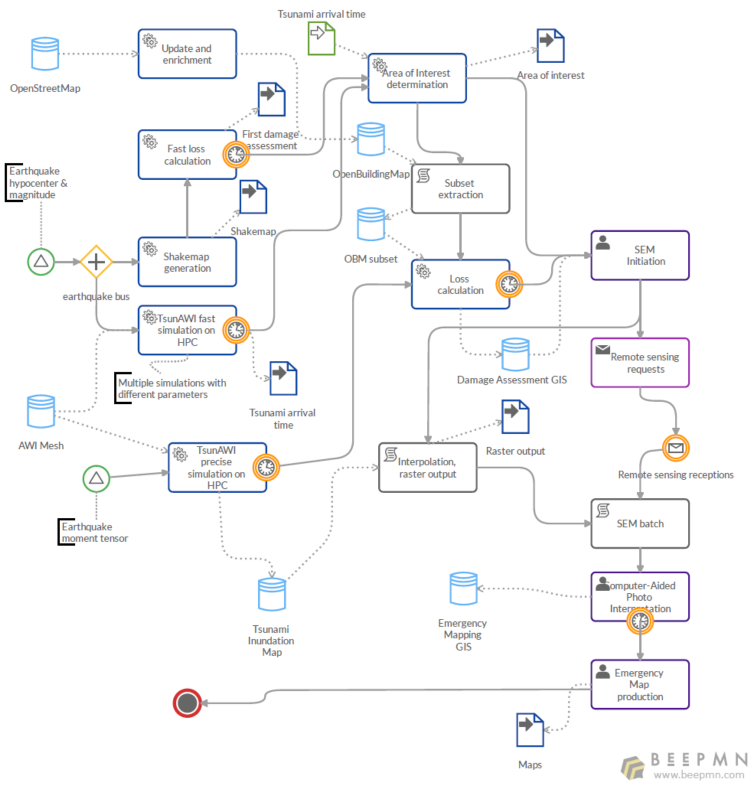

Figure 7: WP6 (Tsunami/Earthquake) pilot workflow chart

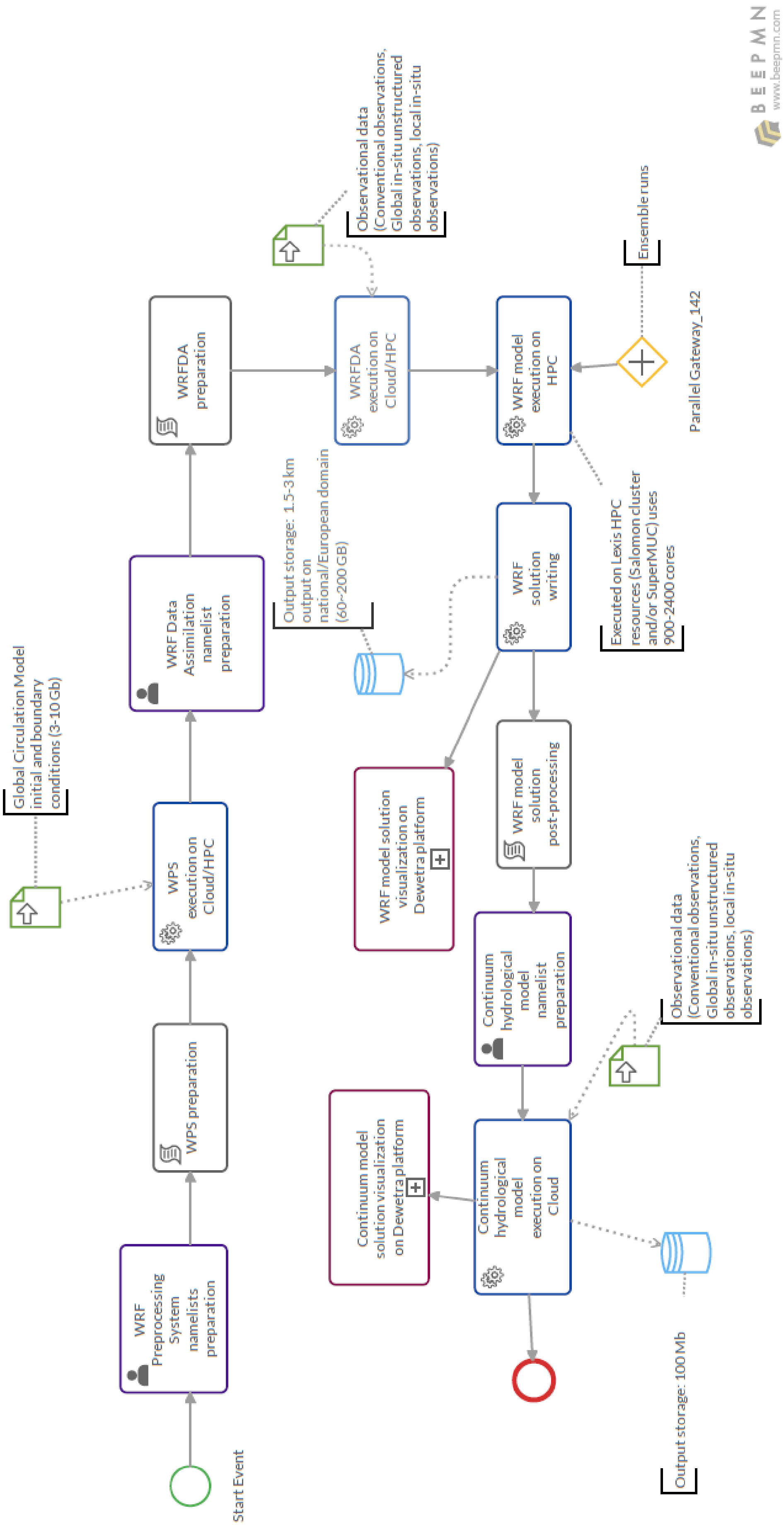

Figure 8: WP7 (Weather/Climate) pilot workflow chart

3.4.2 Deployment of YORC

Even though the final LEXIS cloud hardware is not available yet, an initial deployment of YORC has been decided to not delay system and application integration testing.

In IT4I

3 Virtual Machines were provided on IT4I setup. YORC front-end (Alien4Cloud) and the Orchestrator (YORC) were installed one of these Virtual Machines, the two other virtual machines being used to deploy sample applications through YORC (these Virtual machines being added in a Hosts Pool).

This setup was then configured to delegate its user authentication to the local Keycloak server, with first attempts to test IAM concepts such as realms, roles and authentication backends (see Figure 9).

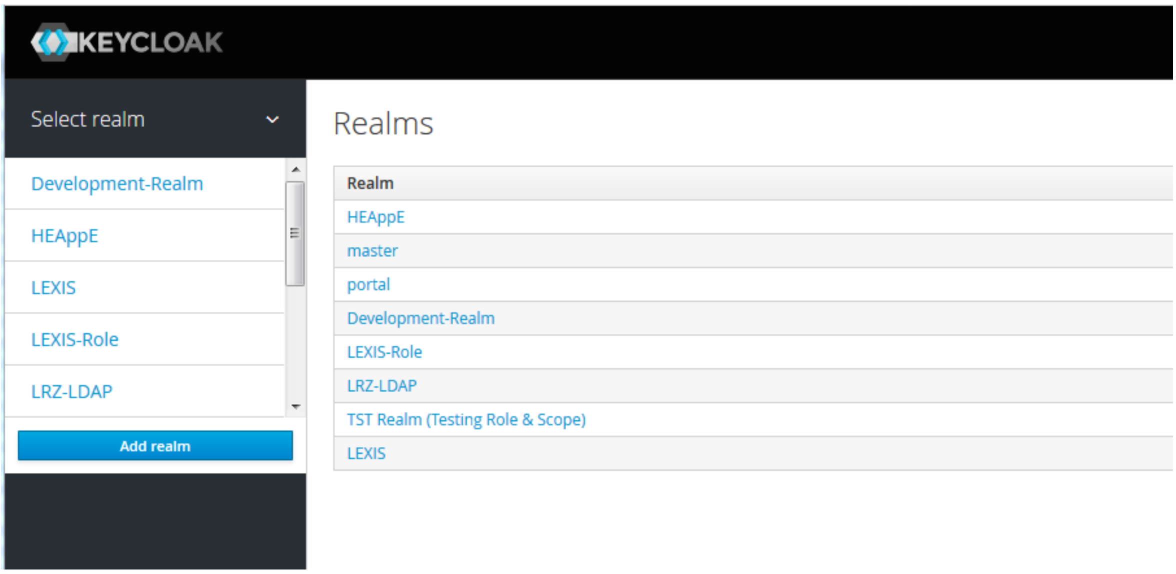

Figure 9: View of the test Keycloak IAM console used to secure YORC requests

In LRZ

3 Virtual Machines were allocated on the OpenStack infrastructure, and the installation procedure was done similarly as in IT4I. This setup should be reinstalled soon to have YORC allocate resources directly from OpenStack on-demand (instead of having VMs statically created in a Hosts Pool like today).

Work with partners

Upon request from WP8 team to WP4 team, knowledge transfer has started, covering Alien4Cloud UI, which roles have defined (Administrator, Architect, Application Manager), actions that can be performed by each role, and a quick description of the internal work performed by the front-end to provide a deployable bundle to the orchestrator. Sample scripts using the Alien4Cloud REST API were also provided (see Figure 10) with a focus on calls that could be used by the LEXIS Portal in a first step:

- How to import new types and application templates in Alien4Cloud,
- How to create an application from this application template,
- How to deploy the application on a given location,
- How to execute a workflow on this application.

