Portuguese **Distributed Computing** Infrastructure

Requirements for INCD phase 1 pilots (Mi4.1)

Project report

State:	**FINAL**
Dissemination:	**PUBLIC**
Authors:	**Anabela Oliveira, Alberto Azevedo, Catarina Ortigão, João Pina, Pedro Lopes, José Barateiro, Jorge Gomes, André B. Fortunato, António Santos, Armando Pinto, Elsa Alves, João Pedro Santos, Juan Mata, Lourenço Mendes, Paula Beceiro, Paula Freire, Rita Salgado Brito, Ricardo Martins, Teresa Viseu**
URL:	**http://www.incd.pt**

Date	Author	Comments
31-04-2018	Oliveira, A. Azevedo, A.	Initial version
14-06-2018	Oliveira, A. Azevedo, A.	1st version for contributions
9-07-2018	Ortigão, C.	Pilots added.
12-07-2018	Pina, J.	Final version

Cofinanciado por:
COMPETE 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional Lisb@20 FCT Fundação para a Ciência e a Tecnologia

Portuguese **Distributed Computing** Infrastructure

Portuguese **Distributed Computing** Infrastructure

Executive Summary

The present report describes the pilot use cases set up intended for the test and validation of the INCD infrastructure and services. For each use case scenario the requirements and feasibility were analysed and a detailed description for each scenario was elaborated.

Portuguese **Distributed Computing** Infrastructure

Introduction

Approach and objectives

The Portuguese National Distributed Computing Infrastructure (INCD) is a scientific digital infrastructure with the overall objective of providing computing and storage services to the Portuguese academic and scientific community in all domains of knowledge, enabling scientific research and development across all smart specialisation domains. INCD will provide a comprehensive portfolio of ICT services through which other research entities can build and/or deploy their own computing and data services. The services portfolio will include cloud computing, grid computing, advanced computing, data storage and many others to be identified jointly with the user community.

The present report aims at identifying the pilot applications and services from the scientific community to be deployed at INCD infrastructure in order to test and validate the INCD infrastructure and associated services.

Description of the Phase 1 pilots

This section describes the first set of pilots selected for implementation in the INCD project. The chosen use cases were selected from try to represent chosen Each of them will be included in the INCD infrastructure service catalogue and will be available for use by any institution from the portuguese scientific and academic community, powered by the resources from the INCD infrastructure. The selected use cases were chosen to be representative from several portuguese scientific and academic community and will be powered by the resources from the INCD infrastructure.

Pilot 1: BridgeGate: Storing and managing experimental data

This pilot study aims at establishing a platform for storing, managing, visualizing and sharing any set and volume of experimental data, acquired in the realm of research or commercial projects, from small data sets to Big Data. In addition to the database systems, the pilot study will focus on developing a toolset for different user profiles, from those which only require a programming engine and a command line, to those which require dashboards fully developed and curated. An example comprising experimental data acquired in bridge structures monitored at LNEC will be used as a showcase.

Potential users for the outcomes of this pilot study comprise researchers and practitioners, of any scientific area or background, with requirements for managing, visualizing, storing and analysing experimental data sets of any size, obtained from sensors, surveys, etc. Research institutes, public or private companies, which have to manage any type of public of private data, regardless of being confidential, are therefore potential partners and customers.

The developments to be made in this pilot study can be of great importance for process and data normalization and preservation within all science fields using experimental data.

Pilot 2: City Energy

The BEE **(**Building Energy Efficiency) service automates buildings' energy-efficiency assessment, on request, for typified and customized buildings by the user and using building energy consumption data for a period of one year defined by the user.

This service aims to automate the adjustment of usage profiles and specific powers or equipment efficiencies, to minimize the deviation of the simulated consumption values against the

consumptions given by the user. With this model, the annual energy performance of the building is then assessed and compared with the regulatory building to assess the building energy efficiency.

To reduce the gap between building energy simulation and real building operation, an optimization tool for Building Energy Model Calibration is under development to increase the accuracy in the occupancy and equipment operation profiles. The performance gap between energy forecast and actual consumption reaches dramatically high thresholds. This difference can reach about +90% considering electricity, ranging to +130% including the thermal energy, and +350% of CO2 emission in comparison with the expected values.

Modelling properly energy behavior of buildings is a key factor to optimize energy management practices during the building lifecycle. However, it meets several barriers given by factors such as the difference between design and as-built data, simulation settings and real parameters, standard operation schedules and actual users' behavior, etc.

The accessibility of a calibrated and validated building energy model is central in propose accurate thresholds of efficiency. In fact, modeling assumptions can turn into concerns, in terms of robustness and risk, when predicting future performance. This is especially significant in techno-economic viability assessments such as cost-optimal analysis and life cycle cost (LCC) analysis, which are fundamental to evaluate energy efficiency investments, or energy performance contracting (EPC)

Different optimization tools have been developed to find the best trade-off between competitive goals: Energy Reduction Vs Thermal comfort. The optimization problem is typical of the design process, where different design solutions must be compared to achieve one or more competitive objectives. The typical optimization method consists in the analysis of all the alternatives' performances, developing a full factorial plan and simulating all the possible options (brute-force approach or case-by-case scenario). However, this process is extremely time consuming. That is why some optimization tools, based on evolutionary algorithms have been developed to address this issue in a more efficient manner.

