E-Infrastructures
H2020- INFRAEDI-2018-2020

INFRAEDI-01-2018: Pan-European High-Performance Computing infrastructure and services (PRACE)

PRACE-6IP

PRACE Sixth Implementation Phase Project

Grant Agreement Number: INFRAEDI-823767

D6.2

Final report on collaboration with other e-infrastructures
Final

Version:	1.0
Author(s):	Cédric Jourdain (CINES), Cristiano Padrin (CINECA), Miroslav Puskaric (HLRS), Miroslaw Kupczyk (PSNC), Javier Bartolome (BSC), Filip Stanek (IT4I), Borislav Pavlov (NCSA), Kiss Zoltán (KIFÜ), Pedro Alberto (UC-LCA), Ezhilmathi Krishnasamy (UL)
Date:	23.09.2021

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: INFRAEDI-823767**	
	Project Title: Final report on collaboration with other e-infrastructures	
	Project Web Site: https://www.prace-ri.eu/about/ip-projects/	
	Deliverable ID: < **D6.22**>	
	Deliverable Nature: <Report>	
	Dissemination Level: PU	**Contractual Date of Delivery:** 30 / 09 / 2021
		Actual Date of Delivery: 30 / 09 / 2021
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level are indicated as follows: **PU** – Public, **CO** – Confidential, only for members of the consortium (including the Commission Services) **CL** – Classified, as referred to in Commission Decision 2005/444/EC.

Document Control Sheet

Document	**Title: Final report on collaboration with other e-infrastructures**	
	ID: D6.2	
	Version: <1.0>	**Status:** ***Final***
	Available at: https://www.prace-ri.eu/about/ip-projects/	
	Software Tool: Microsoft Word 2016	
	File(s): PRACE-6IP-D6.2	
Authorship	**Written by:**	Cédric Jourdain (CINES), Cristiano Padrin (CINECA), Miroslav Puskaric (HLRS), Miroslaw Kupczyk (PSNC), Javier Bartolome (BSC), Filip Stanek (IT4I), Nevena Ilieva-Litova (NCSA), Borislav Pavlov (NCSA), Kiss Zoltán (KIFÜ), Pedro Alberto (UC-LCA), Ezhilmathi Krishnasamy (UL)
	Contributors:	Oliver Keunen (LIH), Petr Nazarov (LIH)
	Reviewed by:	Enver Ozdemir, UHeM; Veronica Teodor, JUELICH & Dirk Brömmel, JUELICH
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	31/05/2021	Draft	Initial document structure
0.2	15/07/2021	Draft	Executive summary, Introduction
0.3	19/08/2021	Draft	Add SKAO, Fenix, CoEs, ELI and LIH paragraphs collected

0.4	02/09/21	Draft	Add demonstrator and CERN part, internal correction and formatting
0.5	06/09/21	Draft	Add Puhuri and PaRI part
0.6	17/09/21	Draft	Update after internal review
0.7	22/09/2021	Draft	Update after second internal review
1.0	23/09/2021	Final version	Ready for external review

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Collaboration, Cooperation, e-Infrastructures, data pilots, Centres of Excellence, large-scale instrument infrastructures, AAI, Data management, Security, Training, HPC resources

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] PRACE https://prace-ri.eu/about/ip-projects/#PRACE6IP
[2] EXDCI https://exdci.eu/collaboration/coe
[3] EGI https://www.egi.eu/
[4] GEANT Project http://www.geant.org/
[5] Fenix-ICEI project https://cordis.europa.eu/project/id/800858
[6] EOSC-hub https://www.eosc-hub.eu/
[7] SKA https://www.skatelescope.org/
[8] CERN https://home.cern/
[9] ELI https://eli-laser.eu/
[10] NeiC PaRI https://neic.no/pari/
[11] Luxembourg Institute of health https://www.lih.lu/
[12] Puhuri project https://neic.no/puhuri/
[13] https://prace-ri.eu/cern-skao-geant-and-prace-to-collaborate-on-high-performance-computing/
[14] High luminosity LHC https://home.cern/science/accelerators/high-luminosity-lhc
[15] HiDALGO https://hidalgo-sproject.eu/
[16] EXCELLERAT https://www.excellerat.eu/
[17] OpenID Connect https://openid.net/connect/
[18] UEABS https://repository.prace-ri.eu/git/UEABS/ueabs
[19] HEP Benchmark Suite https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite/-/tree/qa-v2.0#hep-benchmark-suite
[20] PRACE HPC systems https://prace-ri.eu/hpc-access/hpc-systems/
[21] Summer Of HPC https://summerofhpc.prace-ri.eu
[22] SD2C https://sdc2.astronomers.skatelescope.org/

[23] https://sdc2.astronomers.skatelescope.org/computational-resources/engageska-uclca
[24] https://www.skatelescope.org/wp-content/uploads/2021/07/Contact-8-SKA-Magazine-Spread-Low-Res.pdf
[25] https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html
[26] Horovod https://github.com/horovod/horovod
[27] Tensorflow https://www.tensorflow.org/
[28] Horovod Timeline https://horovod.readthedocs.io/en/stable/timeline_include.html
[29] Data and Computing Challenges at the new ELI institutions and the ELITRANS project. Presenter: T. Gaizer (ELI-ALPS), co-authors: G. Beckett, J. Chudoba et al. (2017)
[30] https://horovod.readthedocs.io/en/stable/timeline_include.html
[31] C. Thaury and F. Quéré, J. Phys. B At. Mol. Opt. Phys. 43, 213001 (2010).
[32] U. Teubner and P. Gibbon, Rev. Mod. Phys. 81, 445 (2009).
[33] S. Mondal, M. Shirozhan, N. Ahmed, M. Bocoum, F. Boehle, A. Vernier, S. Haessler, R. Lopez Martens, F. Sylla, C. Sire, F. Quéré, K. Nelissen, K. Varjú, D. Charalambidis, and S. Kahaly, J. Opt. Soc. Am. B 35, A93 (2018).
[34] and K. V. Dimitris Charalambidis, Viktor Chikán, Eric Cormier, Péter Dombi, Jozsef Fülöp, Csaba Janáky, Subhendu Kahaly, Mikhail Kalashnikov, Christos Kamperidis, Sergei Kühn, Franck Lepine, Rodrigo Lopez-Martens, Sudipta Mondal, Károly Osvay, Laszlo Ovary, in Prog. Ultrafast Intense Laser Sci. XIII (Springer International Publishing, n.d.).
[35] S. Kühn, M. Dumergue, S. Kahaly, S. Mondal, M. Füle, T. Csizmadia, B. Farkas, B. Major, Z. Várallyay, F. Calegari, M. Devetta, F. Frassetto, E. Månsson, L. Poletto, S. Stagira, C. Vozzi, M. Nisoli, P. Rudawski, S. Maclot, F. Campi, H. Wikmark, C. L. Arnold, C. M. Heyl, P. Johnsson, A. L'Huillier, R. Lopez-Martens, S. Haessler, M. Bocoum, F. Boehle, A. Vernier, G. Iaquaniello, E. Skantzakis, N. Papadakis, C. Kalpouzos, P. Tzallas, F. Lépine, D. Charalambidis, K. Varjú, K. Osvay, and G. Sansone, J. Phys. B At. Mol. Opt. Phys. 50, 132002 (2017).
[36] S. Kahaly, S. Monchocé, H. Vincenti, T. Dzelzainis, B. Dromey, M. Zepf, P. Martin, and F. Quéré, Phys. Rev. Lett. 110, 175001 (2013).
[37] H. Vincenti, S. Monchocé, S. Kahaly, G. Bonnaud, P. Martin, and F. Quéré, Nat. Commun. 5, 3403 (2014).
[38] S. Monchocé, S. Kahaly, A. Leblanc, L. Videau, P. Combis, F. Réau, D. Garzella, P. D'Oliveira, P. Martin, and F. Quéré, Phys. Rev. Lett. 112, 145008 (2014)
[39] Jie Ren, Jiaolin Luo, Ivy Peng, Kai Wu and Dong Li. Optimizing Large-Scale Plasma Simulations on Persistent Memory-based Heterogeneous Memory with Effective Data Placement Across Memory Hierarchy Proceedings of the ACM International Conference on Supercomputing ▫ 10.1145/3447818.3460356 ▫ 2021
[40] Human Brain Project https://www.humanbrainproject.eu/
[41] HBP Collaboratory https://wiki.ebrains.eu/bin/view/Main/
[42] https://www.ssc-services.de/
[43] SWAN http://www.ssc-services.de/leistungen/produkte/swan/
[44] Ckan https://ckan.org/
[45] Crawford JM: Original research in pathology: judgment, or evidence-based medicine? *Lab Invest* 2007, 87(2):104-114.