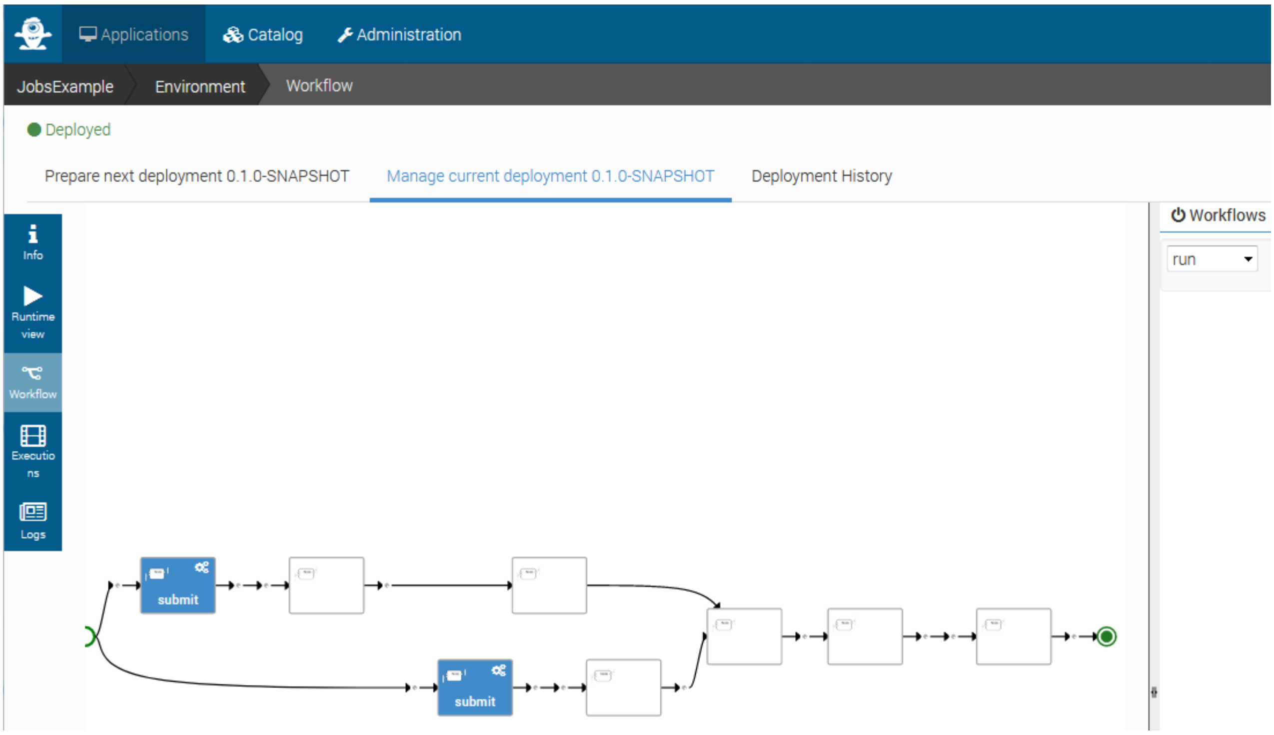

Figure 10: View a workflow deployment in YORC/Alien4Cloud UI (also exists as an API)

Work in progress

As a first step, a lightweight integration of YORC with HEAppE will be performed by implementing a TOSCA component using HEAppE REST API to create/submit/wait for the end/delete jobs. A document should provide a description of this component for TOSCA beginners, and how to to use it from either the Alien4Cloud UI or the REST API.

3.5 LEXIS USER EXPERIENCE AND PORTAL LAYER

The LEXIS project is developing a portal which will facilitate access to the LEXIS system for third parties with specific consideration being given to SME inclusion. The portal will provide a self-service sign-up interface where interested parties can learn about the capabilities of the LEXIS platform, can request access to resources and can interact with the platform in an easy to use manner.

3.5.1 Portal functionalities and roadmap

LEXIS WP8 will develop 4 releases throughout the life of the project with increasing capabilities. The high-level capabilities for each release are shown in Table 7 below. Note that these capabilities are indicative and not precisely defined - which specific capabilities manifest in which release depends on progress through project execution.

Release	Delivery	Capabilities
R1	M09	• Log in to Portal • Obtain information on data sets available within Portal • See Terms and Conditions of use
R2	M15	• Basic support for creating a job via the Portal • Basic support for accounting and billing - see bills for usage over some period
R3	M24	• Estimate cost of job given some basic parameters relating to job • Understanding of job status and some monitoring integrations • Greater accounting and billing support showing real time and historical usage
R4	M30	• Download results via Portal

Table 7: Capabilities of different releases of the LEXIS Portal

3.5.2 Overview of the portal software architecture

The architecture of LEXIS Portal R1 is shown in Figure 11 hereafter. The objective of the first release was primarily to address issues with integration of the different functionalities and technologies, as well as get the respective LEXIS team members working together. As such, the capabilities of LEXIS Portal R1 are modest. The basic capabilities are as follows:

- Log in to the Portal using Keycloak as an OpenID provider,
- View the list of users registered in the system,
- View the list of organizations in the system,
- Create a new organization if the user is not in any specific organization,
- List metadata relating to available datasets stored on an iRODS instance with different levels of granularity.