The present research project illustrates the application of a methodological approach to simulate building performance variability due to occupancy patterns and equipment's use for a standard building model. The optimization tool is supported by an integrated dynamic model combining a design tool (Google SketchUp), a building energy performance simulation tool (EnergyPlus) and a programming language (Python) to develop the parametric analysis with the aim of multi-objective optimization using genetic algorithms. The methodology is based on the combination of EnergyPlus and python script, which communicate via a coupling function written in python. EnergyPlus is used to model the building energy behavior; it works with text-based format inputs (.idf) and outputs (.csv), which allow the communication with python language. The python module encompasses the

optimization engine and the handling of the input/output files of EnergyPlus and the post-processing phase. Although the optimization algorithm reduces the number of simulations to reach optimal solutions, the simulations requires a significant processing power, for that reason, a cloud computing unit is essential to support the project. The tool is designed for Building Energy managers, Qualified energy technicians and for mechanical/electrical engineering academia and research centers.

Pilot 3: OpenFOAM - Fluid modeling in hydraulic structures and maritime and harbour hydraulics using software OpenFOAM

This pilot offers OpenFOAM installed in a powerful cluster to M.Sc. and PhD university students integrated in research groups. The service targets problems in the topics of hydraulic structures and maritime and harbour hydraulics. Additionally, a user-friendly interface facilitates OpenFOAM usage and encourages collaborations between institutions or enterprises.

OpenFOAM is a totally free open-source software to perform numerical simulations of fluids such as water and air. All the code and functionalities are accessible, allowing the user to adapt the code to their needs, which makes it a perfect tool for research and advanced consulting. Furthermore, besides being very powerful, resourceful and matching the most recent commercial programs, is designed to multi-core computing in clusters.

Nowadays, computational fluid dynamics of hydraulic structures is crucial for research and consultancy activities in many areas, particularly regarding dam spillway chutes and interactions of waves and maritime structures. Numerical modelling is an outstanding complement to physical laboratory modelling allowing the study of multiple solution alternatives before building the physical model. Data collections from the physical model are used to calibrate and validate the numerical model. This way, much more solutions are tested in less time and with less resources. Therefore, most recent projects demand both type of modelling.

This service is being developed in the scope of several Ph.D. thesis.

Pilot 4: StreamSim: 1D/2D integrated modelling in urban systems and streamflow simulation in rivers based on model Basement

This platform allows to carry out an integrated 1D/2D modelling in urban water systems and simulation of river runoff using the combined utilization of SWMM (for sewers, in the case of urban systems) and Basement (for surface runoff modelling) programs. The use of these models allows obtaining 1D results (SWMM) for the hydraulic variables in the sewers, such as water level, velocity, flow rate, flow volume and discharge to the exterior and 2D results (Basement) for surface hydraulic variables, such as water level and velocity.
This platform also intends to promote 2D simulation of flows in rivers using the Basement model. With this application the characteristics of river processes, namely, inundations in flood situations, sediment transport and modifications in the river bed will be obtained. These results are fundamental of river rehabilitation projects, environmental impacts characterization and risk studies.
The potential users for this platform, in terms of urban water systems, are the water management utilities, municipalities, consultants and researchers with an interest in superficial runoff modelling. Regarding specifically the simulation of flow in rivers, the potential users are modellers, researchers, engineers, and stakeholders in the areas of river, environmental and territorial engineering.
The superficial runoff modelling can be carried out for different purposes such as, for example, to study drainage solutions (natural or infrastructural), to project and design drainage networks, or to assess flooding occurrence in urban areas. It also allows to support studies concerning the control of undue inflows into drainage systems. This platform interests the civil, sanitary and environmental engineering scientific areas.

The main benefits of developing this platform are to support the compliance with Decree-Law n°. 115/2010 of 22 October, to reckon wastewater discharges to the water environment, to improve flood control and management, to reduce flooding risk, to contribute to the management of population alerts, to contribute to increasing people and property safety, to integrate the dimensions of hydraulic, structural and environmental performance assessment and to support decision-making on alternative solutions for stormwater management.

Currently, integrated 1D/2D modelling is under slow development and application, as it is highly dependent on computational resources and requires high storage capacity. This platform can be an adequate solution to overcome these constraints, as it will be able to integrate tools that currently are independent, reduce developing and simulation times, and simulate a high number of scenarios, thus contributing to research projects robustness.

This platform facilitates the dissemination and application to several cases and projects, nowadays limited with the existing tools, allowing for feasible developments in areas as relevant as resilience and risk management in urban areas and flow interconnections between urban areas and receiving water bodies, among others.