[46] Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, Vemuri VN, Parwani AV, Gibbs J, Agosto-Arroyo E *et al*: Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J Pathol* 2019, 249(3):286-294.

[47] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017, 542(7639):115-118.

[48] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019, 25(8):1301-1309.

[49] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, Schilling B, Haferkamp S, Schadendorf D, Holland-Letz T *et al*: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur J Cancer* 2019, 113:47-54.

[50] Schmauch B, Romagnoni A, Pronier E, Saillard C, Maille P, Calderaro J, Kamoun A, Sefta M, Toldo S, Zaslavskiy M *et al*: A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun* 2020, 11(1):3877.

[51] Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS: Image reconstruction by domain-transform manifold learning. Nature 2018, 555(7697):487-492.

[52] Ronneberger O, Fischer P, Brox T: U-Net: Convolutional Networks for Biomedical Image Segmentation. https://arxiv.org/abs/1505.04597, 2015

[53] Imaging platform at LIH https://imaging.lih.lu

[54] Sergeev A, Del Blaso M: Horovod: fast and easy distributed deep learning in TensorFlow. https://arxiv.org/abs/1802.05799, 2018

List of Acronyms and Abbreviations

AAI	Authentication and Authorisation Infrastructure
aisbl	Association International Sans But Lucratif (legal form of the PRACE-RI)
CoE	Center of Excellence
CERN	European Organization for Nuclear Research
CUDA	Compute Unified Device Architecture (NVIDIA)
DMZ	Demilitarized Zone (physical or logical subnetwork)
DoA	Description of Action (formerly known as DoW)
EC	European Commission
eduGAIN	International interfederation service interconnecting research and education identity federations
EGI	European Grid Infrastructure. International e-Infrastructure set up to provide advanced computing and data analytics services for research and innovation.
ELI	Extreme Light Infrastructure
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum on Research Infrastructures
EOSC	European Open Science Cloud
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network. The current EC-funded project as of 2015 is GN4.
GPU	Graphics Processing Unit
HBP	Human Brain Project
HL-LHC	High Luminosity Large Hadron Collider
HPC	High-Performance Computing; Computing at a high-performance level at any given time; often used synonym with Supercomputing
HTC	High-throughput computing
ICEI	Interactive Computing E-Infrastructure
IdP	Identity Provider
LIH	Luxemburg Institute of Health
MB	Management Board (highest decision-making body of the project)
ML	Machine Learning
MoU	Memorandum of Understanding
MPI	Message Passing Interface
MRI	Magnetic Resonance Imaging
NREN	National research and education network
OAuth2	Industry-standard protocol for authorization, developed by IETF
OIDC	OpenID Connect - a simple identity layer on top of the OAuth2
OrcID	A unique, persistent identifier free of charge to researchers
PA	Preparatory Access (to PRACE resources) NREN
PaRI	Pandemic Research Infrastructure
PTC	PRACE Training Centres
POC	Proof of concept
PB/s	Peta (= 10^{15}) Bytes (= 8 bits) per second, also PByte/s
PIC	Particle-in-Cell
PML	Perfectly Matched Layer is an artificial absorption layer for wave equations
PMO	PRACE Management Office
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
RI	Research Infrastructure
RNA	Ribonucleic acid
RPC	Resistive Plate Chamber

SAML2	Version of the SAML standard for exchanging authentication and authorization identities between security domains
SKA	Square Kilometre Array
SKAO	Square Kilometre Array Observatory, second largest intergovernmental organisation in the world dedicated to astronomy
SoHPC	Summer of HPC
SSH	Secure Shell Protocol
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UFTP	The UDP-based File Transfer Protocol is a communication protocol designed to transfer files to multiple recipients
UEABS	Unified European Applications Benchmark Suite
VM	Virtual machine
WLCG	Worldwide LHC Computing Grid

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BILKENT	Bilkent University, Turkey (3rd Party to UHeM)
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	The Computation-based Science and Technology Research Center (CaSToRC), The Cyprus Institute, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CENAERO	Centre de Recherche en Aéronautique ASBL, Belgium (3rd Party to UANTWERPEN)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3 rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3 rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd Party to PNSC)
DTU	Technical University of Denmark (3rd Party of UCPH)
EPCC	EPCC at The University of Edinburgh, UK
EUDAT	EUDAT OY
ETH Zurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland
GCS	Gauss Centre for Supercomputing e.V., Germany
GÉANT	GÉANT Vereniging
GENCI	Grand Equipement National de Calcul Intensif, France
GRNET	National Infrastructures for Research and Technology, Greece

ICREA	Catalan Institution for Research and Advanced Studies (3rd Party to BSC)
INRIA	Institut National de Recherche en Informatique et Automatique, France (3rd Party to GENCI)
IST-ID	Instituto Superior Técnico for Research and Development, Portugal (3rd Party to UC-LCA)
IT4I	Vysoka Skola Banska - Technicka Univerzita Ostrava, Czech Republic
IUCC	Machba - Inter University Computation Centre, Israel
JUELICH	Forschungszentrum Jülich GmbH, Germany
KIFÜ (NIIFI)	Governmental Information Technology Development Agency, Hungary
KTH	Royal Institute of Technology, Sweden (3rd Party to SNIC-UU)
KULEUVEN	Katholieke Universiteit Leuven, Belgium (3rd Party to UANTWERPEN)
LiU	Linkoping University, Sweden (3rd Party to SNIC-UU)
MPCDF	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3rd Party to GCS)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA2)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
SDU	University of Southern Denmark (3rd Party to UCPH)
SIGMA2	UNINETT Sigma2 AS, Norway
SNIC-UU	Uppsala Universitet, Sweden
STFC	Science and Technology Facilities Council, UK (3rd Party to UEDIN)
SURF	SURF is the collaborative organisation for ICT in Dutch education and research
TASK	Politechnika Gdańska (3rd Party to PNSC)
TU Wien	Technische Universität Wien, Austria
UANTWERPEN	Universiteit Antwerpen, Belgium
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UEDIN	The University of Edinburgh
UHeM	National HPC Center, Turkey
UIBK	Universität Innsbruck, Austria (3rd Party to TU Wien)
UiO	University of Oslo, Norway (3rd Party to SIGMA2)
UL	UNIVERZA V LJUBLJANI, Slovenia
ULIEGE	Université de Liège; Belgium (3rd Party to UANTWERPEN)
U Luxembourg	University of Luxembourg
UM	Universidade do Minho, Portugal, (3rd Party to UC-LCA)
UmU	Umea University, Sweden (3rd Party to SNIC-UU)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UnivPorto	Universidade do Porto, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
WCSS	Politechnika Wroclawska, Poland (3rd Party to PNSC)

Executive Summary

This deliverable describes existing and new collaborations between PRACE-6IP [1] members and other e-Infrastructures, as well as those with the Centres of Excellence (CoEs) [2]. These activities aim to build and develop a Europe-wide integrated e-Infrastructure for PRACE users, through the identification of new user-friendly services and the implementation of pilot projects incorporating innovative workflows and solutions.