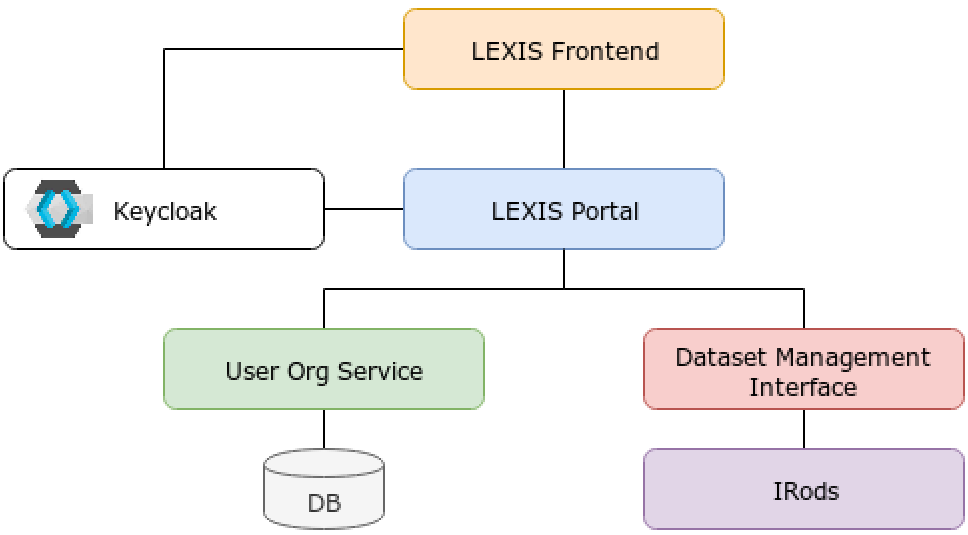

Figure 11: LEXIS portal software architecture for Portal Release R1

No critical issues arose during the R1 development cycle; standard issues associated with using toolsets provided and determining a good working modus were dealt with. As of this writing (Aug 16, 2019) the components are close to being integrated and no specific risks to delivery for M09 are anticipated. The most complex issues which arose within this development cycle are the following:

- Understanding the AAI model to be used in LEXIS and relating it to the WP8 development,
- Developing reasonable solution for authentication and co-ordination between React Front End and Portal,
- Working with iRODS using a non-standard library (the *Gorods* library) - iRODS has a very large C/C++ toolchain and it was awkward to work with using more modern technologies.

Planning for the Portal R2 is ongoing. The current envisaged system architecture is shown in Figure 12 hereafter. Note that this figure emphasizes integration with AAA services - specific linkages to e.g. DDI, Orchestration services (Alien4Cloud) and billing are not explicitly shown.