Pilot 5: On-demand Operational Coastal Circulation Forecast Service - OPENCoastS

OPENCoastS is a generic framework to establish, in an interactive way, on-demand forecast systems for the coastal circulation over the user's region of interest. This service will be maintained in operation for the period specified by the user, with a limit of one month (forecast period extension can be requested). The service generates daily 48-hour predictions of water levels, velocities and wave parameters, based on the numerical simulation of the relevant physical processes.

This service will be important for all players with management responsibilities and economic activity in the coastal regions, including the scientific communities. The service will be supported by a Web tool that will facilitate its use by users without strong computational background.

As the service will promote open access to the generated forecast data (by providing additional advantages to all users that decide to share their work), it is expected that OPENCoastS will contribute to the generation of a database of detailed information for the Portuguese coast. This database can be used for climate studies as well as support new services and products fed by the developed forecasts . The service targets the areas of coastal engineering and oceanography.

Pilot 6: WorSiCa: Cloud-based platform for coastline monitoring based on satellite Sentinel images processing

Coastal zones are vulnerable to the climate changes. The sea level rise and the instability of weather conditions cause damage to the nearby infrastructures. Costa da Caparica has maritime infrastructures nearby, in risk of being overtopped by the sea. A web platform on the cloud is being developed to monitor the coastline, by processing inputs (Sentinel imagery from the user) and storing results (shapelines). Parallel computation is required to process high resolution satellite imagery and auxiliary data (topography, bathymetry, others). The study case for this application will be Costa da Caparica, and will be applied to other coastal zones.

This service allows users to know and follow the historical evolution of the coast line and the possible hazards that will come with its advance. This will help to support plans to mitigate the risk of overtopping infrastructures, and the loss of marginal zones.

This service is useful for all entities who are working on the coastal zone, including the scientific and technical community.

Pilot 7: RESTATE_INCD - Towards a decision support system for evaluation of the safety of large concrete dams

Several critical infrastructures for modern societies, such as dams and bridges, are overcoming its life expectancy. Specific monitoring and maintenance plans are fundamental to guarantee safety

conditions of these infrastructures. For an effective decision making about the assessment of the structural behaviour under operating conditions, confidence in measured data is crucial. Additionally, it must be possible to interpret these data (through adequate data based methods) in order to properly assess the structural behaviour and condition (with the support of reliable numerical models).

This pilot project aims to fill the gap between the development of new methodologies and the availability of computational resources for structural safety assessment and information management of critical infrastructures.

Based on this pilot project, expert users will be able to access and explore a dedicated software environment, allowing them to use and develop advanced algorithms, in the context of both data quality control, and analysis and interpretation of the structural behaviour of critical infrastructures.

In practice, civil engineers, researchers and academics will be able to develop, test and use new advanced algorithms, mainly based on machine learning tools, and taking advantage of the available computational environment.

To demonstrate the potential of the pilot project, an application tool based on a methodology published in a scientific paper [Mata, 2013] will be presented. The purpose of this tool is to identify the effect of daily variation of air temperature on the structural response of a concrete dam.

Pilot 8: SCHISM - Platform for Coastal Circulation Model Service

This pilot targets the implementation of a numerical simulation service of estuarine and coastal processes with the SCHISM model, an open source community model. Based on a circulation core, SCHISM includes modules for marine agitation, morphodynamics and water quality. Some of these modules were developed totally or partially at LNEC. Depending on time and resources availability, the implementation of a coupled wave model is also considered. A supporting interface is considered, to facilitate the use of the model as well as results download and visualization. Some preprocessing and visualization programs, available in the software package or already developed at LNEC, may be used. SCHISM's manual is available at http://ccrm.vims.edu/schismweb/schism_manual.pdf. A non-exhaustive list of publications is available at http://ccrm.vims.edu/schismweb/schism_pubs.html

To demonstrate the usefulness of this platform, a tidal and storm surge model will be developed in the Atlantic NE, and simulations will be carried out over a period of several decades. The domain of calculation will include mainland Portugal and the archipelagos of the Azores and Madeira. The

results will allow to determine extreme levels throughout the Portuguese coast, including the Azores, thus contributing to the new AIR Center.

All entities with responsibility and interest in coastal areas, including the academic, scientific and technical communities are identified as potential users of this service. This pilot is included in the coastal and environmental engineering and oceanography area. SCHISM is routinely used in many research institutions worldwide, in research and consultancy projects. The use of the service platform would therefore be immediate.
While commercial or semi-commercial models often privilege ease-of-use over performance, models such as SCHISM are primarily intended for experienced users. As a consequence, these models can be difficult to use, for example in terms of compilation or generation of some input files. This application will allow the scientific community to enjoy a well-paralleled, process-sophisticated model through a friendly Web tool that will eliminate barriers to its use. Thus, this platform will promote the use and dissemination of SCHISM for all public entities covered by the INCD protocol. This service will also allow the support of disciplines of numerical methods and modeling for courses in Physical Oceanography, Coastal Engineering or Environmental Engineering.