PRACE-6IP is currently collaborating with major European e-Infrastructures (EGI [3], GÉANT [4], Fenix-ICEI [5], EOSC-hub [6]), and continues to consolidate links with CoEs and has started new pilots with large-scale instruments infrastructures (SKAO [7], CERN [8], ELI [9]). Preliminary discussions are also underway with health data infrastructures (PaRI [10], Luxembourg Institute of Health (LIH) [11]). The focus was on continuing and further developing collaboration on identity management and data services to best cover the full life cycle of data.

The collaboration with Fenix-ICEI [5] allows the provisioning of new services, such as active and archived data repositories and virtual machine services. Additionally, meetings and workshops have been held to develop a user-friendly authentication and authorisation to these services through the implementation of a PRACE Identity Provider (IdP) that would allow PRACE users to access the services and resources provided by the Fenix-ICEI infrastructure as seamlessly as possible (using a single set of credentials) and vice-versa. A pilot is being defined to evaluate the potential of this identity federation. Moreover, discussions with the Puhuri project [12] are underway to evaluate the integration of certain functionalities, such as direct access to the resource (ssh), in the future PRACE AAI solution.

The strategy to increase the offer of services focused on the data life cycle has been to establish links with large-scale instrument infrastructures which, by their nature, produce a huge amount of data to be processed, sometimes on the fly. The collaboration with CERN lets us understand the workflows already in place and to propose pilots on PRACE Tier-0 and Tier-1 systems, including the services and technologies used to cover the whole data life cycle. In the same perspective and following the Memorandum of Understanding signed between GÉANT, CERN, SKAO, and PRACE, four demonstrators have been initiated to support and assist them in the challenge of data-intensive science [13].

Finally, due to the COVID-19 pandemic and due to facing a growing demand for data processing and storage by health data infrastructures, PRACE-6IP started new collaborations with some health institutions/projects (LIH, PaRI). Their workflows requiring and drawing on current technologies such as machine learning for large-scale model training for medical imaging applications, or the need for an easy-to-use and secure (e.g. cloud-based) platform where users can upload their data are very interesting issues for the development of interoperable services for HPC users.

1 Introduction

This document summarises the activities carried out in collaboration with other e-Infrastructures during the 6th Implementation Phase of the PRACE project (May 2019 - December 2021).

The guiding principle of our work is to initiate collaborations with e-Infrastructures in the form of pilots, proof-of-concept or mini-projects in order to improve existing services and develop new services for the scientific communities. This is achieved by identifying synergies and common strategies with our collaborators' projects, implementing best practices in our collaborations by drawing inspiration from the working methods and services used by our partners who are at the cutting edge of technology in their respective fields.

An additional principle is to facilitate the interoperability of services by fostering partnerships that implement and use these services using the expertise of collaborators and their feedback.

Section 2 describes collaborations with large-scale facilities such as the Square Kilometer Array (SKA), the Extreme Light Infrastructure (ELI), or the High Luminosity Large Hadron Collider (HL-LHC [14]). Section 3 reports on effort devoted to identity management, in particular the design and development of the future PRACE AAI, a federated Authentication and Authorisation Infrastructure. Section 4 describes the activities undertaken with some CoEs (HiDALGO [15] and EXCELLERAT [16]) and the actions carried out to strengthen our links. Section 5 describes collaborations with health data infrastructures. Conclusions are provided in Section 6.

2 Collaboration with large-scale instruments

2.1 SKAO, GÉANT, CERN, and PRACE

The discussion about a Memorandum of Understanding started in the previous PRACE-5IP, and the final agreement was achieved in July 2020 with the support of PRACE -6IP PMO and WP2 of PRACE-6IP. The kick-off meeting was held at the end of September 2020, approving the main activities of the collaboration. Four tasks have been identified:

- Training & Centre of Expertise
- Authenticated Workflow Demonstrator
- Benchmarking Proof-of-Concept
- Data Access Demonstrator

WP6 was mainly involved in the last two activities, which are described below.

Regular meetings are organised every three months in order to share achieved progress and experienced issues. Unfortunately, the collaboration will end in December 2021, when the regular activities of PRACE-6IP will come to their natural end.

2.1.1 Data Access Demonstrator

The activity aims to explore user-friendly solutions for data access services. CERN and SKAO introduced their workflows, remarking the needs of a data transfer service that should run between the CERN infrastructure or SKA observatories and the selected HPC site. The challenge of the activity is to transfer data in the petabyte range per day.

The first meetings were dedicated to identify which PRACE-6IP partner could be interested to collaborate, and how to adapt the needs to the services provided by PRACE. The individual steps for the data access demonstrator were:

- Identify HPC sites and identify the people who will be directly involved
- Define the authentication and authorisation that will be used
- Define the rules and attributes of the demilitarised zone (DMZ)
- Run scenarios (the first one would be to mirror the file system directly on a different centre (CINECA/CERN))

CINECA was identified as partnering Tier-0 site, and a meeting between CERN and CINECA was held in July 2021 in order to agree on the DMZ configuration. The authentication and authorisation mechanisms will be based on the OIDC solution [17], explored in the PRACE AAI activity.

The tests related to the authentication of CERN's people are expected in September 2021, and the configuration of DMZ is expected in October 2021.

2.1.2 Benchmarking Demonstrator

Following the kick-off meeting, an HPC benchmarking group dedicated to this demonstrator was created to develop a common benchmark suite with CERN and SKAO. This working group aims to implement a benchmark suite that will help organisations to measure and compare the performance of different types of computing resources for astronomy and particle physics data analysis workflows.

CERN and SKAO already have benchmarks that have different specificities in terms of data flow and processing, from those currently in the Unified European Application Benchmark Suite (UEABS), the PRACE benchmark suite [18]. Indeed, CERN's workloads continue to focus on the HTC until the end of the Large Hadron Era (2026), and multi-node work in the software (distributed parallelism) is not planned. The idea for CERN is to move towards benchmarks more representative of their current workflow which would be based on WLCG workloads [19]. SKAO, on the other hand, is planning both HTC and HPC workloads. Therefore, the interest is twofold, for CERN and SKAO to take advantage of PRACE's expertise and resources to homogenise and benchmark on Tier-0 systems and for PRACE to expand, diversify, and modernise its benchmark suite.

Following a presentation of the UEABS benchmark suite mechanisms, the prerequisites for the new benchmark suite were defined:

- Cover all architectures and accelerators where use is intended
- Targeting, in particular, the PRACE systems [20] and also other systems adapted in the context of our collaboration
- Reproducible, verifiable accounting of execution and results (including tests)
- Reporting of results to configurable locations
- Run as a cluster job
- Inclusion of system/environment metadata
- Inclusion of energy cost and efficiency measurement

The working group involved two students for two months, via the PRACE Summer of HPC (SoHPC) programme [21], which offers summer internships in HPC centres across Europe. We have therefore defined a work plan for the student who will be supported by HPC experts from SURF Amsterdam and CERN:

1. Introduction: Familiarisation & running with current tools from CERN and PRACE

2. Investigation: Examine missing coverage from current tools. Includes familiarisation with containerisation, accounting (metadata, energy use included), collection & reporting
3. Proposal: Based on work in (1) and (2), propose an integration route for the final deliverable
4. Integration, testing, report

Preparatory accesses (PA) of type C for 6 months on two Tier-0 systems were considered to provide special assistance from PRACE experts to support the set-up and optimisation but due to lack of funding at the target HPC sites we had to go for PA type B which are intended for code development and optimisation by the user. Students attended a week of training from 30 June to 3 July and student access to the systems was opened on 5 July 2021.