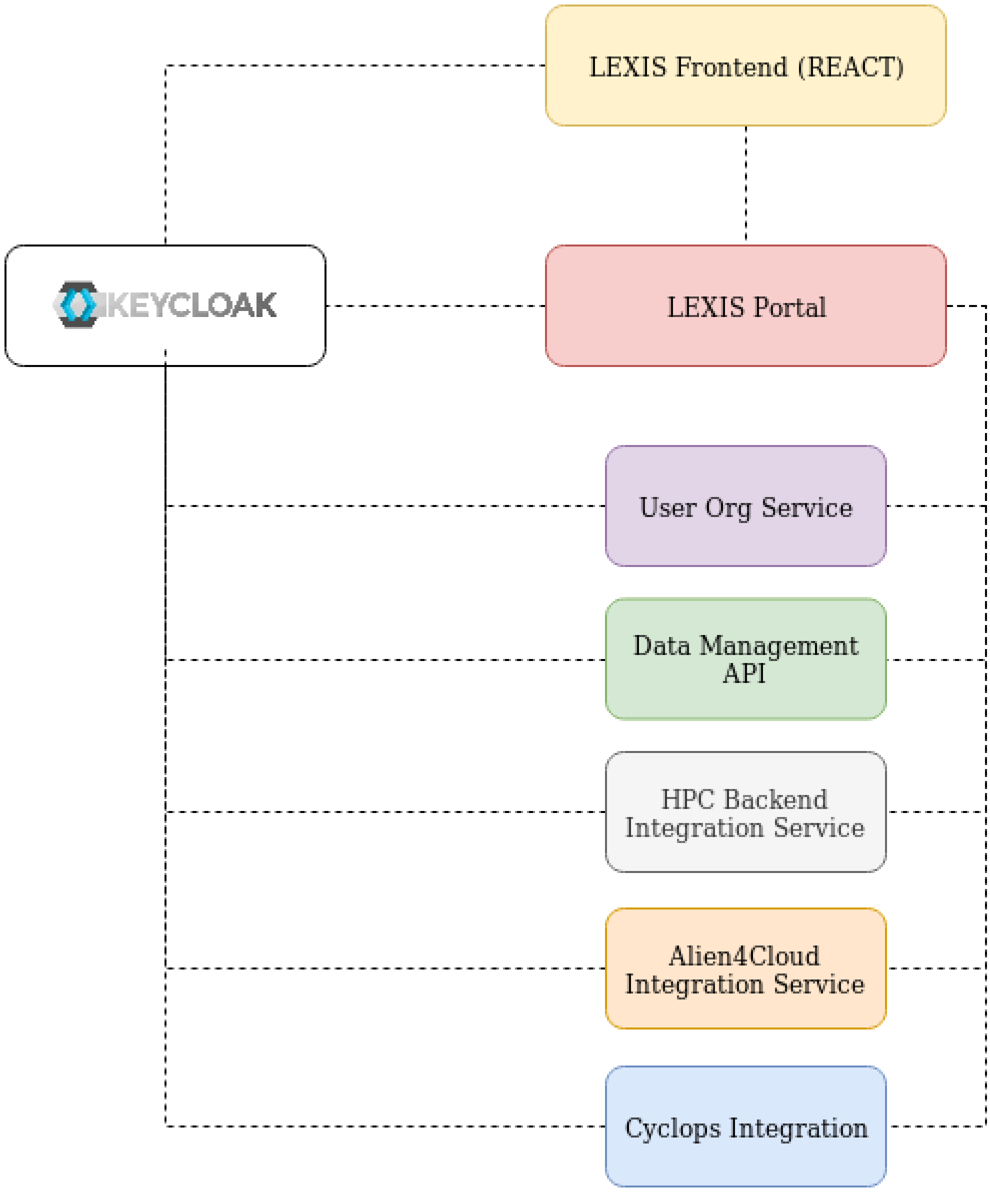

Figure 12: Envisioned Architecture of LEXIS Portal Release R2

The second release will have to interact with many more components, which will significantly increase the complexity of the development; the specific interfaces and capabilities of all these components are still under study and the project is working on devising how they can be integrated meaningfully to produce expected functionality in Release R2.

Release R3 will be planned at the end of the R2 cycle to incorporate learning from this cycle; similarly, R4 will be developed at the end of the R3 cycle.

3.5.3 Implementation

The Portal implementation is still evolving and as with many iterative development processes, all the key implementation issues are only resolved at final delivery.

For R1 and R2, the solution is based on the following:

- Components specific to the Portal:
  - Rich React-based frontend,
  - Golang backend modules serving Swagger based REST APIs.
- Components with which integration is required:
  - Keycloak - using the *Golang Gocloak* library and a Keycloak *Javascript* library,
  - Alien4Cloud - using the A4C REST API,
  - Cyclops Accounting and Billing Framework - using the Cyclops REST APIs,
  - iRODS using the *Gorods* library.

The components are packaged as docker containers; best practices are used for creating the containers - multistage docker builds, building from scratch/alpine images where appropriate etc. This results in a set of docker containers that are typically small (~some 10's of MB) which are stored on the IT4I *Gitlab* container registry.

Docker-compose files are used to bring up the containers, managing dependencies between containers as necessary and providing the service with the appropriate configuration data. *Toml* syntax is predominantly used for configuration information, being easier to read/manage than *yaml* and *json* files and having better typing supports than *ini* file formats. A `config.toml` file is bind mounted into the container at launch to provide the service with appropriate configuration data.

3.6 REMOTE VISUALIZATION TECHNOLOGIES

Remote visualization technologies are one key enabler of scientific applications usage in the cloud because the not only avoid possibly slow and unsafe transfers of large data back to the client, but they also reproduce the acceleration of a high-end 3D workstation on a regular, unaccelerated client host.

Keyboard and mouse events are sent in real time to the graphics servers that render the 3D interactive scene, and the servers sends a compressed 2D video stream version of the 3D scene.

The concurrent sessions can be safely dispatched and controlled using the same job scheduler as the one used for the compute jobs.

3.6.1 IT4I

A 3D remote visualization solution is being deployed in IT4I on a new HPC environment. It is based on the Bull/Atos software XRV (eXtreme factory Remote Visualizer)[21].

In the IT4I context, remove visualization sessions are using the same job scheduler as the HPC jobs, which identifies GPU-accelerated nodes with enough remaining slots before opening the session. As for compute jobs, if no idle graphical slot has been found, the visualization job will wait for one to be released.

[21] XRV: https://atos.net/fr/produits/calcul-haute-performance-hpc/bull-extreme-factory#1499763136525-9fc69891-38d1

This client-server technology enables 3D remoting in various possible ways:

- **Seamless mode:** if a single graphical application is to be used (e.g. Paraview), this mode allows to only open the application's windows in the client's Window manager (they appear to be local) and significant bandwidth is saved.
- **Desktop mode:** the full remote desktop is exported (like most other remote display technologies).
- **HTML5:** the graphics are displayed in a new tab in the client's web browser (no XRV client software needed in this case).

At the time of writing, it is actively investigated in how far this IT4I service can be used in the LEXIS project.