Pilot 9: RESCCUE RAF App: Climate change city resilience calculator

This tool provides a framework to assess urban resilience to climate change, with a focus on water, considering an objective-oriented approach and four resilience dimensions: organizational, considering governance relationships; spatial, covering urban space and environment; functional, focused on strategic services in the city (water, wastewater, stormwater, waste, energy and mobility); and physical, centred on infrastructure of these services. The resilience objectives are described through key criteria, expressing different points of view, which are evaluated by indicators or metrics. In this given scope, the metrics are described and associated to reference values, providing a user-friendly assessment to support a structured diagnosis. The App allows the use of a defined structure based on dimensions / objectives /criteria / metrics, specifically designed to address the referred scope.

The potential users for this tool are the municipalities, utilities of urban strategic services and consultants and researchers with interest in cities or services resilience, climate change or any other multidisciplinary assessment. The App can be used as a tool to support assessment, diagnosis and decision-making as well as the development of resilience plans, to monitor progress of a city or service or to compare different cities or services. This tool is of interest to civil, urban water and environmental engineering, urban resilience and climate change adaptation scientific areas.

The main benefits of developing this platform are to contribute to the resilience assessment of cities and services, to the development of resilience plans, to address the contribution of urban services to the city's resilience and to acknowledge improvement opportunities and monitor progress.

This tool facilitates the dissemination and application to several cases and projects, nowadays limited with the existing tools, allowing for feasible developments in areas as relevant as resilience and risk management in urban areas.

Pilot 10: LHC Experiments - ATLAS and CMS

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator. The LHC consists of a 27-kilometre ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles (hadrons) along the way. It accelerates and collides protons, and also heavy lead ions.

LIP is deeply involved in the CERN LHC endeavor, contributing from the very beginning to the two largest LHC experiments, ATLAS and CMS. The LHC allows to explore the high energy frontier of particle physics, to go into uncharted territory beyond the current knowledge and try to understand the fundamental building blocks and the forces that shape the Universe.

ATLAS is the largest of the four gigantic experiments operating at the LHC where protons and lead ions are made to collide at unprecedented high energies and luminosities.

The Compact Muon Solenoid (CMS) is a general-purpose detector at the Large Hadron Collider (LHC). It has a broad physics program ranging from studying the Standard Model (including the Higgs boson) to searching for extra dimensions and particles that could make up dark matter. Although it has the same scientific goals as the ATLAS experiment, it uses different technical solutions and a different magnet-system design.
As the Tier-2 of the Worldwide LHC Computing Grid (WLCG), INCD supports the LHC experiments, delivering computing and storage capacity to process and analyze data from the ATLAS and CMS groups.

Pilot 11: Cosmic Rays Physics

Cosmic ray physics is an active field of research, with many ongoing experiments addressing questions on their origin, nature, acceleration and propagation. The vast range of energies of cosmic rays implies that different detection methods are used, from space-based experiments to ground-based giant air shower detectors.

Auger Collaboration @ LIP

The Pierre Auger Observatory covers an area of 3000 km2 in the Pampa Amarilla, Argentina. It consists of 1600 detectors separated by 1.5 km that sample the shower of millions of particles produced when the highest energy cosmic rays hit the atmosphere. In dark nights, 27 telescopes detect the ultraviolet light emitted by the showers. The Observatory is taking data since 2004, and some breakthroughs have been achieved. Nevertheless, several open questions remain concerning the nature and origin of the highest energy cosmic rays. The observatory will continue operations until 2025 and is currently being upgraded, to enable a better understanding of the electromagnetic and muonic shower components. R&D for future cosmic ray detectors also takes place at the Observatory site.

The LIP group of the Auger Collaboration is mainly focused on the full exploitation of the particle physics potential of the Observatory, namely on the efforts to understand hadron interactions at high energies through a window that is largely complementary to the LHC.

LATTES – Gamma-Ray detector for astrophysics

The energy threshold of the air shower experiments presently in operation or in construction remains very large and unable to bridge with data from satellite-borne experiments. All the extensive air shower (EAS) experiments presently in operation or under construction are in the Northern hemisphere. The recent observation of the first multi-messenger event combining the detection of gravitational and electromagnetic waves triggered a growing international interest in building such an Observatory in South America.

The goal of LATTES is to design, prototype and construct a ground array able to monitor the Southern gamma-ray sky above 50 GeV, bringing to ground the wide field-of-view and large duty cycle observations characteristic of satellites, with comparable sensitivity and a cost one order of magnitude lower. Such an instrument will be a powerful time-variance explorer covering a missing space in the global multi-messenger network of gravitational, electromagnetic and neutrino observatories. It will be able to issue pointing alerts to IACTs (Imaging Atmospheric Cherenkov Telescopes) and thus fully complementary to CTA. It will collect abundant and highly relevant data and play a fundamental role in the search for emissions from extended regions, as the Fermi bubbles or dark matter annihilation regions.

LATTES proposes an innovative concept: a compact EAS array of hybrid detector units, covering an area of at least 20 000 m2, to be placed at high altitude (about 5 000 m above sea level, a.s.l.) in the Southern hemisphere. Each detector unit combines two autonomous Resistive Plate Chambers (RPCs), with good space and time resolution, with a Water Cherenkov Detector (WCD), ensuring

trigger efficiency and efficient background rejection. This detector concept was simulated using an end-to-end realistic simulation of both gamma and proton (background) showers.