2.2 SKA Data Challenge

The SKA Science Data Challenge 2 [22] was set up by the SKAO Consortium to analyse a simulated datacube of 1 TB in size in order to find and characterise the neutral hydrogen content of galaxies across a sky area of 20 square degrees. Following a collaboration with SKA-Engage, the Portuguese branch of SKAO, UC-LCA was listed as one of the computing centres in Europe which would host members of research groups willing to participate in this data challenge [23].

Forty teams registered to take part in the data challenge with a total of 276 participants representing 80 institutes and 23 countries. Two teams, one from the UK and one from Japan chose UC-LCA and its cluster Navigator to perform the data challenge. The data cube was transferred to Navigator, a project allocation and an account were created, so that things were ready for the teams to complete the challenge, needing only to contact UC-LCA to create the user accounts and learn about the local software environment (module environment, queues, and software installed including Anaconda). In June 2021, we learned that these teams had withdrawn from the challenge. This happened to other teams as well.

UC-LCA has regularly participated in video conference meetings with the SKAO managers for this data challenge. It seems that many SKAO users are not familiar with the usage of HPC resources, so that this must be a concern of SKAO in view of a future data challenge that will occur in 2022. The idea is to create communication channels between computer centres and prospective users so they can be acquainted with what is needed to use an HPC facility in order to process SKA data. The Data Challenge 2 and the computing resources available were described in the July issue of Contact, the internal news magazine of SKAO [24].

2.3 Collaboration with CERN on data certification

The experiments running on the LHC at CERN work in challenging conditions and produce an enormous amount of data to be analysed. The smooth operation of the detectors is vital for the data quality. The harsh radiation environment puts detector modules under stress and in rare cases may lead to malfunctions which in turn can spoil the data quality. The detector performance monitoring as well as the certification of the experimental data as usable for physics analysis is a crucial task to ensure the quality of all physics results. Currently, the detector monitoring and data certification is done by human experts and is extremely expensive in terms of human resources and required expertise. Thus, introducing ML techniques in LHC data analysis and data certification turned out to be an interesting case. We implemented Machine Learning tools to help and facilitate the monitoring and data certification for one of the LHC CMS experiment's subsystem (Resistive Plate Chambers subsystem). PRACE was

identified as a partner able to provide HPC infrastructure in order to preform machine learning on the monitoring data.

2.3.1 Use case

Two important parameters of the Resistive Plate Chambers (RPCs) are used to monitor the detector stability. One of the parameters is the detector current and the other is the detector counting rate. For each of the parameters separate tool has been developed and used. All data has been re-analysed and recorded in a dedicated database to estimate individual detector rates. An additional complication is that the detector-related parameters (rate, current, etc.) and the environment parameters (LHC luminosity, temperature, pressure, etc.) are in separate databases and are not synchronised. No external access is granted to the database and thus as a first step special scripts are run locally at dedicated machines at CERN to extract the data in a convenient format.

2.3.2 Results

Preparatory Access has been granted at CINECA's Marconi100 system, an accelerated cluster based on IBM power9 architecture and Volta NVIDIA GPUs. The tools have been ported and tested on the HPC infrastructure and show promising results. The data corresponding to the full 2017 and 2018 data taking periods has been downloaded and stored in a convenient for ML format and the performance of two different ML tools has been evaluated.

The first tool uses linear regression [25] to correlate detector rate with the accelerator luminosity. The synchronisation of the data (rate and luminosity) and ML is performed on PRACE infrastructure (Marconi100 supercomputer at CINECA) and good scaling is achieved. In Figure 1 we report the performance in terms of speedup, defined as the execution speed, normalised to the execution speed on one node.

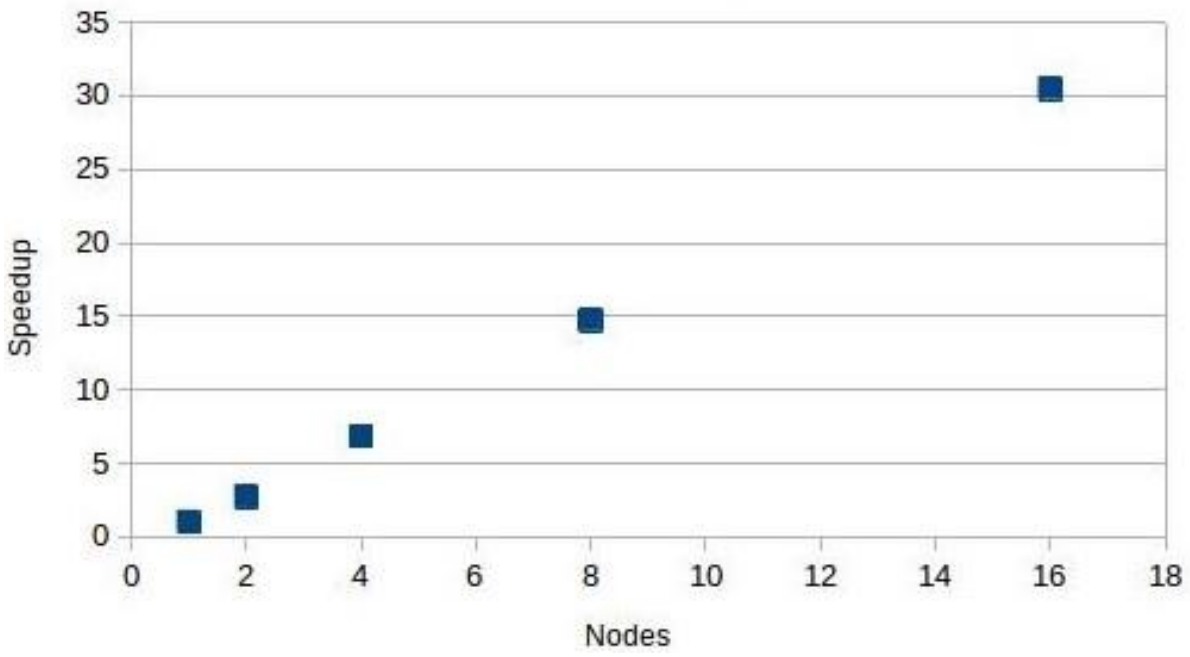

Figure 1: Performance result (speedup) obtained for RPC rate tool.

The second tool is used to monitor the RPC currents. It uses Horovod [26] and TensorFlow [27] in order to be able to use GPUs on several nodes simultaneously. The tool uses an auto-encoder with approximately 700 inputs (and the same number of outputs). The auto-encoder predicts the current of the detectors and comparing the prediction with the measurements helps detector

monitoring, by spotting hardware problems. The code is ported and run on Marconi100 at CINECA. The performance is analysed using Horovod timeline file to find possible bottlenecks [28]. The results are shown in Figure 2. A synchronisation of the machine learning parameters between the nodes is performed at each epoch, this step takes a fixed time and affects the scaling. By increasing the amount of data to be ingested by the algorithm (per epoch), the duration of this synchronisation becomes negligible and thus, for larger data samples, the scaling improves. Horovod's timeline plots confirm this hypothesis.