3.6.2 LRZ

LRZ has quite a strong tradition in offering a remote-visualisation service for its HPC environment. While this might be used in LEXIS as well, LRZ has worked and is working on a special remote-visualisation solution for LEXIS. This solution enables remote visualization of X11 applications running on the LRZ cloud service based on OpenStack[22], over a web service. The technologies used include x11vnc and web sockets.

Technical environment

The base image is GPU-Ubuntu-18.04-LTS-bionic-(cuda/docker) with additional graphical software:

- **Paraview** (open source 3D scientific viewing application matching LEXIS pilots' requirements),
- Remote graphics middleware: **x11vnc, python3-websockify, novnc, xvfb,**
- Basic graphical tools for testing purposes: **xinit, xterm.**

With this base image used, the virtual machines are automatically deployed to those parts of the LRZ Compute-Cloud infrastructure containing NVidia V100 GPUs.

Websockify

Since HTTP allows multiple requests on the same connection, websockify needs to close the connection to x11vnc after the request is finished; otherwise x11vnc gets confused by the trailing data.

Configuration of the image

Network/Security groups

There are different possibilities to connect the user with the remote desktop:

1. SSH tunnel from web portal to cloud, then web sockets to localhost at web portal,
2. TCP connection from web portal to cloud, then web sockets to TCP at web portal,
3. Web socket from client to cloud.

Solution 3 is implemented here for simplicity, but the other solutions can have security or other advantages; they can be implemented when needed.

Apart from the "default" security group's built-in rules, security group issues arose and ports 5900-5902 and 6080 had to be opened.

Instance

The instance should be assigned a floating IP (in our case $IP=138.246.232.76) and a keypair to access it. The instance flavor lrz.small (1VCPU) is sufficient for an initial test.

[22] Remote visualization on the LRZ cloud service based on OpenStack: https://cc.lrz.de

Starting the Remote Visualisation

The steps to run an actual remote visualization scheme on the present LRZ test system are relatively straightforward:

1. SSH connection to access the visualization server image (able to identify a LEXIS user),
2. Mount the required dataset location, taking care of preserving user ID,
3. Create a random x11vnc password using the x11vnc -storepasswd command,
4. Make sure certificates are available in /etc/ssl,
5. Run x11vnc using “/usr/local/bin/x11vnc -shared -ncache 10 -forever -create --usepw”,
6. export the DISPLAY environment variable,
7. Run Paraview with the required CLI parameters to load the dataset and start the job,
8. Run websockify.

A specialised link then makes it possible to access the remote visualization session (as presented in Figure 13):

```
https://dhcp-<ip of the virtual machine>.dynamic.eduroam.mwn.de:6080/vnc_auto.html?password=<password>
```

Future Work

The system described above will be further tested and integrated into the LEXIS infrastructure. Network configuration or other details of our design choices might be somewhat revised in this process.

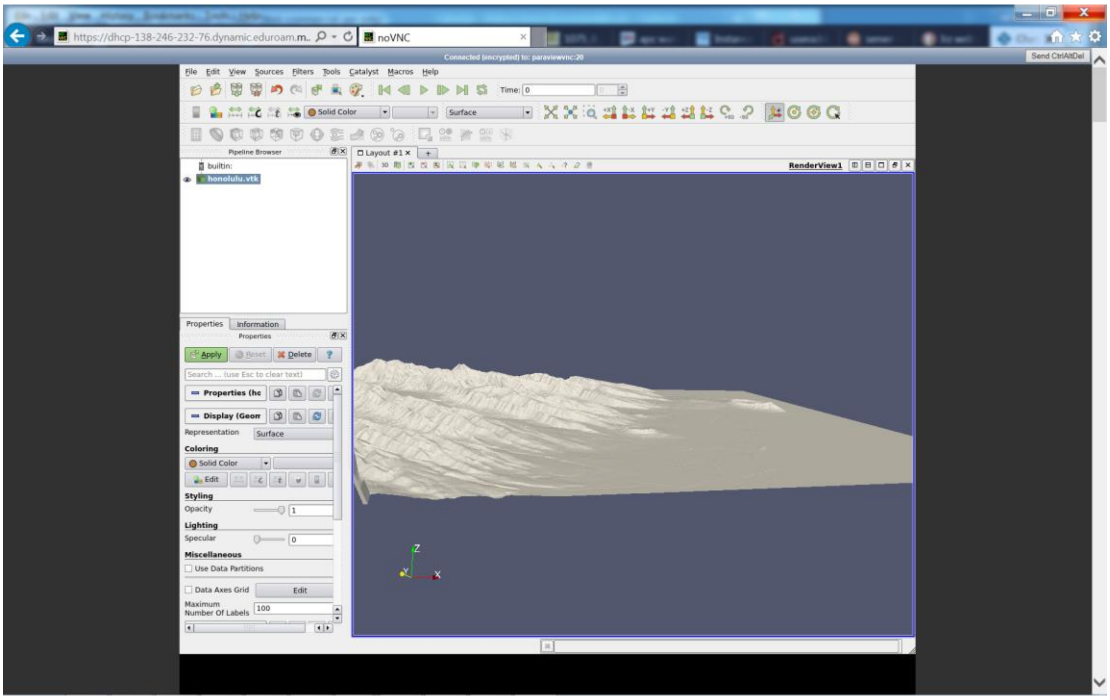

Figure 13: Remote visualization session example applied to Paraview

4 SUMMARY

After 9 months of active codesign (involving all partners: application use cases stakeholders, HPC service providers and technology vendors), the high-level system architecture assessments and plans from D2.1 [1] have been extended, and partially put to the test of practice with a number of prototype systems deployed.