The simulation of an extensive air shower requires the treatment (interaction and propagation) of millions of particles across the atmosphere until reaching the ground level. These particles are then injected in a detailed detector simulation that takes into account all processes involved in their detection. This required a huge amount of computing power and storage resources.

Showers are simulated using CORSIKA and particle interactions with the detector make use of the Geant4 toolkit. In order to produce an accurate sensitivity plot and check LATTES performance more that 10^7 showers need to be simulated and reconstructed taking massive hours of CPU hours and large datasets.

INCD already supports the data analysis of the Auger collaboration as being part of the computing resources in the framework of the collaboration. With this pilot we plan to increase the Auger data analysis adding computational resources and contributing to the design of LATTES by providing HTC and storage resources together with expertise on compilation and tuning of the software applications.

Pilot 12: Space Applications

Space exploration in one of the key areas of application of particle physics instruments and methods, and LIP became a recognized partner in the Space community, and particularly in the Portuguese participation in European Space Agency (ESA).

The LIP group for space applications is focused on the study of space radiation environments and their effects. The competences developed include all the technologies identified on ESA's roadmap for this domain: radiation environment measurement technologies; radiation environment modeling; radiation effects analysis tools; test characterization and radiation hardness assurance of Electrical, Electronic and Electro-mechanical components (EEE). A wide range of activities was developed involving different institutions both in academia and in national industry.

Space is a crucial and competitive market and requires state of the art computing facilities adapted to their needs. INCD supports the work of this group providing computation through the HTC/HPC cluster and support to the data analysis workflows.

Pilot 13: Neutrino Physics

The Sudbury Neutrino Observatory (SNO) measured the oscillations of solar neutrinos, i.e., their transformations from one type to another (Nobel prize in Physics 2015). The detector is located 2 km deep underground, in SNOLAB, Canada. An acrylic sphere with 12 m diameter and 5 cm thickness, that contained 1000 tons of heavy water, is surrounded by 9500 light sensors. The SNO+ experiment follows from SNO, replacing the heavy water with liquid scintillator to increase the sensitivity to other neutrino physics signals. The LIP Neutrino Physics group joined the SNO experiment in 2005, and is a founding member of the SNO+ international collaboration. The main goal of the experiment is the search for neutrinoless double-beta decay, by loading the scintillator with large quantities of Tellurium. Several other low-energy, low-background, physics topics are also part of its program: antineutrinos from nuclear reactors and the Earth's natural radioactivity, solar and supernova neutrinos, and searches for new physics.

Like the remaining High Energy Physics experiments this type of studies have complex simulations and data analysis workflows which requires large amount of computational hours and dedicated storage. INCD supports this work by providing HTC and storage resources.

Pilot 14: GBIF - Portuguese biodiversity data portal

GBIF is an intergovernamental initiative where countries collaborate promoting free and open access to biodiversity data. Therefore, the Portuguese Node launched in 2016, the *Biodiversity Data Portal of Portugal*, using the *Atlas of Living Australia* platform, implemented on the INCD Cloud Service. It enables the access to data on species occurrence, publishers, images and more. It currently supports more than three million biodiversity occurrences, from which 2 506 860 published by Portuguese institutions. The aim is to support users to share, search and use Portugal's biodiversity data.

GBIF is a successful story at the Iberian level and in the next few months it will add new modules (new species catalogues, species lists and regional mapping) but also plans to integrate the Portuguese and Spanish portals into a single Iberian platform allocated at INCD Cloud Service.

Pilot 15: Computational Multi-omics

Our mission is to understand the (epi)genome biology and its impact on cancer and other diseases using computational multi-omics approaches. Such methods rely on the statistical analysis and integration of large scale data (high-throughput sequencing, microarrays, proteomics, high-throughput screening) and clinical/phenotypic data.

The wealth of genomic data generated by the International Cancer Genome Consortium (ICGC), Roadmap Epigenomics, ENCODE and GTEX has allowed the genome, epigenome and transcriptome profiling of a wide range of normal and pathological tissues. Given the accumulation of this huge amount of biological information, handling and mining big data is becoming increasingly mandatory for major research. However, more than the data processing and integration, the key challenge relies on extracting information and generating scientific knowledge. Therefore, we aim to achieve biomedical advances through novel ways of combining large multi-omics and phenotypic data. Hence, we integrate large-scale profiles at different levels: genome (information within the DNA sequence and mutations), epigenome (DNA methylation, histone modifications, nucleosome positioning and chromosome conformation), transcriptome (RNA expression and isoforms). Integration and visualization of such complex data sets is crucial for interpretation and decoding of the underlying biology associated pathological conditions.

This pilot will use the HPC infrastructure and very large memory (VLM) machines provided by INCD.