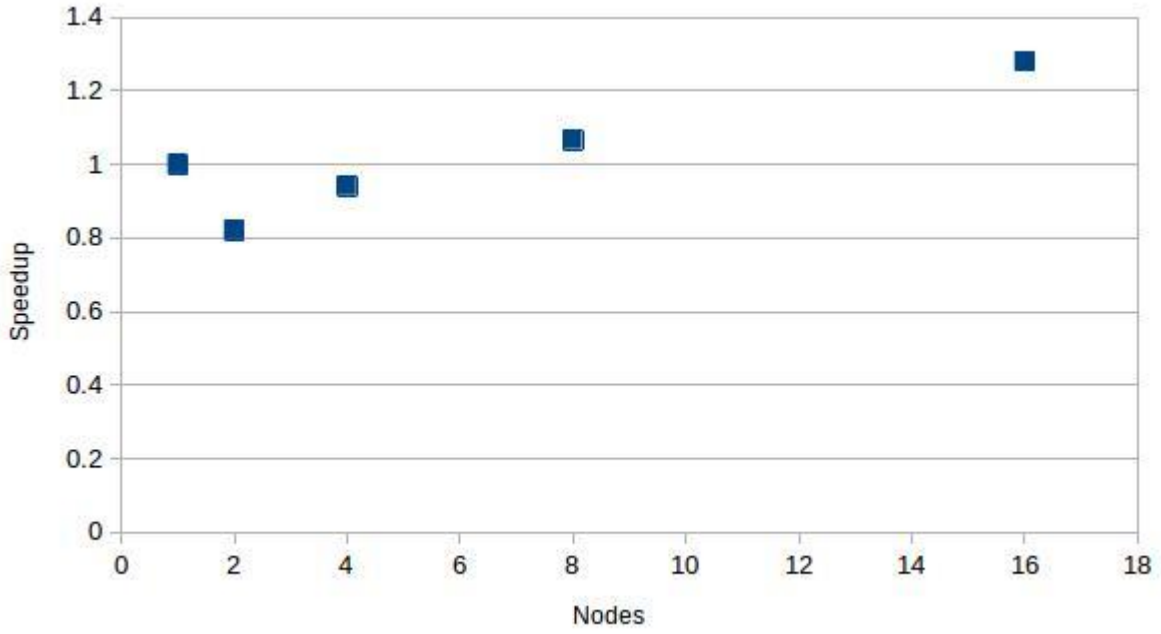

Figure 2: Performance result (speedup) obtained for RPC currents tool.

2.4 ELI

The Extreme Light Infrastructure (ELI) is one of the most ambitious European Strategy Forum on Research Infrastructures (ESFRI) projects built in Europe, accommodating some of the most intense lasers in the world. After the efforts of building the infrastructure, ELI European Research Infrastructure Consortium (ERIC) was established to operate facilities and is determined to work together with other European e-Infrastructures to solve its data storage and computational challenges.

These facilities generate an extreme amount of data (10 PB/year [28]) needed to be pre- and post-processed to gain insights on the reactions of material caused by different lasers requiring large data and computing capacities not available onsite. PRACE was identified as a computational partner, while GÉANT being a network partner offering high bandwidth connection between ELI and PRACE sites.

2.4.1 Use case

Surface high-order harmonics generation (SHHG) by the interaction of high-intensity short-pulse laser with plasma mirrors leads to attosecond sources that provide a unique approach to utilize relativistic control on charge dynamics for attoscience [31][32]. SPA has several unique features and are now being implemented in the two beamlines named SHHG-SYLOS & SHHG-HF at the ELI-ALPS facility [32][33][34][35]. The lasers have unique specifications and would allow different possibilities of interaction that did not exist before. It is essential to identify conditions of interaction that are optimal for the generation of coherent XUV radiation. Multiple

characteristics affect the interaction: irradiation conditions, target geometry and plasma mirror properties, etc. [35][36][37][38]. This necessitates virtual experiments by conducting sets of a fully relativistic 3D particle-in-cell (PIC) simulations pertinent to the situation.

2.4.2 Simulation Methodology

The WarpX PIC code was proposed to be used in this pilot. WarpX is an advanced electromagnetic PIC code that supports dynamic mesh refinement (AMReX), boosted-frame simulation and perfectly-matched layers (PML). It is an extremely scalable, highly optimised code that is under very active development and is available both for CPU and GPU platforms. It features hybrid OpenMP/MPI parallelisation combined with advanced vectorisation techniques and load-balancing capabilities.

2.4.3 Outcome

Under this particular project, we would use such numerical PIC simulation tools to study this unique parameter space and aim to identify the optimal conditions. Within its scope, we would specifically test the plasma parameters and shape for optimal SHHG generation with optimal spectral, temporal properties and maximum energy. The results are expected to lead to publications and at the same time very important for the forthcoming SPA-beamlines at ELI-ALPS, Hungary.

2.4.4 Results

Early tests showed that WarpX is highly scalable on GPU-accelerated HPC systems [39], making it ideal for using it as proof of concept (POC) between ELI and PRACE, as it requires a high number of modern GPUs to work.

Preparatory Access type B has been granted for JUWELS Booster at Forschungszentrum Jülich and the code has been successfully compiled to do the scaling tests.

The results of this POC suggests that PRACE is able to quickly provide access and test specific code required by ELI to establish further projects when ELI would serve as essential infrastructure for large scale international laser research.

3 Collaboration on identity management

For several years and during the Implementation Phase projects (4IP, 5IP and 6IP), PRACE has been closely following the various initiatives in the field of identity management. PRACE's use of existing IdP federations for authentication and authorisation services, also known as Authentication and Authorisation Infrastructure (AAI), has become one of the major focuses of the PRACE-6IP WP6. This is for good reason, since relying on these federations means facilitating the interoperability of PRACE's research infrastructure with other international or national projects, but also increasing the overall level of security offered to its users.

3.1 Fenix

The Fenix infrastructure aims to provide a federated infrastructure of data repositories, virtual machine services and scalable supercomputing systems in a well-integrated environment. An initial version of this infrastructure is currently being realised through the ICEI project (Interactive Computing E-Infrastructure), which is part of the European Human Brain Project (HBP) [40], being one of the initial prime users of this research infrastructure.

Fenix is offering the next portfolio of services:

- Interactive Computing Services: quick access to single compute servers to analyse and visualise data interactively, or to connect to running simulations, which are using the scalable compute services.
- Scalable Computing Services: massively parallel HPC systems that are suitable for highly parallel brain simulations or for high-throughput data analysis tasks.
- Virtual Machine Services: service for deploying virtual machines (VMs) in a stable and controlled environment that is, for example, suitable for deploying platform services like the HBP Collaboratory [41], image services or neuromorphic computing front-end services.
- Active Data Repositories: site-local data repositories close to computational and/or visualisation resources that are used for storing temporary replicas of data sets. In the near future they will typically be realised using parallel file systems.
- Archival Data Repositories: federated data storage, optimised for capacity, reliability and availability that is used for long-term storage of large data sets which cannot be easily regenerated. These data stores allow the sharing of data with other researchers inside and outside of HBP

One of the interesting features of Fenix for PRACE is that all these services are provided by the different Fenix partners using a federated identity management system that will permit that any user access/use resources assigned to different sites.

Collaboration of PRACE with Fenix has been performed for two aspects:

1. PRACE-ICEI coordinated calls
2. Identity management.

3.1.1 PRACE-ICEI coordinated calls

In 2018, a collaboration started between PRACE and Fenix to organise calls for resources for ICEI (Fenix Resources) in a coordinated manner, in 2019 there was an ICEI pilot call inside the 18th PRACE Tier-0 call.

Since March 2020, there have been six special PRACE-ICEI calls for Fenix resources, these calls have permitted that PRACE user community could make use of the extended services apart from the HPC (scalable computing) resources and other types of resources such as the ones described before. More generally, since the launch of these calls, we have received an average of 8.8 requests per PRACE-ICEI call, of which 77% have been accepted and the requested resources have been provided.

Most of this collaboration has been performed inside PRACE-6IP WP6, consisting of a scientific review panel in conjunction with a technical assessment to ensure that the application fits with the resources and type of services that Fenix is offering.

3.1.2 Identity management

Most of the collaboration related with Fenix and WP6 has been performed in the identity management service that Fenix has developed, studying the possibility of the use of that technology or the possible integration of PRACE authentication and authorisation process inside the Fenix identity management technology.

Though some initial discussions were performed previously, during the PRACE-6IP WP6 F2F in November 2019 in Slovenia, the technology and idea that Fenix is using for federated identity management was presented to all PRACE partners. Several discussions occurred during the

meeting and finally it was agreed to start a pilot to demonstrate the usability of that technology for PRACE infrastructure.