The result of this codesign phase is a coherent and secured LEXIS federated system architecture spanning 2 HPC & data service providers (IT4I in Czech Republic and LRZ in Germany) and one Weather & Climate data provider (ECMWF in UK). All substantial technical choices at hardware, platform and middleware levels have been made and LEXIS teams have started to work on deploying them or testing them.

In the present D2.2 deliverable, we have described:

- Key technologies for LEXIS that have been or are being deployed on the existing infrastructure in Section 2,
- New key technologies that have been identified as required for LEXIS and that will be deployed soon in Section 3.

These results are globally encouraging, technologically promising and on-schedule which confirms the LEXIS project's good health.

As the new technology obtained agreement among all partners and are already being implemented, the next steps are the following:

- Adoption by the other technical work packages of these technologies in the design of the main LEXIS platform building blocks such as:
    - LEXIS AAI services and API
    - LEXIS data management APIs
        - General purpose one based on iRODS
        - WCDA for Weather & Climate curated Data management API
    - LEXIS orchestration services
    - LEXIS portal
- Modelize the pilot use cases as workflows in the YORC orchestrator (involving all the pilot teams and the WP4/YORC experts)

A REFERENCES

[1] L. Deliverable, *D2.1 Requirements Definition and Architecture Design,* April 2019.

[2] [Online]. Available: https://doku.lrz.de/display/PUBLIC/SuperMUC-NG.

[3] [Online]. Available: https://www.lrz.de/services/compute/supermuc/systemdescription/.

[4] [Online]. Available: https://www.lrz.de/services/compute/linux-cluster/overview/.

[5] L. Deliverable, *D3.1 Local Storage Solutions Report,* 2019.

[6] L. Deliverable, *D4.1 Analysis of Mechanism for Securing Federate Infrastructure.*

[7] „Keycloak Open Source Identity and Access Management (www.keycloak.org)“.

[8] L. Deliverable, *7.1 Architectural Requirements and System Design for Interchange of Weather & Climate Model Output between HPC and Cloud Environments.*

[9] John Chevers, „Géant - Linking Prace,“ 2016.

B APPENDIX: SAMPLE WIREFRAME DESIGNS OF LEXIS PORTAL

This appendix shows some sample wireframes for the LEXIS portal, giving preliminary indications on what it will look like.