Pilot 16: INNUENDO platform

MRamirez Lab from Instituto de Medicina Molecular IMM (http://im.fm.ul.pt), is developing the INNNUENDO platform (http://www.innuendoweb.org/) for the analysis of high throughput sequencing (HTS) data in the context of clinical microbiology and genomic epidemiology. High Throughput Sequencing created a paradigm shift in several research areas. In Microbiology, it allowed the sequencing of a bacterial genome in a matter of days, and the subsequent analysis of the raw data allows a never-before achieved level of detail on the bacterial genome structure. This has direct applications for studying the evolution of antibiotic resistance, surveillance of bacterial clones that are known to cause disease or evade vaccines and allows the tracking of microbial outbreaks, which much more detail than any of the previously used technologies. However, the needed bioinformatics analyses are computational intensive and existing software needs dedicated computing environments and know-how to be effectively run, causing the need for specialized personal. The goal of the INNUENDO platform is to allow the streamlined use of that software, so it can be used in research and reference laboratories, with minimal training, to effectively perform bacterial strain surveillance and outbreak detection, using standardized and validated analytical approaches.

The final goal of this pilot is to provide an INNNUENDO platform available to the portuguese scientific community. The use of the INCD infrastructure is paramount for this goal, since it allows testing different scenarios of the setting up the platform and its subsequent configuration and optimization.

Summary of pilots

Table 1 presents a summary of the phase 1 pilots and their services, including the tools used.

Table 1 - Summary of phase 1 pilots

Pilot number	**Pilot acronym/short name**	**Software**	**Scientific areas**
1	BridgeGate	programming dashboard	Civil Engineering
2	City Energy	Energy Plus model + programming dashboard	Computational Mechanics
3	OpenFOAM	OpenFOAM model	Hydraulics
4	StreamSim	Basement SWMM models	Hydraulics
5	OPENCoastS	WIFF platform + SCHISm model	Physical and Operational Oceanography
6	WorSiCa	Copernicus data scripting	Earth Observation, Coastal engineering
7	RESTATE_INCD	programming dashboard	Civil Engineering
8	SCHISM	SCHISM modelling system	Physical Oceanography, Coastal Engineering and Environmental Engineering
9	RESCCUE RAF APP	Resilience Assessment Framework	urban engineering, urban resilience, climate change adaptation
10	CERN/LHC Experiments - ATLAS and CMS	process and analyze data from the ATLAS and CMS experiments. HEP software foundation.	Particle Physics; Accelerators; High energy Physics

11	Cosmic Rays Physics	CORSIKA simulator and Geant4 toolkit	Cosmic Rays Physics
12	Space Applications	space radiation analysis tools and models	Physics - space radiation.
13	Neutrino Physics	simulation and data analysis workflows	Neutrino Physics
14	GBIF	Platform to share, search and use Portugal's biodiversity data	Biology - biodiversity
15	Multi-omics	statistical analysis and integration of large scale data tools	Biology - cancer genomics
16	INNUENDO	Software for bioinformatics analyses	Microbiology and genomic epidemiology

Pilots requirements of infrastructure usage

The requirements for each of the pilots from the previous section is listed here, regarding both external and infrastructure resources. Regarding INCD, this preliminary list of requirements will help infrastructure managers to organize and manage access for pilots execution.

Pilot 1: BridgeGate: Storing and managing experimental data

Based on the analysis of the management and storage systems currently in place in LNEC, the setup of the pilot study and of the showcase will roughly require:

- processing: 8 VCPU (70K CPU hours / year);
- RAM: 16 Gb;
- storage: 4 TB;

Taking into account the confidential nature of some applications, this service requires mechanisms for authentication and access control.

Pilot 2: City Energy

This pilot will be based on the work carried out under a Ph.D. thesis work. Besides EnergyPlus, an open-source model, no other external resources are anticipated. The usage of infrastructure resources cannot be defined at this stage, depending on the extent of the analysis to be carried out

in the demonstration of the service. However, the experience of LNEC in the use of EnergyPlus, in the scope of the ADENE-funded ECO-AP project, anticipates that these resources should be small both in terms of computing power and storage.

Pilot 3: OpenFOAM: Fluid modeling in hydraulic structures and maritime and harbour hydraulics using software OpenFOAM

A typical hydraulic structure simulation has the following estimated requirements:

- Dedicate services: 6 VCPU (52K CPU hours / year);
- processing: 32 to 256 cores and 2 GB per core;
- Total processing: 1M CPU hours / year ;
- Local storage: 6 TB.

Pilot 4: STREAMSIM: 1D/2D integrated modelling in urban systems and streamflow simulation in rivers based on model Basement

The requirements in terms of INCD resources for this platform are computing capacity, storage, graphic display and user authentication with differentiated access. Given the development to be done in the scope of this pilot (integrated in an ongoing Ph.D. thesis), an estimate of the resources is not yet possible.