In the course of 2020, several discussions on how integration in terms of identity management could be performed between Fenix and PRACE.

The first proposal was to include PRACE identity management as a new IdP inside the Fenix federated identity management. This approach was not feasible in a short-term period as trust-relation between IdP and their policies inside Fenix made it difficult to extend that to all PRACE partners in time.

To continue studying this collaboration several actions were performed, a more technical workgroup team was formed inside task 6.1 and task 6.3 to study in more detail the PRACE site policies in terms of authentication and authorisation technologies, and the integration or the potential use of a technology like the one Fenix is using for identity management. This effort is coordinated by CSCS and meetings are regularly scheduled each month. A technical contact from GÉANT who is responsible in Fenix for the identity proxy service between Fenix sites is also invited to these meetings.

On the other hand, several studies were performed to document and update all PRACE user workflows and services which could benefit from a centralised federated identity management solution.

3.2 Puhuri

The Puhuri project established by NeIC is responsible to implement the Puhuri platform. The Puhuri platform is a set of integrated services for managing access to shared resources in a federated manner. This covers the resource allocation, identity management and service provision of HPC but also other compute resources (e.g. cloud services). One of its targets is the EuroHPC LUMI pre-exascale system, where resources are shared by the consortium of ten European countries and the EuroHPC Joint Undertaking (JU). Because of the natural links between the LUMI consortium, EuroHPC, and PRACE through its members and mostly the similar topics addressed in other PRACE activities (e.g. the mentioned identity management in 3.1.2) a presence of PRACE in the Puhuri Working Group project was established.

During the four meetings that already took place the main topics to address were:

- Establish a common framework of the federated AAI approach that would enable in the future interoperability between the projects
- Sharing new developed technologies like token based SSH access to HPC systems
- Coordination of EuroHPC calls organised by PRACE

The common framework of the federated AAI is based on the GÉANT technologies and services like SaToSa proxy and MyAccessID/eduGAIN to further enable interoperability with other projects like Fenix and if possible, enable a pan-European access for all major HPC activities. The SaToSa proxy is used for translation between different authentication protocols (SAML2, OpenID Connect, OAuth2, social networks like Facebook, Google, OrcID etc.) and allows the manipulation of attributes and flows. The MyAccessID identity and access management service is provided by GÉANT with the purpose of offering a common identity layer for infrastructure service domains (especially to Fenix, Puhuri, and LUMI at start) and delivers a discovery service to let users to choose their home IdP, link their different identities, and guarantee a persistent and unique identification towards the connected infrastructure service domains. The Puhuri project has its own meetings with Fenix on top of the PRACE ones to ensure the interoperability with enough detail for their own implementation.

Since the common access to HPC resources is still via SSH and also because the proposed new PRACE federated AAI is looking at OpenID Connect standards for user authentication a solution for the end users following such technology is more than desirable. The Puhuri project has developed an approach based on OpenID token authentication for SSH which would naturally fit into the new PRACE AAI and is willing to share this technology with PRACE. The aim is to evaluate this technology in the PRACE Security Forum and if possible, also to include it into the pilot of the new PRACE AAI as defined in Task 6.1.

Since one of the goals of Puhuri is to enable EuroHPC JU projects on the LUMI system, a need for technical interoperability between the PRACE Peer Review Tool (PPRT), Puhuri components like AAI Proxy for identity management and Puhuri portal for the projects management was identified. After a set of bilateral discussions between Puhuri with PRACE-6IP, PRACE-6IP with EuroHPC JU, and EuroHPC JU with PRACE BoD, an agreement was made to establish a technical working group between PRACE and Puhuri to identify how to integrate the PPRT with the Puhuri components in the LUMI use case enabling a common identity for the actors of the review and projects and to enable a machine-to-machine exchange about the proposals and projects.

4 Collaboration with the CoEs

Centres of Excellence gather professionals with similar interests to exchange ideas, best practices, and to collaborate on various research activities. The CoEs considered in this task require high-performance computing infrastructure and corresponding services for their research. Collaboration with the PRACE project, an entity with the experience and know-how in this field, could lead to several fruitful activities with potential synergy. The European Centres of Excellence EXCELLERAT [16] and HiDALGO HPC and Big Data Technologies for Global Systems [15] have been contacted and collaboration activities have been discussed. The PRACE staff have assessed a requirement form which helped WP6.3 team to better understand the CoE's goals and activities, and to proactively suggest and offer PRACE services. It was decided to actively support both CoEs to apply for core hours provided by PRACE as a part of the grant for CoEs, which ended up successfully and core hours were granted in the scope of the 21st and 22nd Call for Proposals for Project Access, as presented in Table 1 and Table 2. The latter provide computational resources to research projects in academia and industry. The CoEs have afterwards been provided with additional user and technical support.

Awards	Marconi 100	Joliot-Curie Rome	Joliot-Curie KNL	Joliot-Curie SKL	Juwels Cluster	Juwels Booster	HAWK	MareNostrum4	SuperM UC-NG	Piz Daint	Total
EXCELLERAT	875.000	350.000	240.000	170.000	45.000	55.400	550.000	240.000	86.000	250.000	2.861.400
HiDALGO	200.000	350.000					600.000	240.000	86.500	250.000	1.726.500

Table 1: Core hours provided by PRACE (Call 21)

Awards	Marconi 100	Joliot-Curie Rome	Joliot-Curie KNL	Joliot-Curie SKL	Juwels Cluster	Juwels Booster	Hawk	MareNostrum4	SuperM UC-NG	Piz Daint	Total
EXCELLERAT	275.000	150.000	93.750	87.000	40.000	38.000	280.000	100.000	65.000	510.000	1.638.750
HiDALGO	200.000	210.000					280.000	100.000	87.000	510.000	1.387.000

Table 2: Core hours provided by PRACE (Call 22)

With the computing infrastructure provided, further communication followed regarding which of the PRACE operational services could support CoEs activities. What seems to be a common challenge is the management of the large quantities of data, especially transferring them either

between the HPC sites or between the user and the HPC site. Data service is one of the core PRACE operational services where the High-Performance Computing Center Stuttgart (HLRS) acts as a service leader. Service activities include the integration of the data transfer tools and monitoring their availability within the PRACE network. The service leader has expertise with using data transfer tools GridFTP, UDP-based File Transfer Protocol (UFTP), and wrapper scripts for GridFTP commands. The expertise also includes troubleshooting data transfers by determining bottlenecks during the transfer process, tuning server network parameters and determining optimum data transfer command parameters.

4.1 EXCELLERAT

The EXCELLERAT expertise consists mainly in data-driven engineering using HPC, particularly in data management, data visualisation and data analytics. One of the partners in the project, SSC Services GmbH [42], a company based near Stuttgart, has been present in this field for a longer time. One of their products is a data exchange platform SWAN (System für den weltweiten Austausch von Nutzdaten) [43]. It is designed for a safe exchange and sharing of confidential development data with the possibility of logging all the modifications and access requests. The user interface is provided by the web browser. Since the beginning of the project, the developers have been adopting the platform for use with the HPC infrastructure and for hosting it in the HPC site environment. During the discussion, we have identified UFTP as the data transfer tool which could contribute to the SWAN platform operation. One of the advantages is the authentication procedure which requires the user to provide only an SSH public key. HLRS, as the HPC site where the SWAN platform is being hosted, already provides UFTP for transferring data to the file system attached to the HPC system. Even though it is planned to connect multiple HPC sites with this platform, namely HLRS and the Barcelona Supercomputing Center (BSC), by the time of writing this text, this has not been realised yet. This use case was also discussed with the colleagues from the SSC Services, and when this idea is realised, GridFTP will be considered. The advantage of the latter is the support of third-party transfers, that is, data transfer between two servers that are controlled by the third party. Using servers at HPC sites as endpoints brings advantages such as better network components and better uplink bandwidth to the national research and education network (NREN). When the EXCELLERAT activities reach this goal, HLRS will provide support in implementing and testing the GridFTP.