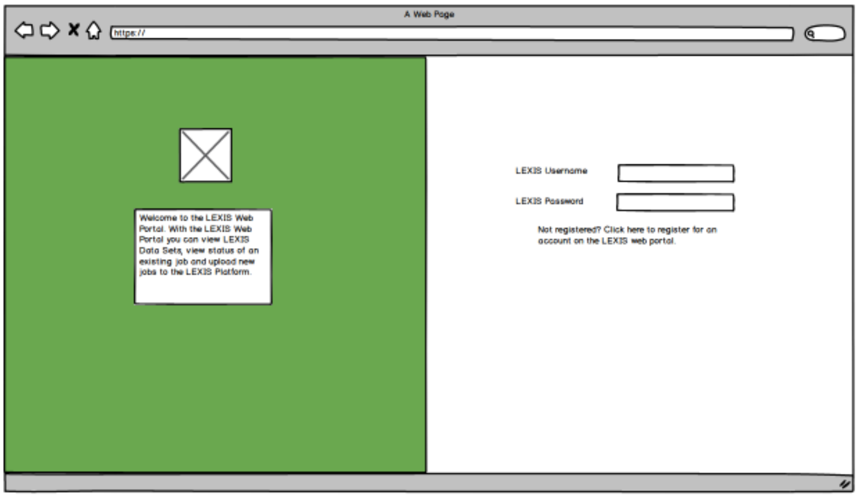

Figure 14: LEXIS portal login page

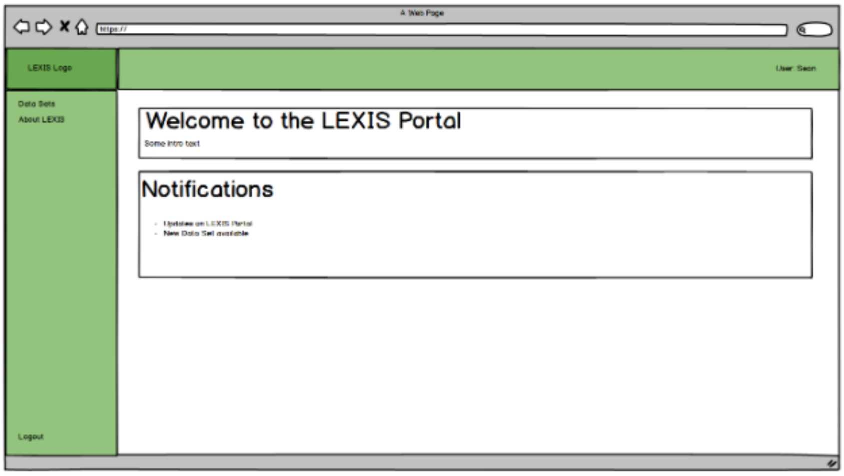

Figure 15: LEXIS portal entry page

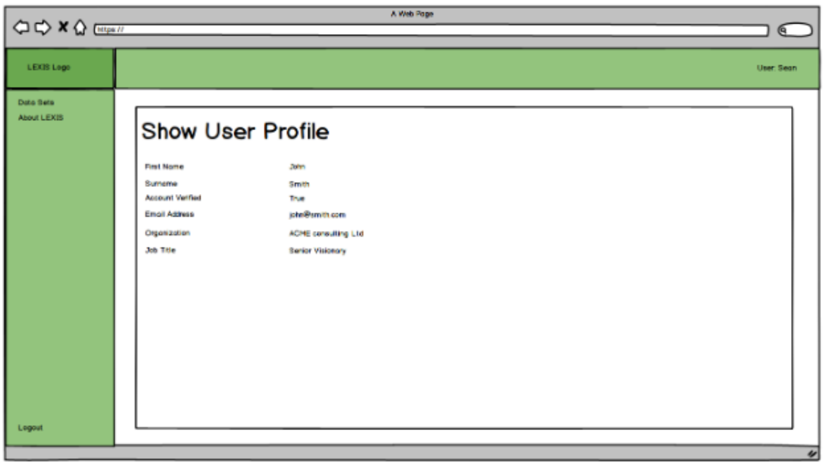

Figure 16: LEXIS portal user profile view

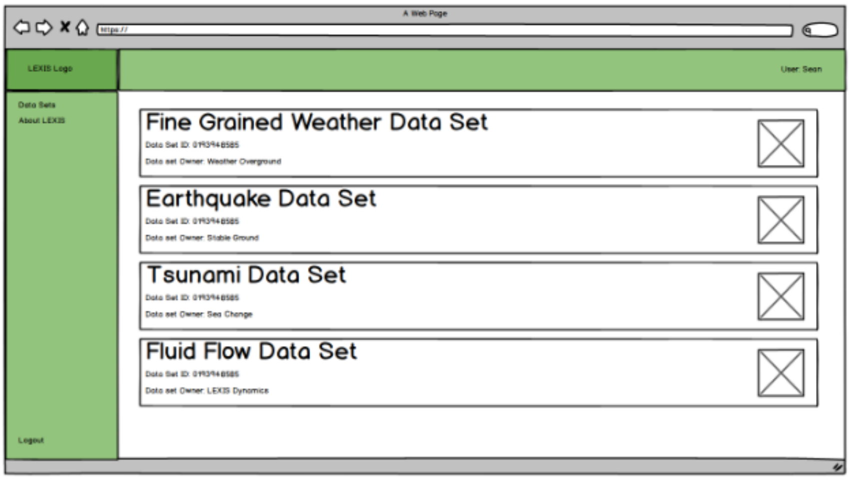

Figure 17: LEXIS available datasets view page

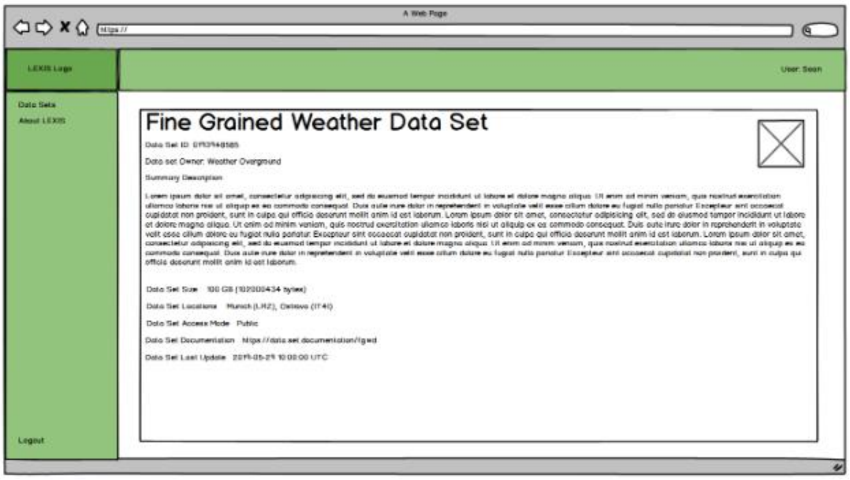

Figure 18: LEXIS specific dataset view page