Pilot 5: On-demand Operational Coastal Circulation Forecast Service-OPENCoastS

The requirements in terms of INCD resources for this platform are computing capacity, storage, graphic display and user authentication with differentiated access. Taking into account on previous studies for the Tagus river the estimated computacional requirements are the following:

- Dedicate services: 4 VCPU (35K CPU hours / year);
- Total processing: 300K CPU hours / year (HTC and Cloud computing);
- RAM: 2 Gb per core;
- storage: 1 TB ;

This service also the following external providers:

- sea and turmoil forecasts from LNEC;
- weather forecast from: NOAA, ECMWF, MeteoGalicia
- data or climatology of river inflows

Pilot 6: WorSiCa: Cloud-based platform for coastline monitoring based on satellite Sentinel images processing

The input data of this service will be Sentinel imagery to generate a waterline, and in-situ information (from sensors and models) to correct that waterline and generate a correct coast line. Sentinel data is freely available. The estimated computacional requirements are the following:

- Dedicate services: 4 VCPU (35K CPU hours / year);
- Data processing: 1 to 12 cores with 2 Gb per core;
- GPU processing : 2400 hours GPU time (K80 Nvidia);
- Total processing: 235K CPU hours / year;
- Local storage: 10 TB;

This service also the following external providers:

- Satellite imagery from the ESA hub repositories.
- In-situ information (sensors and models), topography and bathymetry data from external sources.

Pilot 7: RESTATE_INCD - Towards a decision support system for evaluation of the safety of large concrete dams

Hardware requirements will depend on the specific tools to be built by the user and cannot be accurately estimated at this time. Desirable software tools to be made available to the final users include:

- The R Project for Statistical Computing;
- Python;
- Latex;
- Jupiter;
- Paraview;

Taking into account the confidential nature of some applications, this service requires mechanisms for authentication and access control.

Pilot 8: SCHISM: Platform for Coastal Circulation Model Service

The input data of this service will be an unstructured simulation grid and meteo-oceanographic data to impose the initial and boundary conditions. The platform will be developed in a Linux/Unix environment.The estimated computacional requirements are the following:

- processing: 1 to 8 cores with less than 1 Gb per core;
- total processing: 250K CPU hours / year;
- Local storage: 10 TB;

This service also the following external providers:

The imposition of boundary conditions usually requires data or results from other large-scale models. For example, for the Portuguese coast, river flow data are available on the SNIRH, meteorological hindcasts on NOAA and ECMWF, oceanographic hindcasts on ECMWF and Copernicus, etc. Some institutions also have results of hindcasts or own forecasts, as is the case of LNEC, IH, IPMA, IST, etc. It would be very useful to include in the platform transparent ways of attaching some of these databases to the model. Several scripts to facilitate this task (to download files and transform their formats) are already developed and can be provided.

Pilot 9: RESCCUE RAF APP: Resilience Assessment Framework Tool

Hardware requirements will depend on the specific tools to be built by the user and cannot be accurately estimated at this time.

- Dedicate services: 2 VCPU (17K CPU hours / year);
- local storage: 100 GB;

Pilot 10: LHC Experiments - ATLAS and CMS

The INCD lodges and operates the national Tier-2 The portuguese T2 provide disk storage and processing capacity for simulation and data analysis of ATLAS and CMS according to the Memorandum of Understanding [WLCG_MoU] signed between Portugal and Cern.

The T2 it must be integrated in the Worldwide LHC Computing Grid (WLCG), EGI and IBERGRID distributed e-infrastructure and the amount of resources supplied are negotiated a yearly base by CERN and national representatives. This agreement also implies that part of the Tier-2 resources are shared with external communities. The Tier-2 operation is ruled by signed Service Level Agreements (SLA) with EGI and WLCG. The commitment of the SLA impose to all services a good quality of service monitored by the reliability, availability and ticket response time. This quality factor are monthly reported by both of the infrastructures. Each T2 must also have a security team responsible to act in case of vulnerabilities discovered in the infrastructure.

According to the 2017-2018 signed agreement the portuguese Tier-2 must deliver the following resources:

- 3200 HSPEC to both ATLAS and CMS (6400 in total) equivalent to 7M CPU hours / year
- 220 TB of online storage to ATLAS
- 200 TB of online storage to CMS

Also a Tier-2 his required to run the following services:

- APEL-SSM and APEL-client for accounting
- ARGUS which is an authorization service
- BDII system information based on ldap.
- CREAM-CE: computing element service responsible to accept and submit jobs to the local batch system.
- Frontier-Squid: squid cache system
- PerfSonar: network monitoring service
- StoRM: storage resource manager
- SGE: local batch system that accepts requests from the computing element
- VOBOX: service used for the CMS data transfers
- xRootD: service that allows low latency and scalable data access in federated environments. Used as failover mechanism for transfers.