4.2 HiDALGO

At the beginning of the collaboration activities, colleagues from the HiDALGO CoE kindly provided us with the filled-out requirement form [15]. As one of the expected results of the collaboration, increased accuracy of the simulation models and hence simulation results has been mentioned. Furthermore, real-time data ingestion into the running simulations as a means of achieving this goal has been identified. Specifically, data ingestion from the remote location into the HPC environment where the simulation is running. This can be achieved by using a good performing data transfer tool. The colleagues have therefore modified the existing data management software CKAN [44] by implementing the GridFTP tool. The Keycloak was deployed for the proper management of user authentication. Besides, for computing purposes, the idea was then to use the granted access to HPC infrastructure to also take advantage of the high performing storage infrastructure and PRACE network for improving the data transfer rates between the PRACE sites.

Before proceeding with the tests, PRACE staff supported the HiDALGO applicants with applying for access grant to the Joliot Curie system at TGCC and staff of PSNC with the

valuable supporting of computational PRACE Tier-1 resources. The following case studies were considered in HIDALGO and partially the initial deployment has been done of the following cases on Tier-1 EAGLE machine:

- Migration: The main goal was to improve agent-based simulation framework in terms of accuracy, resolution, clarity, and performance, and to incorporate a range of relevant phenomena in its computations. The models were developed incorporating the precipitation data from ECMWF (European Centre for Medium-Range Weather Forecasts), and plan to exploit telecommunications data provided by the Moonstar Communications GmBH to help validate our simulations. In addition, the new techniques have been created in order to speed up the construction of our simulations (e.g., by automatically extracting and converting geographical data) and to establish better ways to visually explore the simulation output. Enabling simulations on a large scale to accurately forecast where displaced people, coming from various conflict regions of the world, will eventually arrive to find safety. The approach could assist in case of a global crisis in a number of crucial ways. Firstly, it could forecast refugee movements when a conflict erupts. Secondly, it could acquire approximate refugee population estimates in regions where existing data is missing or incomplete. Finally, it could investigate how border closures and other policy decisions are likely to affect the movements and destinations of refugees.
- Urban Pollution: The main part of the case is HPC-framework for simulating the air flow in cities by taking into account real 3D geographical information of the city, applying highly accurate computational fluid dynamics (CFD) simulation on a highly resolved mesh (1-2 m resolution at street level) and using weather forecasts and reanalysis data as boundary conditions. Emission is computed from the weakly coupled traffic simulations and general emission data of other sources. For the demonstration area, the city of Győr, Hungary, a traffic monitoring sensor network with a plate recognition camera system have been used.
- Social Networks: One of the main goals of this use case is to understand the spread of messages as well as the influence of social networks. As a second goal was to identify false/malicious messages, which intend to change the behaviour of a substantial number of users. The countermeasures were proposed to develop on the algorithmic level in order to prevent the spread of such false messages on a large scale. The developing of a highly scalable simulation framework for such stochastic processes in real-world networks, in order to be able to analyse and predict the impact of these processes on the society is in progress.

The codes were granted on PRACE Tier-1 machine 0.5 million core-hours each. The work on the simulations will be continued after this deliverable due date. PRACE staff have been asked afterwards to provide support in interpreting the achieved transfer rates and to list all possible bottlenecks in the process.

5 Collaboration with health infrastructures

5.1 Luxemburg Institute of Health

The Luxembourg Institute of Health (LIH) is a public biomedical research organisation in Luxembourg. One of the main goals of the organisation is to support the development of new diagnoses, preventive strategies, innovative therapies, and clinical applications that impact healthcare. They have many specialised departments and research platforms. Most of their research needs a computational platform (both hardware and software) to conduct and support their research activities.

In this context, University of Luxembourg will be helping to carry out LIH research project activities using the PRACE computational resource. The work will be carried out for six months in total. In addition, two research subtopics within this project will be considered for addressing the research questions.

The work has started in July 2021 and will be continued until the end of December 2021. The computational resource is granted from the CSCS, and the Piz Daint (CPU-GPU heterogeneous) system will be used for the computation. The main objective of this work is medical image analysis which focuses on a new type of machine learning and using a heterogeneous architecture, in particular GPU accelerators. This collaboration will consider two research topics: (i) digital pathology and (ii) Magnetic Resonance Imaging (MRI) reconstruction. Both of these topics will study and analyse new possibilities of machine learning and image analysis techniques. For example, splitting medical images across the compute nodes and use multiple GPU compute nodes. More detailed information about the work is explained below.

5.1.1 Digital pathology

The analysis of histopathological images is the gold standard in cancer medicine. It is mainly because of its precise diagnostics capabilities, and histopathology helps to identify tumor subtypes and defining efficient treatment. Nevertheless, it would require lots of work for a pathologist to investigate the images and slides thoroughly. In addition, these kinds of work might vary from one pathologist to another, depending on experience and subjective conclusions [45]. However, the recent developments in high-quality scanning systems, image analysis tools, and computational methods of histopathological image analysis brought a novel domain of digital histopathology [46][47][48]. The deep learning neural network (DLNN) is one of the promising tools and methodology to analyse medical images. It has been shown to match and even exceed experts' performance in specific image classification tasks [49]. This makes DLNNs extremely useful virtual assistants, especially in high throughput settings. Today, the cutting edge of the research in digital histopathology is at combining digital histopathology and high-throughput molecular data (for example – RNA sequencing data in transcriptomics). For example, the recent work of Schmauch et al. showed that a DLNN could predict transcriptomes from whole-slide histopathological images, though with rather a low accuracy [50].

The LIH recently proposed to improve this approach by predicting biological signals or activated biological processes instead of individual gene expression. The approach developed by LIH so far is presented in Figure 3(a). It includes colour stabilisation and scaling of all images to the same level of magnification, cutting slides into smaller informative pieces – tiles, and analysis of the tiles by a DLNN model. The network outputs could be used as features of the tiles and could be united into features of the slides using either clustering or a weakly supervised learning approach [48][50]. Such slide-level features, as shown the preliminary analysis, could be used to predict biological signals in the tissue with high accuracy, see Figure 3 (b).

We will implement analysis at several levels of magnification (x20, x5, and x1) and combine the resulting features as shown in Figure 3 (c). In addition, we will also investigate two DLNN models: weakly supervised predictions (supervised class labels are assigned at the level of slides, not tiles) and convolutional auto encoders that generate features without supervision.

The networks will be trained and tested on publicly available data, including healthy tissues from the GTEx database (14k images) and the TCGA cancer database (9k images). Transcriptional signals corresponding to biological processes were already extracted earlier for both these datasets in the frame of FNR project DEMICS (C17/BM/11664971).

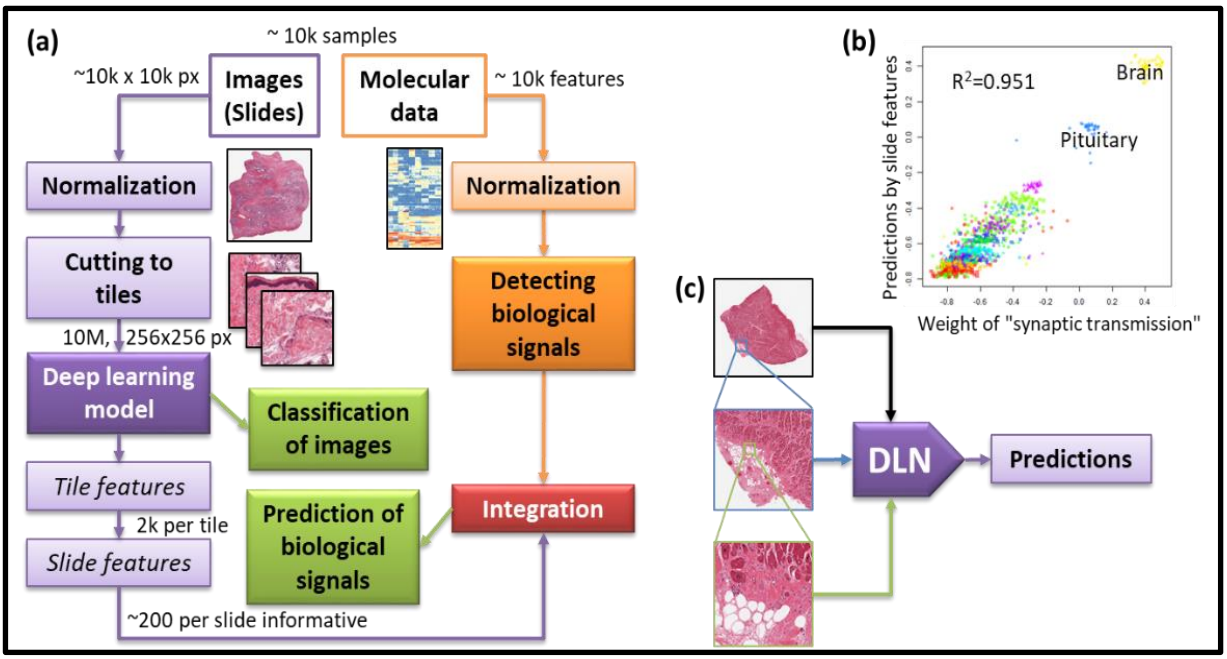

Figure 3: (a) Integration of histopathological images and molecular profiles. (b) Concordance of the image-based prediction to biological signals in normal GTEx dataset. (c) Multi-scale application of DLN.

5.1.2 MRI Reconstruction

The process of reconstructing medical images in Magnetic Resonance Imaging (MRI) is typically done with the intent to provide general-purpose images with various contrasts, suitable for a wide variety of radiological tasks. In the context of brain tumours for instance, these images can be used for diagnosis, prognosis, treatment planning, or to assess the effect of a given therapy. Post-processing techniques that may include image registration, contrast optimisation, segmentation, and quantification of various markers are often needed to complete the task, see Figure 4.

We aim to investigate if, using deep learning approaches, models to reconstruct images from sensor data and models that implement advanced image post-processing techniques can be combined to synergistically improve the performance of radiological tasks.

Models used for image reconstruction can include models that provide a direct mapping of the sensor to the image domain [51], generative adversarial networks and encoder-decoders algorithms. For a tumor segmentation task for instance, deep convolutional networks inspired by UNet [52] have proven to be highly efficient. When high-resolution images are used in a combined reconstruction and segmentation task, the models can quickly reach sizes that exceed the amount of memory available on a single GPU, rendering the training of such models lengthy and costly. The models will be trained on existing preclinical MRI data generated on the in vivo imaging platform of LIH [53].

We would like to experiment various scenarios of splitting data and models on multiple GPUs, using Horovod [54] or similar libraries, to circumvent single GPU limitations and develop state-of-art task-oriented images reconstruction models.

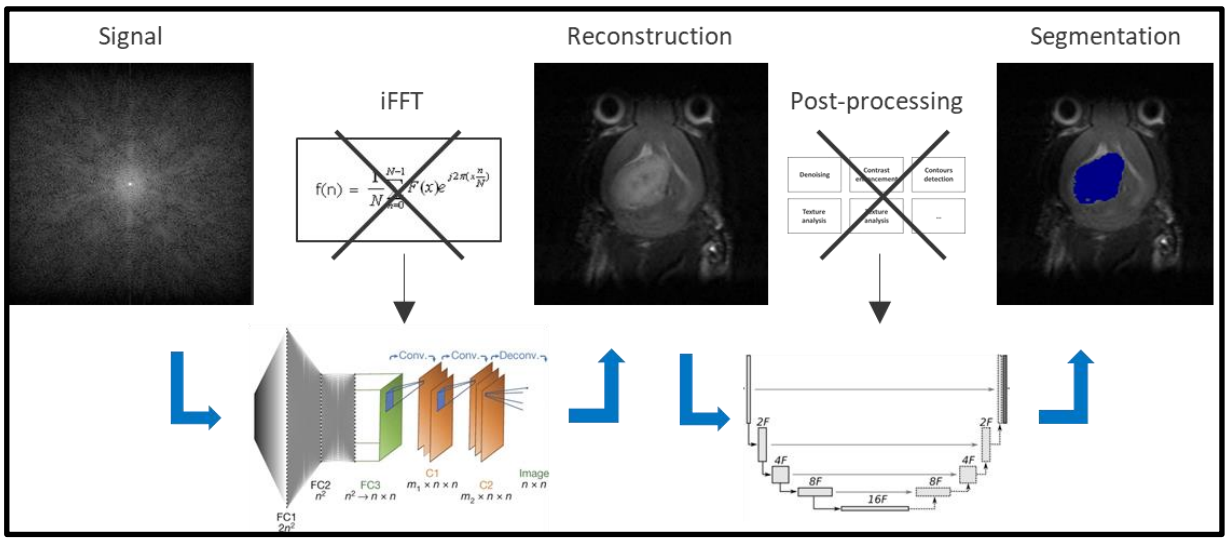

Figure 4: Principle of Task-Oriented Images Reconstruction.

5.2 PaRI

The Nordic Pandemic Research Infrastructure (PaRI) project established by NeIC aims to align with other European COVID-19 initiatives, such as ELIXIR and EOSC-Life, thus contributing to an effective response to pandemics (COVID-19 and future). The challenges faced by this project are interesting for the HPC world but also very related (use of dashboard, data processing, data security aspects, etc.). In addition, the interest in using computing resources by the community (e.g. vaccine research) was highlighted during the pandemic. A connection with this infrastructure has been established, notably through the presence of PRACE members in the follow-up of their activity.

A mailing list and a reference group have been set up by the NeIC/PaRI Project Manager and service leader in PRACE-6IP WP6.2. Regular meetings are held approximately every two months since January 2021 and allow the different communities involved to follow the overall efforts to prepare the ground and to manage future pandemics more effectively. The objective of PRACE, for the time being, is to follow closely the different initiatives in this field in order to better understand the challenges and to bring its expertise if needed.

6 Conclusions

This deliverable reports on existing and new collaboration with other national or European projects, other e-Infrastructures and CoEs. The team of task 6.3 exposes how, through technical understanding of projects, workflows and technical solutions, identification of test cases, POCs, or pilot projects, they develop or design services that are useful for each parties. Sometimes simply providing support, showing the services that exist within PRACE is enough to create links with e-Infrastructures that may then lead to more formal collaborations.

Task 6.3 has done its best to address the planned activities, i.e. the development of new data services, in particular by developing our links with infrastructures using large-scale instruments (SKAO, CERN, or ELI). This target pilot projects and demonstrators that allow implementation of innovative techniques on Tier-0 or Tier-1 systems, such as on-the-fly data processing, or the implementation of machine learning techniques for data processing.

A particular effort of the whole WP6 has been made to implement a pilot project in collaboration with the Fenix-ICEI project to evaluate the potential of integrating a federated

AAI service into PRACE. Moreover, this activity will continue beyond the end of the PRACE-6IP project, the results of this pilot will be used for the future AAI service to be used after PRACE.

Due to the impact of the COVID-19 pandemic, a strong highlight on the digital services needs for health data infrastructures occurred. A growing interest between the HPC community and the health domain has oriented part of our activities towards discussions with health data infrastructures/projects (PaRI, LIH) in order to reflect together on our synergies and make PRACE's computational services more visible for the interested stakeholders and communities in the future.