Pilot 11: Cosmic Rays Physics

The cosmic ray community uses different data analysis techniques and software. Like the LHC experiments, AUGER experiment have a distributed production which makes use of the portuguese T2 (explained in pilot 10) but in opposition of the LHC all of the user analysis is all made locally. The requirements are the following:

- Dedicate services: 10 VCPU (87K CPU hours / year).
- Total processing: 700K CPU hours / year;
- Online Storage: 5 TB;
- Local Storage: 20 TB;

Pilot 12: Space Applications

The Space Applications make use of the GEANT4 toolkit in order to study the space radiation environments effects in humans and electronics. The estimated computational requirements are the following:

- Total processing 300K CPU hours / year;

- Local Storage 5TB local storage;

Pilot 13: Neutrino Physics

Like the cosmic ray studies use case (pilot 11) the neutrinos community from SNO have a distributed production which makes use of the portuguese T2 and all of the user data analysis is made locally. The estimated computacional requirements are the following:

- Total processing: 100K CPU hours / year;
- Online Storage: 2 TB;
- Local Storage: 20 TB;
- Dedicated services: 20 VCPU (175K CPU hours / year);

Pilot 14: GBIF - Portuguese Biodiversity Data Portal

The GBIF biodiversity portal it's based on the Atlas of Living Australia and currently supports over 3 millions biodiversity occurrences, from which 2.5 million published by Portuguese institutions. This services makes use of cloud service with an estimated computacional requirements:

- Dedicate services 50 VCPU (438K CPU hours / year);
- Local Storage: 2 TB;

Pilot 15: Computational Multi-omics

The analysis of genome data are not only very CPU intensive (HPC) but also requires very large memory (VLM), higher than 500GB of RAM. The estimated computacional requirements are the following:

- processing: 400K CPU hours / year (HTC) with VLM;
- processing: 500K CPU hours / year (HPC);
- Total processing: 900K CPU hours / year (HPC and HTC) ;
- Local Storage: 2 TB local storage.

Pilot 16: INNUENDO platform

The analysis of high throughput sequencing (HTS), like for the pilot regarding multi-omics data analysis, is very CPU intensive (HPC) but also requires very large memory (VLM), higher than 500GB of RAM, computing resources. For this pilots special computing machines are required and the estimated computacional requirements are the following:

- Dedicate Services: 30 VCPU (250K CPU hours / year);
- processing: 300K CPU hours / year (HPC) with VLM;
- Total processing: 560K CPU hours / year (HPC and Cloud computing);
- Local Storage: 2 TB;

Summary of requirements

Table 2 - Summary of phase 1 pilots' requirements for infrastructure usage.

Pilot number	Pilot acronym/short name	Technology requirements	INCD resources
1	BridgeGate	Cloud resources and online storage	8 VCPU 4 TB online storage
2	City Energy	local job cluster	to be defined latter
3	OpenFoam	cloud resources and local cluster HPC jobs	6 VCPU 1M CPU / hours / year 6 TB local storage
4	StreamSim	local job cluster	to be defined latter
5	OPENCoastS	Cloud resources and local jobs processing	4 VCPU 300K CPU hours / year 1 TB local storage
6	WorSiCa	Cloud resources and local HPC and GPU jobs processing	2400 hours GPU time (K80 Nvidia) 235K CPU hours / year 10 TB local storage
7	RESTATE_INCD	local job cluster	to be defined latter
8	SCHISM	local cluster HPC jobs processing	250K CPU hours / year 3 TB online disk
9	RESCCUE RAF APP	Cloud resources with DB storage	2 VCPU 50 Gb storage
10	LHC Experiments - ATLAS and CMS	Grid middleware and local jobs processing	6400 HSPEC or equivalent in computing CPU hours (7M CPU hours / year) 600TB online storage 100TB local storage 100 VCPU for services

11	Cosmic Rays Physics	Grid middleware and local jobs processing	700K CPU hours / year 5 TB online disk 20 TB local storage 10 VCPU for supporting services
12	Space Applications	local cluster jobs processing	300K CPU hours / year 5TB local storage
13	Neutrino Physics	Grid middleware and local cluster jobs processing	100K CPU hours / year 2 TB online disk 20 TB local storage 20 VCPU for supporting services
14	GBIF	Cloud resources and local storage	50 VCPU 2 TB local storage
15	Multi-omics	Access to very large memory machines (VLM)and local cluster HPC jobs processing	400K CPU hours for VLM 500K CPU hours for HPC jobs 2 TB local storage
16	INNUENDO	Cloud resources and local jobs processing	30 VCPU 2 TB local storage 300K CPU hours for HPC jobs

Conclusions

The Phase 1 set of pilots to be developed in the scope of the INCD project was presented herein. Infrastructure resources for each pilot are identified along with a summarized description of the service, its tools and its importance for both Portuguese research and society.

The selected pilots cover a range of areas and instruments, including numerical models, simulators and tools for facilitated data processing and preservation. The selection was based only the usefulness of the services as well as the importance of its supporting tools, preference being given to those widely used in the scientific community.

Implementation of these pilots will be done by LNEC, responsible for its thematic development and operational implementation for INCD resources, and LIP, responsible for its integration in the

infrastructure. Associação INCD will handle service usage and monitoring, in particular the accounting of the associated scientific production.

References

Mata, J.; Tavares de Castro, A.; Sá da Costa, J. “Time-frequency analysis for concrete dam safety control: Correlation between the daily variation of structural response and air temperature”. Engineering Structures. Elsevier. 48(3):658–665. 2013. doi:10.1016/j.engstruct.2012.12.013